Brief Study on
Prostate Cancer Proteomics
using
Protein-Protein Interaction Networks

A thesis

submitted in partial fulfillment of the requirement for the

Degree of

Master of Computer Science and Engineering

of

Jadavpur University

By

Shiladittya Paul

Registration No.: 140759 of 2017-2018

Examination Roll No.: M4CSE19006

Under the Guidance of

Dr. Ram Sarkar

Department of Computer Science and Engineering

Jadavpur University, Kolkata-700032

India

2019

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

Certificate of Recommendation

This is to certify that the dissertation entitled "**Brief Study on Prostate Cancer Proteomics Using Protein-Protein Interaction Networks**" has been carried out by **Shiladittya Paul** (University Registration No.: 140759 of 2017-18, Examination Roll No.: M4CSE19006) under my guidance and supervision and be accepted in partial fulfillment of the requirement for the Degree of **Master of Computer Science and Engineering**. The research results presented in the thesis have not been included in any other paper submitted for the award of any degree in any other University or Institute

……

Dr. Ram Sarkar (Thesis Supervisor)

Department of Computer Science and Engineering

Jadavpur University, Kolkata-32

Countersigned

…….

Dr. Mahantapas Kundu

Head, Department of Computer Science and Engineering,

Jadavpur University, Kolkata-32.

…….

Prof. Chiranjib Bhattacharjee

Dean, Faculty of Engineering and Technology,

Jadavpur University, Kolkata-32.

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

Certificate of Approval*

This is to certify that the thesis entitled **"Brief Study on Prostate Cancer Proteomics Using Protein-Protein Interaction Networks"** is a bonafide record of work carried out by **Shiladittya Paul** in partial fulfillment of the requirements for the award of the degree of **Master of Computer Science and Engineering** in the Department of Computer Science and Engineering, Jadavpur University during the period of August 2017 to May 2019. It is understood that by this approval the undersigned does not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose for which it has been submitted.

……

Signature of Examiner 1

Date:

……

Signature of Examiner 2

Date:

*Only in case the thesis is approved

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

Declaration of Originality and Compliance of Academic Ethics

I hereby declare that this thesis entitled "**Brief Study on Prostate Cancer Proteomics Using Protein-Protein Interaction Networks**" contains literature survey and original research work by the undersigned candidate, as part of his Degree of **Master of Computer Science and Engineering**.

All information has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name: Shiladittya Paul

Registration No: 140759of 2017-18

Exam Roll No.: M4CSE19006

Thesis Title: Brief Study on Prostate Cancer Proteomics Using Protein-Protein Interaction Networks

…..……..

Signature with Date

Acknowledgment

I would like to start by thanking my family, my friends and my teachers for helping me deploy all the right resources and for shaping me into a better human being.

I would like to express my deepest gratitude to my advisor, **Dr. Ram Sarkar**, Department of Computer Science and Engineering, Jadavpur University for his admirable guidance, care, patience, mental support and for providing me with an excellent atmosphere for doing research.

Words cannot express my indebtedness to **Mr. Sagnik Sen**, Senior Research Fellow, Department of Computer Science and Engineering, Jadavpur University for his amazing guidance and supervision. I am deeply grateful to him for the long discussions that helped to enrich the content of this thesis. Without his enthusiasm, encouragement, support and continuous optimism this thesis would hardly have been completed. I would like to thank Prof. **Mahantapas Kundu**, Head, Department of Computer Science and Engineering, Jadavpur University, for providing me with moral support at times of need. I would also like to thank Prof. **Mita Nasipuri**, Department of Computer Science and Engineering, Jadavpur University for her amazing supervision and guidance.

I am grateful to **Dr. Brijesh Sriwastava**, Assistant Professor, Department of Computer Science and Engineering, Government College of Engineering and Leather Technology, Kolkata for sharing the data used in this study and also his valuable ideas towards this thesis.

I am also thankful to **CMATER** Laboratory for giving me the proper laboratory facilities for carrying out my work.

I would like to thank my junior **Mr. Manosij Ghosh** and **Mr. Kushal Kanti Ghosh** without them this thesis could not be completed as they played a major role by solving some issues with algorithms and methods.

I would like to specially mention my senior **Mr. Nirmal Das**, who provided me with the template for this thesis. I am highly grateful to **Mr. Neelotpal Chakraborty** who provided valuable technical knowledge.

I am also thankful to **Dr. James Coker**, Associate Professor, University of Maryland, University College for his great lectures for a deeper understanding with proteomics.

I would like to specially mention **Ms. Mousumi Ghosh** for her love, support and motivation.

I am also thankful to my friends, **Mr. Sankar Bhattacharjee, Mr. Mahasis Sigha Roy, Mr. Ayan Karmakar and Mr. Madhusudan Ghosh** for their support and Motivation.

This thesis would not have been completed without the inspiration and support of a number of wonderful individuals — my thanks and appreciation to all of them for being part of this journey and making this thesis possible.

……..

Shiladittya Paul

Registration No: 140759 of 2017-18

Exam Roll No.: M4CSE19006

Department of Computer Science & Engineering

Jadavpur University

Contents

List of figures

List of Tables

Abbreviations

PPI - Protein Protein Interaction

PPIN – Protein Protein Interaction Network

HGP - Human Genome Project

PCa - Prostate Cancer

BPH - Benign Prostatic Hyperplasia

PSA - Prostate-Specific Antigen

FCR - Ferritin-Creatinine Ratio

HSP - Heat Shock Proteins

PDI - Protein Disulfide Isomerase

RCC - Renal Cell Carcinoma

SCLC - Small Cell Lung Cancer

AR – Androgen Receptor

Abstract

The biggest challenge to the scientists of the current decade is cancer. At the pre-stage cancer is treatable. So the molecular level research is going on. To detect cancer at the primary stage, suitable protein biomarker is needed to analyze. so now the major focus of cancer research is to determine correct biomarkers to solve it by a proper clinical solution. By using suitable biomarkers, it is possible to detect and track cancer on the right path. Till today, continuous going UniProt database has 156637804 entries, on that 171065 entries belongs to human and only 20421 entries of human are manually verified. In this large database to check each and every protein manually that either it is suitable as a cancer biomarker or not, is a very difficult process as well as very time taking. To make this task easier and faster, the computational approach is needed. Using existing biomarkers which have biological and clinical evidence this thesis proposed a new approach to find out influential proteins. So that the scientist no need to check the whole database, instead of checking the biological nature of that influential protein to find out new biomarker. In this thesis paper there is a brief study on prostate cancer and conclude that, by applying this computational approach for biomarkers finding, some influential proteins will come out as a result. From this influential proteins, it is possible to find new biomarkers by checking biological significance. After implementing the method on prostate cancer known biomarkers, a deep biological survey check is performed and find that 60% influential protein related to cancer.

Chapter 1

Introduction

By many biological experiments, it has been revealed that Proteins are the main agents of biological function to determine the phenotype of all organisms. From the previous study on molecular biology, we saw that Proteins rarely act alone as their functions tend to be regulated. Many molecular processes within a cell are administered by molecular machines that are constructed by their PPIs from a large number of protein components. So understanding PPIs in normal and disease states is crucial to understanding cell physiology.

Before the 20th era, several revolutionary discoveries in biological science were made. The completion of sequencing the human genome, The Human Genome Project was an international scientific research project with the goal of determining the sequence of nucleotide base pairs that make up human DNA, and of identifying and mapping all of the genes of the human genome from both a physical and a functional standpoint certainly belongs to the key tasks successfully completed in the year 2003, represents a milestone in the bioinformatics [1].

Now after the accomplishment of the complete genome also brings along a new, even more challenging task for the researcher: The characterization of the human proteome. Proteomics is the large-scale study of proteins, the main tool for proteome research, is a relatively new and extremely dynamically evolving branch of science, focused on the evaluation of gene expression at proteome level [2]. Current proteomics deals with different issues due to the specific properties of proteins. This field incorporates technologies that can be applied to serum and tissue to extract important biological information in the form of biomarkers to help clinicians and scientists to understand the dynamic biology of their system of interest, such as a cancer patient. Also, the interaction between proteins plays a vital role here.

In this thesis, a brief study on prostate cancer proteomics is done using computational approaches PPIN which helps to find influential protein in cancer research.

1.1 Biological Overview

The life of an organism requires a coordinated function of different organs or tissues. Even in the unicellular system, a series of molecular events, mostly controlled by signal transduction. Signal transduction is a series of phosphorylation events performed by protein kinases. Additionally, messenger molecules coordinate different organs, which are also proteinous in nature.

The central dogma (Figure 1.1.a) of molecular biology describes the two-step process, transcription and translation [3], by which the information in genes flows into proteins:

Deoxyribonucleic acid (DNA) → Ribonucleic acid (RNA) → Protein.

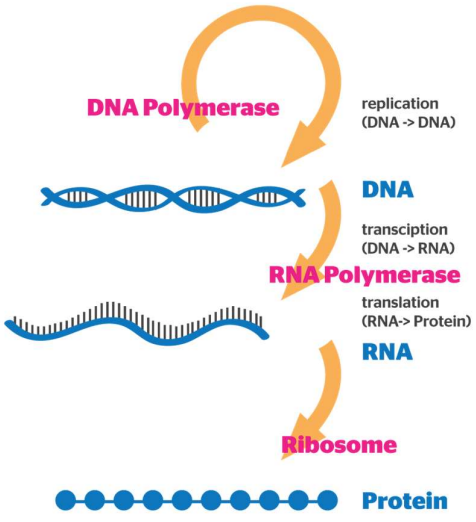

Figure 1.1.a. The Central Dogma,
Source:https://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology

In eukaryote organisms, the DNA is found inside a special area of the cell called the nucleus (where it is called nuclear DNA), but a small amount of DNA can also be found in the mitochondria (where it is called mitochondrial DNA or

mtDNA). The DNA information is stored as a code consisting of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Human DNA consists of about 3 billion bases, and in all people, more than 99 percent of those bases are the same. DNA bases pair with each other to form units called base pairs, A with T and C with G. A sugar molecule and a phosphate molecule are also attached to each base. A base, sugar, and phosphate are called a nucleotide together. Nucleotides are arranged in a spiral called a double helix in two long strands. Because the cell is very small and because organisms have many molecules of DNA per cell, each molecule of DNA has to be packed tightly. This packaged DNA form is known as a chromosome. In 1953, Francis Crick and James Watson described the molecular shape of DNA as a "double helix." Double-stranded DNA is composed of two linear strands that run opposite to each other, known as anti-parallel strands; these strands twist together to form a double helix [4]. The structure of DNA can also be described as a ladder.

Transcription is the process of replicating the information in a DNA strand into a new messenger RNA (mRNA) molecule. DNA safely and stably stores genetic material as a reference or template in the nuclei of cells. Meanwhile, mRNA is comparable to a copy of a reference book because it carries the same information as DNA but is not used for long-term storage and can leave the nucleus freely.

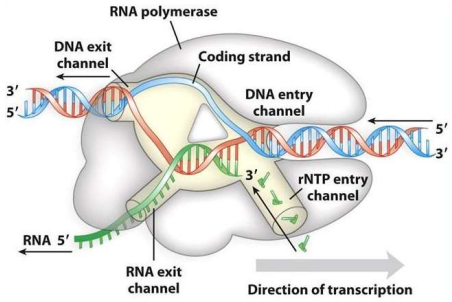

Figure 1.1.b. Transcription Process

Source: https://en.wikipedia.org/wiki/transcription_process

Transcription is performed by an RNA polymerase enzyme and a number of accessory proteins called transcription factors (Figure 1.1.b). In order to recruit RNA polymerase to a suitable transcription site, transcription factors can bind to specific DNA sequences called enhancer and promoter sequences. The

transcription factors and RNA polymerase together form a complex called the initiation of transcription. This complex initiates transcription, and by matching complementary bases to the original DNA strand, the RNA polymerase begins mRNA synthesis. The mRNA molecule is elongated and transcription is terminated once the strand has been fully synthesized.

At this point, the RNA needs to be edited before it can be translated into a protein. This editing process is called splicing, which involves removing non-coding regions called "introns", leaving only, and protein-coding "exons." Splicing begins with the assembly of factors at the intron/exon borders, which act as beacons to guide small proteins to form a splicing machine, called the spliceosome. The gene's newly formed mRNA copies then serve as protein synthesis blueprints during the translation process.

mRNA contains a sequence of amino acids used to make protein chains that read three at a time. Each of the three bases sets is called a codon [5].

A codon is a sequence of three nucleotides of DNA or RNA that corresponds during protein synthesis to a particular amino acid or stop signal. DNA and RNA molecules are written in a four nucleotide language; meanwhile, there are 20 amino

AA	Codons	AA	Codons
Ala	GCT, GCC, GCA, GCG	**Leu**	TTA, TTG, CTT,CTC, CTA, CTG
Arg	CGT, CGC, CGA, CGG, AGA, AGG	**Lys**	AAA, AAG
Asn	AAT, AAC	**Met**	ATG
Asp	GAT, GAC	**Phe**	TTT, TTC
Cys	TGT, TGC	**Pro**	CCT, CCC, CCA, CCG
Gln	CAA, CAG	**Ser**	TCT, TCC, TCA, TCG, AGT, AGC
Glu	GAA, GAG	**Thr**	ACT, ACC, ACA, ACG
Gly	GGT, GGC, GGA, GGG	**Trp**	TGG
His	CAT, CAC	**Tyr**	TAT, TAC
Ile	ATT, ATC, ATA	**Val**	GTT, GTC. GTA, GTG
Start	ATG	**Stop**	TAA, TGA, TAG

Table 1.1.a. Multiple codons for amino acids.

acids in the protein language. Codons provide the key for translating these two languages. Each codon corresponds to a single amino acid (or stop signal), and the

full set of codons is called the genetic code. The genetic code includes 64 possible permutations, or combinations, of three-letter nucleotide sequences that can be made from the four nucleotides. Of the 64 codons, 61 represent amino acids, and three are stop signals. Codon respect to the amino acid is shown in Table 1.1.a.

For example, the amino acid glutamine is represented by the codon CAG, and TAA is a stop codon. The genetic code is described as degenerate, or redundant, as more than one codon can encode a single amino acid. They are read in succession when codons are read from the nucleotide sequence and do not overlap with each other.

The translation is the process of translating the sequence of a messenger RNA (mRNA) molecule to a sequence of amino acids during protein synthesis. The genetic code describes the relationship between the sequence of base pairs in a gene and the corresponding amino acid sequence that it encodes.

1.1.1. Proteins

Proteins are the end products of the process of decoding that begins with cellular DNA information. Proteins compose structural and motor elements in the cell as workhorses of the cell, and they serve as catalysts for virtually every biochemical reaction in living things. This incredible array of functions derives from an incredibly simple code that specifies a wide variety of functions. Protein building blocks are amino acids, small organic molecules consisting of an alpha (central) carbon atom linked to an amino group, a carboxy group, a hydrogen atom, and a variable component called a side chain. Within a protein, peptide bonds connect multiple amino acids, forming a long chain. Peptide bonds are formed by a biochemical reaction that extracts a molecule of water as it joins the amino group of one amino acid to a neighboring amino acid carboxyl group [6]. The linear sequence of a protein's amino acids is considered the protein's primary structure (Figure 1.1.1.a).

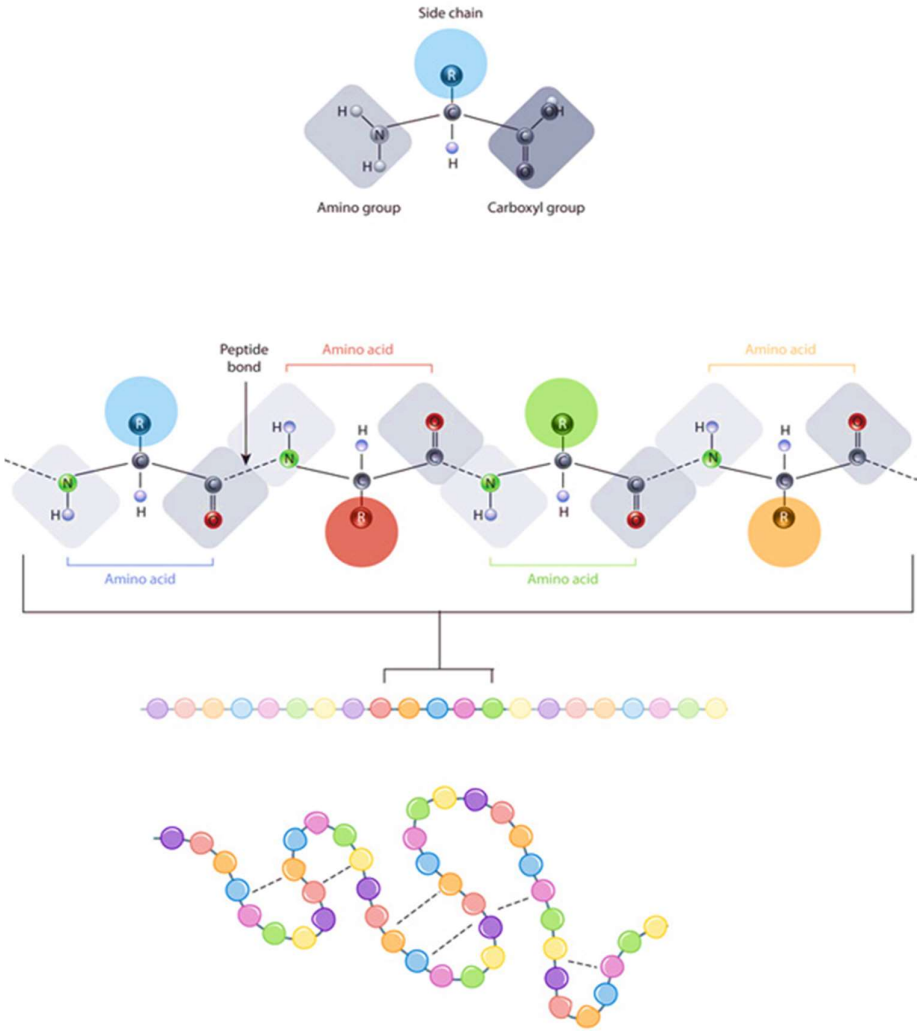

Figure 1.1.1.a. The relationship between amino acid side chains and protein conformation

The defining feature of an amino acid is its side chain (at top, blue circle; below, all colored circles). When connected together by a series of peptide bonds, amino acids form a polypeptide, another word for protein. The polypeptide will then fold into a specific conformation depending on the interactions (dashed lines) between its amino acid side chains.

Proteins are made from just twenty amino acids, each with a unique side chain. The amino acid side chains have various chemistries. Nonpolar side chains are the largest group of amino acids. Several other amino acids have positive or negative side chains, while others have side chains that are polar but uncharged. Amino acid

side chains chemistry is critical to protein structure as these side chains may bind to each other to hold a protein length in some form or conformation. Charged side chains of amino acids may form ionic bonds, and polar amino acids may form bonds of hydrogen. With weak van der Waals interactions, hydrophobic side chains interact with each other. The overwhelming majority of bonds that these side chains form are noncovalent. In fact, cysteines are the only amino acids that can form covalent bonds with their particular side chains. The sequence and location of amino acids in a particular protein guides the bends and folds in that protein due to side chain interactions.

The primary protein structure— its sequence of amino acids— drives the folding and intramolecular bonding of the linear amino acid chain ultimately

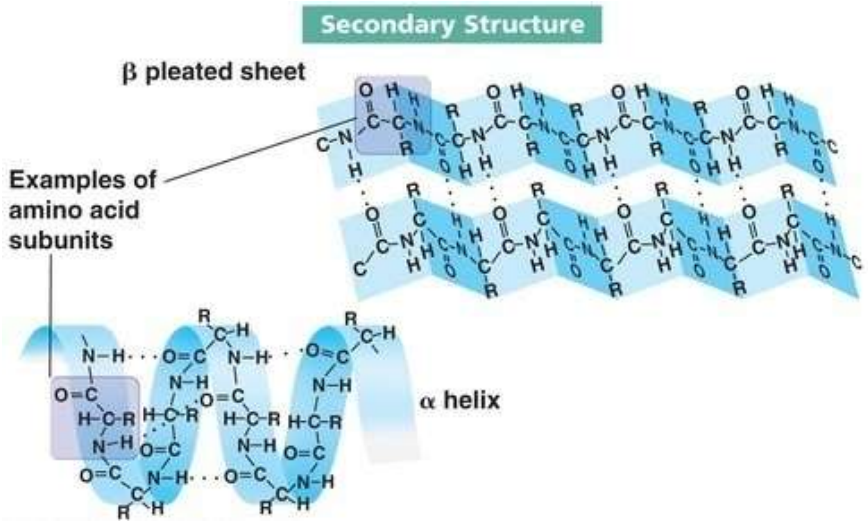

Figure 1.1.1.b. Alpha helix and Beta Sheet.
Source: https://en.wikipedia.org/wiki/alpha_helix

determines the unique three-dimensional shape of the protein. Hydrogen bonding in neighboring regions of the protein chain between amino groups and carboxyl groups sometimes causes some folding patterns to occur. These stable folding patterns, known as alpha helices and beta sheets, constitute the secondary structure (Figure 1.1.1.b) of a protein. In addition to other less common patterns, most proteins contain multiple helices and sheets. In a single linear chain of amino acids — sometimes called a polypeptide— the ensemble of formations and folds constitutes the tertiary structure of a protein. Finally, the quaternary structure of a

protein refers to those macromolecules with multiple polypeptide chains or subunits. Figure 1.1.1.c. shows the four level of Protein Structure.

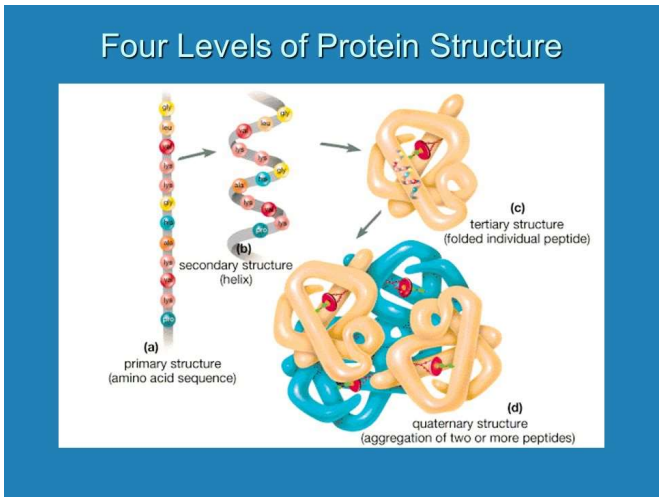

Figure 1.1.1.c. Four Levels of Protein Structure

Typically the most energetically favorable shape adopted by a newly synthesized protein. As proteins fold, before they reach their final form, which is unique and compact, they test a variety of conformations. Thousands of noncovalent amino acid bonds stabilize folded proteins. Furthermore, chemical forces between a protein and its immediate environment contribute to the shape and stability of the protein. For example, the proteins that are dissolved in the cell cytoplasm have hydrophilic (water-loving) chemical groups on their surfaces, whereas their hydrophobic (water-averse) elements tend to be tucked inside. In contrast, the proteins that are inserted into the cell membranes display some hydrophobic chemical groups on their surface, specifically in those regions where the protein surface is exposed to membrane lipids. It is important to note, however, that fully folded proteins are not frozen into shape. Rather, the atoms within these proteins remain capable of making small movements.

Although proteins are considered macromolecules, even with a microscope, they are too small to visualize. So, to figure out what they look like and how they are folded, scientists must use indirect methods. X-ray crystallography is the most common method used for studying protein structures. This method places solid

crystals of purified protein in an X-ray beam and uses the pattern of deflected X-rays to predict the positions of thousands of atoms within the protein crystal. Proteins are constructed as amino acid chains that then fold into unique three-dimensional forms. Bonding within protein molecules helps to stabilize their structure, and the ultimate folded protein forms are well adapted to their functions.

1.1.1.1. Protein Families

To complete their tasks, all proteins bind to other molecules, and a protein's precise function depends on how its exposed surfaces interact with those molecules. Proteins with related forms tend to interact similarly with certain molecules, and therefore these proteins are considered a family of proteins. Within a specific family, the proteins tend to perform similar functions within the cell.

In their primary structure, proteins from the same family also often have long stretches of similar sequences of amino acids. These stretches were preserved through evolution and are vital to the protein's catalytic function. For example, cell receptor proteins at their binding sites contain different amino acid sequences that receive chemical signals from outside the cell, but are more similar in amino acid sequences that interact with common intracellular signaling proteins. Protein families may have many members, and they have probably evolved from duplications of ancient genes. These duplications led to modifications of protein functions and expanded the functional repertoire of organisms over time.

A family of proteins is a group of proteins sharing a common evolutionary origin, reflected in sequence or structure by their related functions and similarities. Protein families are often arranged in hierarchies, with proteins divided into smaller, more closely related groups that share a common ancestor [7]. In this context, it is sometimes used the terms superfamily (describing a large group of distantly related proteins) and subfamily (describing a small group of closely related proteins). A hypothetical protein family hierarchy is illustrated in Figure 1.1.1.1.a.

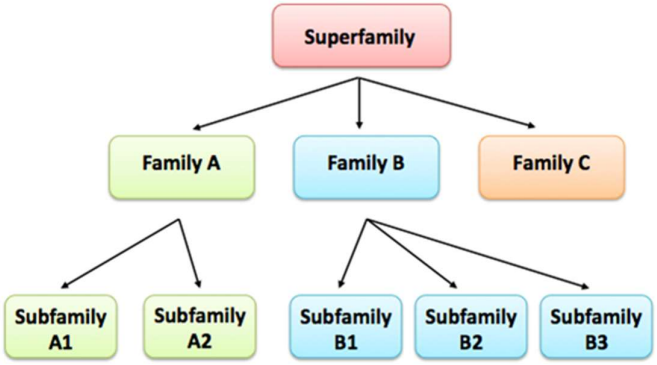

Figure 1.1.1.1.a A hypothetical family hierarchy of proteins that shows the relationships between members of the superfamily, family and subfamily. Directional arrows indicate that one group is another's subgroup.

1.1.2. Proteomics

Proteomics involves the use of technologies to identify and quantify a cell, tissue or organism's overall protein content. It supplements other "omics" technologies such as genomics and transcriptomics to expose an organism's protein identity and to recognize a particular protein's structure and functions. Proteomics-based technologies are used for different research settings in different capacities such as detection of different diagnostic markers, candidates for vaccine production, understanding pathogenicity mechanisms, alteration of patterns of expression in response to different signals and interpretation of functional protein pathways in different diseases [8]. Proteomics is virtually complex because it includes analyzing and categorizing a genome's overall protein signatures. The centerpiece of current proteomics is mass spectrometry with LC–MS-MS and MALDI-TOF / TOF being widely used equipment. However, the use of proteomics facilities including equipment software, databases, and skilled staff requirement substantially increases costs, thus limiting their wider use, particularly in the developing world.

Since the initial stages of biological research, the dynamic role of molecules in supporting life is documented. To demonstrate the importance of these molecules, Berzelius gave the title "protein" from the Greek word "proteios" in 1838, meaning "first rank." The "proteome" can be defined at a particular time as the overall protein content of a cell characterized by its location, interactions, post-translation modifications and turnover. In 1996, Marc Wilkins first used the term "proteomics" to refer to the "PROTein complement of a genome." The proteome characterizes most of the functional gene information. The eukaryotic cell proteome is relatively complex and has a wide range of dynamics. In addition, prokaryotic proteins are responsible for pathogenic mechanisms, but their analysis is challenging due to the enormous diversity of properties such as dynamic quantity range, molecular size, hydrophobicity and hydrophilicity.

Proteomics is critical for early diagnosis, prognosis and disease development monitoring. It also plays a vital role as target molecules in drug development. Proteomics is proteome characterization, including at any stage expression, structure, functions, interactions and protein modifications. In response to external stimuli, the proteome also fluctuates from time to time, from cell to cell. Proteomics in eukaryotic cells is complex due to post-translational modifications, which arise at different sites in numerous ways.

Proteomics is one of the most important methodologies for understanding gene function, although it is much more complex than genomics. Fluctuations in the level of gene expression can be determined to discriminate between two cell biological states by analyzing the transcriptome or proteome. Microarray chips were developed to analyze the entire transcriptome on a large scale. Increasing mRNA synthesis, however, cannot directly measure by microarray. Proteins are biological function effectors and their levels depend not only on the corresponding levels of mRNA, but also on the control and regulation of host translation. A huge volume of proteomics data is gathered with the support of high-throughput technologies (Figure 1.1.2.a).

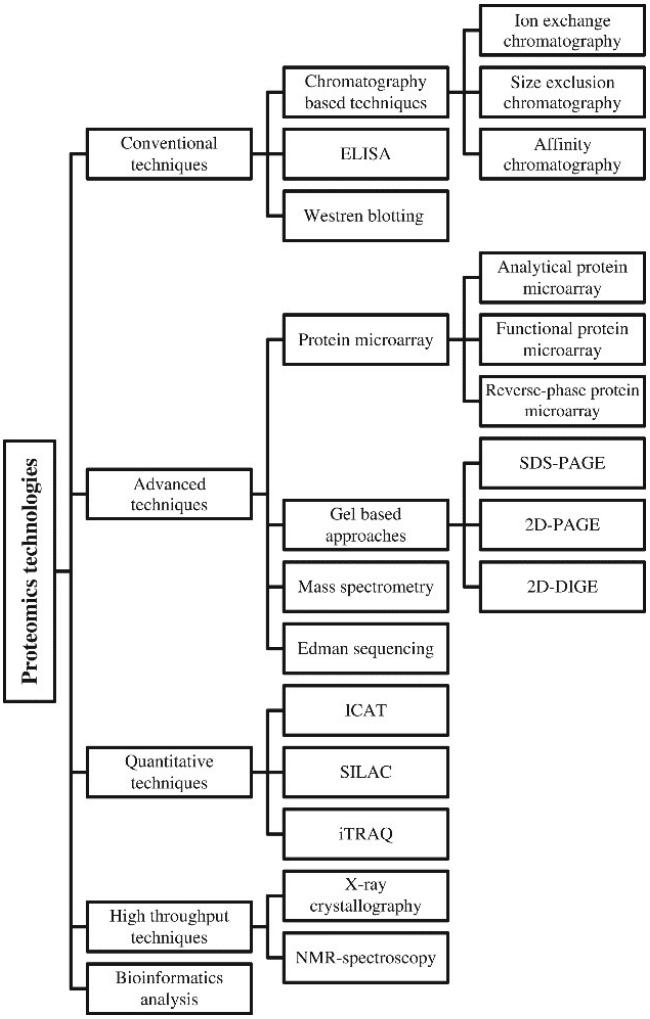

Figure 1.1.2.a an overview of proteomics technologies.

Bioinformatics databases are set up to handle and store huge amounts of data. Different bioinformatics tools are developed for 3D structure prediction, protein domain and motif analysis, rapid protein-protein interaction analysis and MS data analysis. Alignment tools are useful for sequence and structure alignment to detect evolutionary relationships. Proteome analysis provides a complete representation of cell structural and functional information as well as a cell response mechanism to different types of stress and drugs using single or multiple proteomics techniques. Today's major bottlenecks in proteomics research are linked to

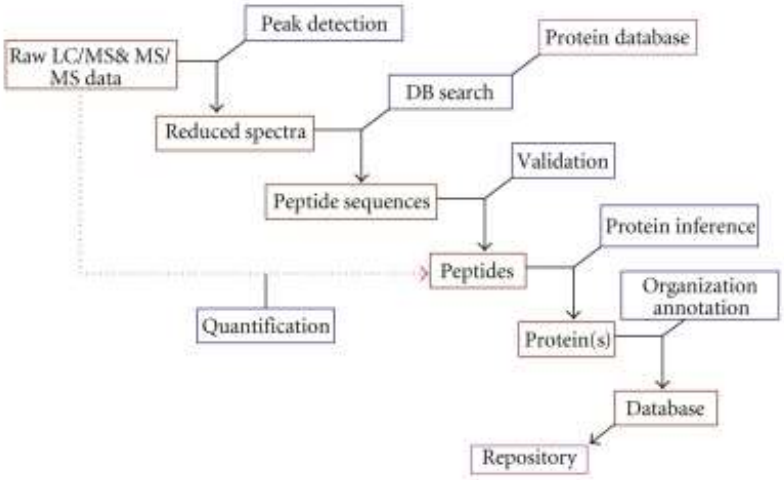

Figure 1.1.2.b. Schematic representation of the different modules constituting a data analysis pipeline.

data analysis in order to create an environment in which computer scientists and biologists and data collectors can work closely together to develop the necessary analytical tools to help interpret the data. Processing and analyzing data from proteomics is a very complex multi-stage process (Figure 1.1.2.b). Because of the lack of standards for data formats, data processing parameters and data quality assessment, the meaningful comparison, sharing, and exchange of data or analysis results obtained on different platforms or by different laboratories remains cumbersome. Precise, consistent and transparent data processing and analysis are integral and critical parts of the workflows of proteomics. It can be generated huge amounts of data now, and there is a huge challenge at the moment to figure out how to actually analyze this data and generate real biological insights. Therefore, the need for an integrated pipeline to process and analyze complex data sets of proteomics has become critical.

1.1.2.1. Proteomics in Early Detection of Cancer

Cancer proteomics involves identifying and quantitatively analyzing differentially expressed proteins relative to healthy tissue counterparts at various stages of the disease, from preneoplasia to neoplasia. Proteomic technologies can also be used to identify cancer diagnostic markers, monitor the progression of the

disease, and identify therapeutic goals [9]. In the discovery of biomarkers, proteomics is valuable because the proteome reflects both the cell's intrinsic genetic program and the impact of its immediate environment. Expression and function of proteins are subject to transcription modulation as well as post-transcription and translation events. Through a differential splicing process, more than one RNA can result from one gene. In addition, there are more than 200 post-translation modifications that proteins could undergo that affect function, the interaction of protein-protein and nuclide-protein, stability, targeting, half-life, and so on, all contributing to a potentially large number of protein products from one gene. During the transformation of a healthy cell into a neoplastic cell, distinct changes occur at the protein level, ranging from altered expression, differential protein modification, and changes in a specific activity to aberrant location, all of which can affect cell function[10][11]. The underlying themes in cancer proteomics are the identification and understanding of these changes. The results include identifying biomarkers that are useful for both early detection and therapy determination.

While proteomics traditionally dealt with quantitative protein expression analysis, more recently, proteomics has been considered to include protein structural analysis. Quantitative proteomics aims to investigate changes in protein expression in various states, such as healthy and diseased tissue or at various stages of the disease. This allows state-and stage-specific proteins to be identified. Structural proteomics tries to uncover protein structure and unravel and map interactions between protein and protein.

1.1.3. Protein-Protein Interactions

From the previous study on molecular biology, we saw that Proteins rarely act alone as their functions tend to be regulated. Many molecular processes within a cell are carried out by molecular machines that are built from a large number of protein components organized by their PPIs [12]. So understanding PPIs is crucial for understanding cell physiology in normal and disease states. PPINs are

mathematical representations of the physical contacts between proteins in the cell. The protein interactome describes the full repertoire of a biological system's PPIs. PPINs are practical means to abstract basic knowledge and to improve biological and biomedical applications. Diseases are often caused by mutations affecting the binding interface or leading to biochemically dysfunctional allosteric changes in proteins. .Therefore, the molecular basis of diseases can be enlightened through protein interaction networks, which in turn can appraise methods for prevention, diagnosis, and treatment [13]. Although our current knowledge of the interactome is both incomplete and noisy because the interaction detection methods have limitations which leads to some false positive and false negative results. Detecting functional modules in PPINs may shed light on cellular functional organization and thereafter underlying cellular mechanisms. In classical approaches of Graph theory there exists many existing module identification algorithms aim to detect densely connected groups in a network. However, based on this simple topological criterion of 'higher than expected connectivity', those algorithms may miss biologically meaningful modules of functional significance, in which proteins have similar interaction patterns to other proteins in networks but may not be densely connected to each other.

There are three fundamental assumptions underlying the identification of disease modules. Firstly, entities forming dense clusters within the interactome (topological modules) are involved in similar biological functions (functional modules). Secondly, molecules associated with the same disease, such as disease-associated proteins, tend to be located in close proximity within the network, which defines the disease module. Thirdly, disease modules and functional modules overlap. Thus, a disease relates to the breakdown of one or more connected functional modules [14].

1.2. Literature survey

In this section, there is a discussion about some required topic such as Biological Challenges and Computational approaches to encounter some issues. Then we go through some relevant topics such as representation and structural

property of PPINs and overview of functional module detection method, Protein databases, mouse and human proteomes, Prostate Cancer and biomarkers.

1.2.1. Proteomics and Cancer

Cancer is a multifaceted disease resulting from deregulated normal cellular signaling networks that control cell behaviors such as proliferation and apoptosis caused by cell or tissue-level genetic, genomic and epigenetic changes.

Cancer mortality does not arise from the lack of available remedies per se, but rather from the diagnosis of such conditions at stages that are too late for remedies to be effective. Prevention, early detection and early intervention are the primary goals of oncologists and cancer biologists. If genes are considered to be the master controllers of cellular behavior, proteins are the effectors and are considered to be the most effective. Specifically, cancer-related proteins expressed direct tumor growth, invasion, metastasis, interaction with surrounding cells, and therapy response. Uncovering the protein signaling network changes, including cell cycle gene network in cancer, aids in understanding the molecular mechanism of carcinogenesis, cancer progression and metastasis and thus identifies the characteristic signaling network signatures unique for different cancers and specific cancer subtypes. Signaling network alterations accumulate at each stage of carcinogenesis that results from genetic, epigenetic and environmental changes and is viewed as a multi-step model of carcinogenesis.

Oncoproteomics is a branch of proteomics that uses proteomic technologies to study proteins and their interactions in a cancer cell. The use of proteomics to foster an improved understanding of cancer pathogenesis, develop new tumor biomarkers for diagnosis, and early detection using a proteomic portrait of samples is of great interest [9]. Oncoproteomics has the potential to revolutionize clinical practice, including cancer diagnosis and proteomic platform screening as a complement to histopathology, individualized selection of therapeutic combinations targeting the entire cancer-specific protein network [15], real-time evaluation of therapeutic efficacy and toxicity, and rational therapy modulation based on changes in cancer.

Most currently available cancer screening tests lack high sensitivity and specificity to be useful in screening the general population, so it is still a clinical challenge to differentiate between some benign and malignant tumors. The advent of oncoproteomics gave hope to discover new biomarkers for use in screening, early diagnosis, and therapy response prediction. Like normal cells, most cancer cells use multiple redundant intracellular signaling pathways to ensure that functions that are critical to their survival are maintained and viable. Thus, potential targets for therapeutic intervention are cellular pathways that are integral to cell function, survival, proliferation, and receptor expression. Based on the proteomic profile of an individual patient, clinicians may recommend combinations of molecularly targeted agents and other therapies [16].

1.2.1.1. Challenges

The field of proteomics has yielded a set of technologies and analytical techniques that are significantly advancing the field of cancer diagnostics. These technologies are found to be an efficient means of identifying new biomarkers for the early detection of cancer and promise hope of new serological screening methods for diagnosis. While proteomics is found to complement genomics-based approaches, providing additional information, it presents various technical, data collection and inference challenges [16]. For example, for amplification of low-abundance proteins, there is no technique equivalent to a polymerase chain reaction, so a detection range from one to several million molecules per cell is needed.

Some technological processes, especially the separation and analysis of proteins, are inherently skill-based and remain difficult to automate [17]. Separation techniques such as capillary electrophoresis may be more automation-friendly, but their superior resolution power is unlikely to replace two-dimensional electrophoresis [18]. Once the proteins are identified, bioinformatics plays an important role in expanding the initial protein information, thus making it a crucial step, in which mishandling of data should be avoided to prevent further mishaps.

With the introduction of national proteomic funding initiatives, proteomics-based approaches should be enabled to realize their potential in biomedical research and translation into clinical practice, making it a powerful tool to fight the disease and maintain health.

1.2.2. Mouse and Human Proteomes

Comparing the proteomes of the mouse and human for several reasons is important. It is likely that the extent to which mouse genetic models recapitulate human disease features will be influenced by the similarity of the protein networks in which that gene works. Overall, mouses and humans share almost the same gene set. So far, nearly every gene found in one species has been found in the other in a closely related form. Less than 10 of the approximately 4,000 genes studies are found in one species but not in the other [19].

There are about 3.1 billion base pairs (or chemical letters) in both the mouse and human genomes. Only about 5% of the sequence consists of regions (genes) that encode proteins. More than 90% of the genome is non-coding DNA that has no known function, sometimes called "junk" DNA. Because of the vast amount of non-coding DNA, simply looking at one sequence alone makes it very difficult to recognize the genes; even the best of today's computational programs fail to identify many coding sequences and misidentify others. The identification of regulatory regions within DNA — the "switches" that turn gene expression on or off, up or down — is similarly difficult as they only exist as poorly defined "consensus" sequences.

The protein-coding regions of the mouse and human genomes are 85% identical on average; some genes are 99% identical while others are only 60% identical. These regions are preserved evolutionarily because the function requires them. The non-coding regions, on the other hand, are much less similar (only 50% or less). Thus, when comparing the same human and mouse DNA region, the functional elements clearly stand out due to their greater similarity. Scientists have

developed computer software that aligns human and mouse sequences automatically, making the protein coding and regulatory regions evident.

Around 80 million years ago, human, mouse and other mammals shared a common ancestor. The genomes of all mammals are therefore comparatively similar. It would be quite informative to compare the dog or cow's DNA sequence with that of the human theoretically. The mouse, however, has a significant advantage in being a well-established experimental model. Not only can genes be found easily in the sequence of the mouse genome, but the function of those genes in the mouse can also be experimentally tested. Thus, scientists can mimic in mice the effect of DNA alterations that occur in human diseases and carefully study the consequences of these DNA misspellings. Mouse models also afford the opportunity to test possible therapeutic agents and evaluate their precise effects.

1.2.3. Protein-Protein Interaction Networks

Protein-protein interactions (PPIs) are essential for nearly every process in a cell, so understanding PPIs is critical to understanding normal cell physiology and disease states. As drugs can affect PPIs, it is also essential in drug development. Interaction networks between proteins (PPIN) are mathematical representations of physical contacts between proteins in the cell. These contacts: are specific; occur between defined protein binding regions; and have a specific biological significance (i.e. they serve a specific function).

To form PPINs, it should be needed to detect protein-protein interactions. Protein-protein interaction detection methods are categorically classified into three types, namely, in vitro, in vivo, and in silico. In in vitro techniques, a given procedure is performed in a controlled environment outside a living organism. In in vivo techniques, a given procedure is performed on the whole living organism itself. In silico techniques are performed on a computer via computer simulation [20]. In Vitro methods: Tandem affinity purification-mass spectroscopy (TAP-MS), Affinity chromatography, Coimmunoprecipitation, Protein microarrays (H), Protein-fragment complementation, Phage display (H), X-ray crystallography,

NMR spectroscopy, etc. In Vivo methods: Yeast 2 hybrid (Y2H) (H), Synthetic lethality. In Silico Methods: Ortholog-based sequence approach, Domain-pairs-based sequence approach, Structure-based approaches, Gene neighborhood, Gene fusion, In silico 2 hybrid (I2H), Phylogenetic tree, Phylogenetic profile, Gene expression-based method [20][22][23]. There exists many PPI databases like BioGrid, Database of Interacting Proteins (DIP), HitPredict, MINT, IntAct, APID, BIND, Biomolecular Object Network Databank (BOND) PINA2.0, etc.

A PPI network is typically represented by an undirected graph G = (V, E) with a set of nodes V and a set of edges E, where V and E represent proteins and interactions between proteins, respectively. The weights on the edges can be used to describe the properties of the PPI network, such as topological or functional features. PPI networks are highly dynamic and structurally complex [24]. They are thus characterized by the inherent properties of complex systems. Additionally, PPI networks manifest the following three topological features: Scale-free distribution, Small-world property, Functional modular network. From the database, we can form two types of network. The first one is binary PPIN and the second one is edge weighted PPIN. In the case of binary PPIN, there is an unweighted network is formed without any weight function [25]. For edge weighted PPIN, there exists a weight function. Gene expression data, Structural distance, Phylogenetic distance can be used for giving weight to edges between two interacting proteins.

With the fast development over the past decade, now there are many different types of approaches for detecting functional modules from PPI networks. Based on computational models, these approaches can be classified into six categories: graph-theoretic approaches, flow simulation-based approaches, spectral clustering-based approaches, supervised clustering, core attachment-based clustering, and swarm intelligent-based approaches, where the graph theoretic approaches can be further classified as density based, hierarchical-based, and partition-based approaches. Table 1.2.3.a illustrates the classification of these detection approaches.

Categories		Main characteristics	Typical algorithms
Graph-theoretic approaches	Density-based	identify densely connected groups of proteins	MCODE[26],MINE[27],CFinder[28] and DPClus[29] LCMA[30],PCP[31],SCAN[32],[33],Ovrlp[34] and PE-WCC[35]
	Hierarchy-based	based on hierarchical nature of modularity	[36], [37], UVCluster[38], Jerarca[39] and [40]
	Partition-based	separate all sparsely connected nodes	RNSC[41], [42] and [43]
Flow simulation-based		simulate a biological or functional flow	TRIBE-MCL[44],[45],STM[46],CASCADE[47],[48] and [49]
Spectral clustering-based		utilize the methodology of matrix analysis	[50], [51], [52] and ADMSC[54][53]
Supervised clustering		use known features to induct clustering	SCI-BN[55]
Core attachment-based		employ the core-attachment protein relations	CORE[56], COACH[57] and [58]
Swarm intelligence-based		exploit swarm intelligence optimized mechanism	ACOPIN[59],[60],NACO-FMD[61] and ACO-MAE[62]

Table 1.2.3.a. The Classification of the Detection Methods of the PPI.

1.2.3.1. Challenges

Even though many practical and effective algorithms have been proposed among a variety of the published approaches, it is a fact that some topological features of the PPI network may hamper the functional module detection. For example, the scale-free distribution of the PPI network can result in the highly uneven size of the clusters; the frequent connections among proteins in different functional groups can generate the blurriness of the modular boundaries. All these problems will worsen the algorithm performance of the functional module detection in PPI networks. Along with the growth of the practical needs of bioinformatics in the postgenomic era, the following problems and challenges are emerging:

1. The unreliability of the interaction data: The PPI data gained from the biological experiments are both incomplete and noisy. The reason is two folded: On the one hand, the false-positive ratio of the high throughput PPI data is much higher compared with the small-scale data; on the other hand, limited by the experiments performed, some practical existing interaction data in the data banks may be lost (false negative). Therefore, to get a better mining result, the detection algorithms should improve their robustness. Even though some effective approaches based on various information fusions now are available, how to reduce the negative effects brought by the noisy data on the detection quality is still an important problem in PPI functional module detection.

2. The efficiency of detection algorithms: A PPI network is usually composed of thousands of proteins and even more interactions. It is a large-scale complex

network. So, a reasonable time complexity is a practical requirement for the detection approaches. However, most detection algorithms based on the computational approaches have high time complexity, which will limit their developments and practical applications. Therefore, researchers should pay more attention to this problem.

3. The overlapping of functional modules: Many existing clustering approaches have difficulties for analyzing the PPI data mainly due to the fact that a protein can have several different functions. Namely, a protein may be involved in one or more functional modules. Therefore, overlapping clusters need to be accurately detected in PPI networks. In spite of some specific strategies have been proposed over recent years, this problem is still a challenging topic.

1.2.4. Prostate Cancer and Biomarkers

The term "biomarker," a portmanteau of "biological marker," refers to a broad subcategory of medical signs that can be accurately and reproducibly measured– that is, objective indications of medical condition observed from outside the patient. In contrast to medical symptoms, medical signs are limited to indications of health or disease perceived by patients themselves. In the literature, there are several more accurate definitions of biomarkers and they are fortunately significantly overlapping. In 1998, a biomarker was defined by the National Institutes of Health Biomarkers Definitions Working Group as "a characteristic objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to therapeutic intervention." [63]. A joint venture on chemical safety, the International Programme on Chemical Safety, led by the World Health Organization (WHO) and in coordination with the United Nations and the International Labor Organization, has defined a biomarker as “any substance, structure, or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease” [64].

An even broader definition takes into account not only the incidence and outcome of disease, but also the effects of treatments, interventions, and even unintended exposure to the environment, such as chemicals or nutrients. The WHO stated in its report on the validity of biomarkers in environmental risk assessment that a true definition of biomarkers includes "nearly any measurement that reflects an interaction between a biological system and a potential hazard that may be chemical, physical or biological. The measured response may be functional and physiological, biochemical at the cellular level, or a molecular interaction."[65]. Biomarker examples include everything from pulse and blood pressure through basic chemistries to more complex blood and other tissue laboratory tests. Medical signs have a long history of use in clinical practice— as old as medical practice itself— and biomarkers are just the most objective, quantifiable medical signs that modern laboratory science allows us to reproducibly measure. The use of biomarkers is somewhat newer in clinical research, and the best approaches to this practice are still being developed and refined. The key issue at hand is determining the relationship with relevant clinical endpoints between any given measurable biomarker.

Prostate cancer (PCa) is the second most common cancer in men worldwide, with an estimated incidence of 1.1 million new cases in 2012. The highest incidence rate in Western countries was observed, at least partially due to the widespread use of prostate-specific antigen (PSA) testing. Thus, in the United States, PCa is the most common tumor with 1, 64,690 new estimated cases for 2018. Actually, a large proportion of PCa is latent, never intended to progress or affect the life of patients. Klotz3 has estimated that between 50 percent and 60 percent of newly diagnosed cases are the percentage of patients with low risk of progression [66].

1.3. Scope of work

- We saw that Proteins rarely act alone as their functions tend to be regulated. Many molecular processes within a cell are carried out by molecular machines that are built from a large number of protein components organized by their Protein-protein interactions (PPIs). Now it is obvious for a disease, normally a set of protein interaction is responsible. So, it is possible to find the interacting proteins with some disease-related biomarkers previously found by some biological experiments.
- By using protein family it is easy to know their interacting nature with other proteins.
- A candidate protein that has the greatest influence on that module can easily be found in the interaction modules for the biomarkers. This can lead to new biomarkers in further research since a particular disease is caused by the functional module.
- Mouse and people have the same ancestor, so some evolutionary links to cancer can be found.
- It is easy to find the most influential proteins in a functional module for prostate cancer or some other types of cancer.

1.4. Motivation

Once the human genome has been completed, it is a very difficult task to characterize the protein encoded through the sequence. If new therapeutic medicines and new biomarkers for early diagnosis are to be developed, the study of the whole protein complement of the genome "proteome," known as proteomics, is essential.

With the emergence of genomic profiling technologies and selective molecular targeted therapies, biomarkers play an increasingly important role in the clinical management of cancer patients. Single gene/protein or multi-gene “signature”-based assays have been introduced to measure specific molecular pathway deregulations that guide therapeutic decision-making as predictive biomarkers. Genome-based prognostic biomarkers are also available for several cancer types for potential incorporation into clinical prognostic staging systems or practice guidelines. However, there is still a large gap between initial biomarker discovery studies and their clinical translation due to the challenges in the process of cancer biomarker development.

Because of the large number of protein interactions, it is almost impossible to study each interaction as a biological experiment. So, some computational simulation-based approach is needed to study the nature of interactions. There are many computational methods for studying protein interactions and finding some influential protein. There is a new computational approach in this thesis to find most influential proteins in the proteomics of prostate cancer using previously valid biomarkers and weighted protein-protein interaction network module as well as some evolutionary link analysis between mouse and human.

1.5 Organization of the thesis

In Chapter 1, A basic introduction to the work done in this thesis will be presented along with the medical obligations for this type of work. The related biology is explained in the beginning, followed by the clinical concerns that led to the importance of such work. The chapter ends with a brief survey of the literature, the scope of work and motivation for this thesis.

In Chapter 2, Initial Biomarkers Selection process is discussed from the clinically proven protein set for prostate cancer. A brief literature survey is required for initial biomarkers selection, is a vital step in this present work.

In Chapter 3, there is a discussion on required Databases and Bio-informatics theories. Some sequence similarity measurement methods have also been discussed which plays a great role for thesis goal.

In Chapter 4, Basics of network centrality are discussed. Eigenvector centrality method, method of finding influential nodes from PPIN for our current work is also discussed.

In Chapter 5, a methodology of fulfilling the overall aims of this study has been discussed.

In Chapter 6, Biological significance analysis of the final outcome of this work is done and discussed.

In Chapter 7, an overall summary of the work related to its advantage, shortcomings and future scopes is also discussed and concluded.

Chapter 2

Initial Biomarkers Selection

A biomarker selection should have a biological or therapeutic basis, or at least the biomarker should indicate a reliable correlation with the cancer's presence, characteristics, or aggressiveness. There should also be an evaluation of the strength of the marker in relation to the outcome of the disease, which, together with other factors, should be performed as an independent predictor in a multivariable assay. An ideal biomarker should be quick, consistent, economical and quantifiable in an accessible biological fluid or clinical sample (e.g. plasma, urine or prostatic fluid) readily interpretable by a clinician. Its expression in the related disease condition should be significantly increased (or decreased) and there should be no overlap between healthy control subjects and untreated patients in the biomarker levels.

Biomarkers are essential factors for clinical and biological research. The identification of a new candidate biomarker is followed by a thorough operational evaluation to validate its application in the clinical setting. Biomarkers that have been scientifically scrutinized must pass several proposed practical tests before being accepted for clinical practice. Five conceptual phases of biomarker development were suggested: I preclinical exploratory, (ii) clinical testing and validation, (iii) longitudinal retrospective, (iv) prospective screening, and (v) cancer control [67].

For this work, for further computational simulation, there is a requirement to select some valid protein biomarkers. Through a brief literature survey, we find some clinically proven valid protein biomarkers.

2.1 Ferritin light chain

Ferritin light chain is a protein encoded by the FTL gene in humans [68] [69] [70]. It is abnormally expressed in IVF and ICSI fetuses, which can contribute to the increased risk of birth defects in these ARTs [71]. This gene encodes the ferritin protein's light subunit. Ferritin is the major protein stored in prokaryotes and eukaryotes in intracellular iron. It is made up of 24 heavy and light ferritin chains subunits. Variation in the composition of the ferritin subunit may affect the iron uptake and release rates in various tissues. The storage of iron in a soluble and non-toxic state is a major function of ferritin. There are multiple pseudogenes in this gene [70].

Immunohistochemical tissue characterization was performed in PCa and BPH patients. Qiang Su et al. found representative immunohistochemical expression of FTL with high intensity of staining in PCa and low intensity of staining in BPH. They conclude that Ferritin is a potential urinary biomarker to discriminate between patients with PCa and BPH [72]. They detected FTL expression in the tissues of patients with PCa and BPH for immunohistochemistry. Both FTL showed high stain intensity (++) in patients with PCa and low stain intensity (+) in patients with BPH.

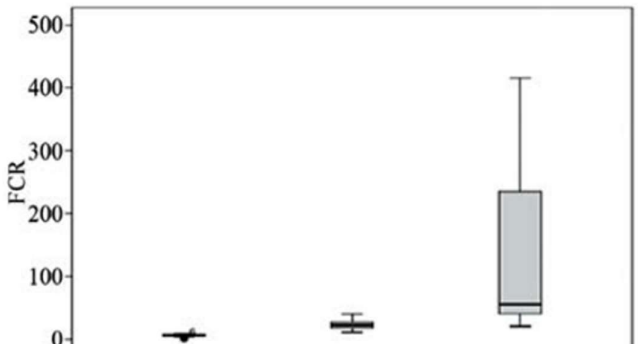

Figure 2.1.a. Box chart of urinary FCR in three groups. FCR was evaluated by ferritin-creatinine ratio (FCR) and was significantly highly expressed in PCa group compared to the BPH) and control groups. There was no significant difference between BPH and normal controls. [72]

Figure 2.1.a provides a box chart of the FCR among the three groups. In BPH and Normal Control, the FCR is not dissimilar, but there are drastic changes for PCa. So, Ferritin light chain plays a very important role for PCa and our subject of interest.

2.2. Mitochondrial 60 kDa heat shock protein

This gene encodes a chaperonin family member. The encoded mitochondrial protein in the innate immune system may function as a signaling molecule. This protein is essential in the mitochondria to fold and assemble newly imported proteins. This gene works as a bidirectional promoter adjacent to a related family member and the region between the 2 genes. This gene has been associated with several pseudogenes. For this gene, two transcript variants encoding the same protein were identified [76].

HSP overexpression signals a poor prognosis in terms of survival and response to therapy in specific cancer types. Elevated HSP expression in malignant cells plays a key role in protection from spontaneous apoptosis associated with malignancy as well as the apoptosis generated by therapy, mechanisms which may underlie the role of Hsp in tumor progression and resistance to treatment [75].

Heat Shock Proteins (HSPs), a family of genes with key roles in proteostasis, have been extensively associated with cancer behavior. Proteomic comparison of prostate cancer cell lines LNCaP-FGC and LNCaP-r reveals 60 kDa heat shock protein as a marker for prostate malignancy by Björn Johansson et al. In non-malignant prostate, HSP60-staining was in the glandular compartment, particularly basal epithelial cells. In prostate cancer, most epithelial cells showed moderate-strong staining without apparent correlation between staining intensity and Gleason grade. For PCa, 60 kDa heat shock protein has an important role. So, we select the 60kDa Heat Shock Protein for further analysis [74].

2.3 Protein disulfide-isomerase

The protein disulfide isomerase (PDI) family is a group of multifunctional endoplasmic reticulum (ER) enzymes that mediate the formation of disulfide bonds, catalyze the cysteine-based redox reactions and assist the quality control of client proteins. Recent structural and functional studies have demonstrated that PDI members not only play an essential role in the proteostasis in the ER but also exert diverse effects in numerous human disorders including cancer and neurodegenerative diseases. Increasing evidence suggests that PDI is actively involved in the proliferation, survival, and metastasis of several types of cancer cells. Although the molecular mechanism by which PDI contributes to tumorigenesis and metastasis remains to be understood, PDI is now emerging as a new therapeutic target for cancer treatment. In fact, several attempts have been made to develop PDI inhibitors as anti-cancer drugs [77].

By exploring microarray and proteomic data, Shili and et al. reported that PDI expression was significantly upregulated in brain and CNS cancers, lymphoma, kidney, ovarian, prostate, lung and male germ cell tumors, and that this upregulation correlates with cancer metastasis and invasion [78]. Over the past decades, structure and domain architecture, biochemical redox reactions, physiological roles, and PDI involvement in multiple diseases have been extensively studied. Dysregulation of PDI gene expression, post-translational modification, or enzymatic activity has resulted in various human diseases. However, the relationship between PDI and cancer has only recently been documented.

Gene expression microarray studies provide an important tool for assessing PDI expression levels in different cancer types. D. Singh et al. 2002. Gene expression correlates of clinical prostate cancer behavior [79]. Walsh also made other significant contributions towards a better understanding of hereditary aspects, pathogenesis and genetic susceptibility to prostate cancer [80]. He showed the value of serial prostate-specific antigen measurement as a means for improving

the diagnosis of prostate cancer and predicting its outcome. By analyzing published microarray data sets, In Figure 2.3.a, there is a comparison between PDI expressions in prostate cancer type with that in normal tissue.

Shili et al. also conclude that PDI is highly expressed in select cancer types, supports tumor growth and is associated with clinical outcomes. Therefore, PDI is a potential drug target for cancer therapy and also our subject of interest.

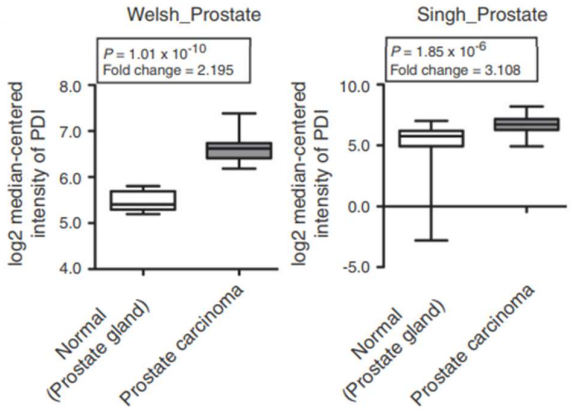

Figure 2.3.a. Protein disulfide isomerase (PDI) is highly expressed in multiple cancer types compared with respective normal tissues.

Cancer types analyzed: prostate Data sets were obtained from OncomineTM (http://www.oncomine.com) with filtering thresholds as P < 105 and fold change >2, and analyzed using Prism 5 (GraphPad Software, Inc). The Student t-test was used for statistical analysis to compare gene expression levels between normal and cancer tissues.

2.4. Sialic acid synthase

Sialic acid synthase is an enzyme that in humans is encoded by the NANS gene [81] [82]. This gene encodes an enzyme that functions in the biosynthetic pathways of sialic acids. In vitro, the encoded protein uses N-acetylmannosamine 6-phosphate and mannose 6-phosphate as substrates to generate phosphorylated forms of N-acetylneuraminic acid (Neu5Ac) and 2-keto-3-deoxy-D-glycero-D-

galacto-nononic acid (KDN), respectively. However, it exhibits much higher activity toward the Neu5Ac phosphate product.

Cong Zhang et al. were retrospectively collected and analyzed the data from 540 patients who were newly diagnosed with PCa or BPH between November 2014 and March 2018 [83]. Pretreatment SA levels were compared across various

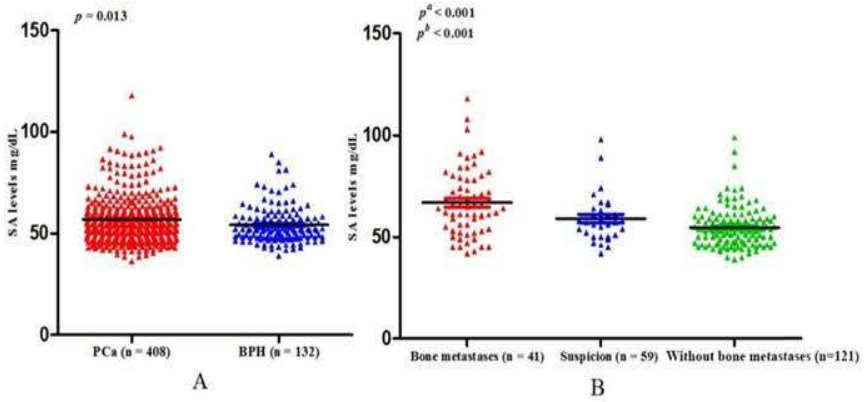

Figure 2.4.a. A) Comparison of SA levels between PCa and BPH patients; (B) Comparison of SA levels among bone metastases, suspicious bone metastases, and without bone metastases in PCa patients (pa: Without bone metastases versus suspicion versus bone metastases; pb: Without bone metastases versus bone metastases).[83]

groups. Also, the associations between SA levels and clinic parameters of patients were analyzed as well (Figure 2.4.a). Univariate and multivariate logistic regression analyses were further used to identify independent associations. As a result Cong Zhang et al. seen that the mean SA levels in patients with PCa were significantly higher than with BPH. Also conclude that Elevated SA level is an independent predictor of prostate cancer as well as its bone metastases. Therefore, the SA level may be a promising diagnostic and prognostic biomarker for prostate cancer and bone metastases.

2.5 Annexin A2

Annexin A2 is a 36-kDa protein interfering with multiple cellular processes, especially in cancer progression. Annexin 2 is involved in diverse cellular processes such as cell motility (especially that of the epithelial cells), linkage of membrane-associated protein complexes to the actin cytoskeleton, endocytosis, fibrinolysis, ion channel formation, and cell-matrix interactions. It is a calcium-dependent phospholipid-binding protein whose function is to help organize exocytosis of intracellular proteins to the extracellular domain [84]. Annexin A2 is a pleiotropic protein meaning that its function is dependent on place and time in the body.

The protein Annexin A2 has been investigated as a prognostic marker because of its widespread presentation in several cancer forms. The protein is strongly expressed in normal prostatic epithelial glands. Although commonly underexpressed in prostate cancer, the association of reduced expression with pathological grade and the stage is unknown [85].

The use of proteomic analysis reveals reduced or lost expressions of annexins a1 and a2 in prostate cancer cells. Shogo Senga et al. observation concurs with other data showing reduced annexins a1 and a2 expression in prostate cancer cells *in vivo [87]*. Chetcuti *et al.* recently reported that 100% of prostate cancer specimens (31 cases in total) examined lacked the 36 kDa annexin a2 immunostaining and these cancer tissues also lacked annexin a2 mRNA expression. Southern analysis of cancer DNA, however, did not reveal any noticeable deletions/mutations in the annexin a2 gene suggesting that the loss of annexin a2 protein expression results from transcriptional or post-transcriptional suppression [86].

Prostate cancer cells showed reduced levels as well as altered expression patterns of Annexin A2.So, Annexin A2 is taken for the further computational operation.

2.6 Fatty acid-binding protein 5

Epidermal or cutaneous fatty acid-binding protein is an intracellular lipid-binding protein, also known as FABP5, and its expression level is closely related to cancer cell proliferation and metastatic activities in various types of carcinoma. However, the molecular mechanisms of FABP5 in cancer cell proliferation and its other functions have remained unclear. Shogo Senga and his team clearly revealed that FABP5 activated expression of metabolic genes (ATP5B, LCHAD, ACO2, FH, and MFN2) via a novel signaling pathway in an ERRα (estrogen-related receptor α)-dependent manner in prostate cancer cell lines [87].

This gene encodes the fatty acid binding protein found in epidermal cells and was first identified as upregulated in psoriasis tissue. Fatty acid binding proteins are a family of small, highly preserved cytoplasmic proteins that bind long-chain fatty acids and other hydrophobic ligands. It is thought that FABPs roles include fatty acid uptake, transport, and metabolism.

Five of prostate carcinomas were investigated by Angelika Tölle and her team. In prostate cancer cell lines, a strong reduction of FABP 4 and FABP 5 mRNA was observed [88].

2.7 Protein S100-A11

The S100 gene family is the largest subfamily of calcium-binding proteins of EF-hand type [89]. To date, at least 25 distinct members of this subgroup have been described. Of these genes, 22 are clustered at chromosome locus 1q21. Interestingly, 14 of 22 members localized in the epidermal differentiation complex (EDC) on chromosome 1q21 [90] [91]. S100 proteins form either homodimeric or heterodimeric complexes with one another [92]. Upon calcium binding, most S100 proteins undergo a conformational change, thus allowing the protein to interact with the different protein targets, thereby exerting a broad range of intracellular

and extracellular functions. Intracellular functions include regulation of calcium homeostasis, cell cycle, cell growth and migration, phosphorylation, cytoskeletal components and regulation of transcriptional factors. In contrast to intracellular function, extracellular S100 proteins act in a cytokine-like manner by binding to cell surface receptors such as the receptor for advanced glycation end products (RAGE) and Toll-like receptors (TLRs) [90] [93].

More recently, there is a growing interest in the S100 proteins and their relationship with different cancers because of their involvement in a variety of biological events which are closely related to tumorigenesis and cancer progression. The association between S100 proteins and cancer can also be explained by several observations: firstly, most of S100 genes are clustered on human chromosome 1q21, a region prone to genomic rearrangements, supporting that S100 proteins may be implicated in tumor progression. Secondly, several S100 members show altered expression in various malignancies. Finally, a number of S100 proteins have been shown to interact with and to regulate various proteins involved in cancer and exert different effects on specific target proteins such as NF-κB, p53, and β-catenin. In a review article, Hongyan Chen and Chengshan Xu discuss the important roles of S100 proteins in tumorigenesis, cancer metastasis, tumor microenvironment, maintenance of pluripotency and their potential implications as biomarkers and prognostic factors [94].

Finally, seven biomarkers are selected for further research interest. (Table 2.7.a)

Table 2.7.a. Selected protein biomarkers and their Uniprot ID

Uniprot ID	Protein Name
P02792	Ferritin light chain
P10809	Mitochondrial 60 kDa heat shock protein
P30101	Protein disulfide-isomerase
Q9NR45	Sialic acid synthase
P07355	Annexin A2
Q01469	Fatty acid-binding protein 5
P31949	S100-A11

Chapter 3

Databases and Bio-informatics theories

To achieve the goal of this thesis, we need some databases for some vital jobs like finding protein family, finding protein similarity, sorting family-specific proteins, finding species-specific proteins, and forming networks of protein-protein interaction. Some bioinformatics concept such as FASTA, Gap penalty scoring method is also needed for protein similarity finding. This chapter discusses the database used and also briefly discusses the FASTA and Gap penalty scoring method.

3.1. Protein Database Used

In order to support protein-related information management, generation of data-driven hypothesis and discovery of biological knowledge, many publicly available data repositories and resources have been developed. In recent decades, the use of high-throughput technologies to study molecular biology systems has revolutionized biological and biomedical research, enabling researchers to systematically study organism genomes (Genomics), the set of RNA molecules (Transcriptomics), and the set of proteins including their structures and functions (Proteomics).

The richness of proteomics data enables researchers to ask complex biological questions and gain new insights into science. To support the generation of data-driven hypotheses and the discovery of biological knowledge, many protein-related bioinformatics databases, query facilities and software tools for data analysis have been developed to organize and provide biological protein annotations to support sequence, structural, functional and evolutionary analyzes in the context of the biology of pathways, networks and systems. A discussion of all necessary databases is available in this section.

3.1.1. Human Protein Reference Database (HPRD)

The HPRD is the result of an international collaborative effort between the Bioinformatics Institute in Bangalore, India and the Pandey laboratory at Johns Hopkins University in Baltimore, USA. HPRD contains manually curated scientific information on the biology of most human proteins. Information on human disease proteins is annotated and linked to the database of Online Mendelian Inheritance in Man (OMIM). The National Information Center for Biotechnology provides links to HPRD through its databases on human proteins [95].

This resource depicts information on human protein functions including protein-protein interactions, post-translational modifications, enzyme-substrate relationships and disease associations. Protein annotation information that is catalogued was derived through manual curation using published literature by expert biologists and through bioinformatics analyses of the protein sequence. The protein-protein interaction and subcellular localization data from HPRD have been used to develop a human protein interaction network.

3.1.2. Pfam

Pfam is a database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models [96].

The general purpose of the Pfam database is to provide a complete and accurate classification of protein families and domains. Originally, the rationale behind creating the database was to have a semi-automated method of curating information on known protein families to improve the efficiency of annotating genomes. The Pfam classification of protein families has been widely adopted by

biologists because of its wide coverage of proteins and sensible naming conventions. For each family in Pfam one can:

1. View a description of the family.
2. Look at multiple alignments.
3. View protein domain architectures.
4. Examine species distribution.
5. Follow links to other databases.
6. View known protein structures.

3.1.3. STRING

STRING is a database of known and predicted protein-protein interactions. The interactions include direct (physical) and indirect (functional) associations; they stem from computational prediction, from knowledge transfer between organisms, and from interactions aggregated from other (primary) databases [97]. The STRING database currently covers 9,643,763 proteins from 2,031 organisms. Interactions in STRING are derived from five main sources:

1. Genomic Context Predictions
2. High-throughput Lab Experiments
3. Co-Expression (Conserved)
4. Automated Text-mining
5. Previous Knowledge in Databases.

3.1.4 Universal Protein Resource (UniProt)

The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. The UniProt databases are the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc). European Bioinformatics Institute and Swiss Institute of Bioinformatics together used to produce Swiss-Prot and TrEMBL, while PIR produced the Protein Sequence Database (PIR-PSD). These two data sets coexisted with different protein sequence coverage and annotation priorities. TrEMBL (Translated EMBL Nucleotide Sequence Data Library) was originally created because sequence data was being generated at a pace that exceeded Swiss-Prot's ability to keep up. Meanwhile, PIR maintained the PIR-PSD and related databases, including iProClass, a database of protein sequences and curated families. In 2002 the three institutes decided to pool their resources and expertise and formed the UniProt consortium [98].

3.1.5 A Simple Modular Architecture Research Tool (SMART)

SMART allows the identification and annotation of genetically mobile domains and the analysis of domain architectures. More than 500 domain families found in signaling, extracellular and chromatin-associated proteins are detectable. These domains are extensively annotated with respect to phyletic distributions, functional class, tertiary structures and functionally important residues. Each domain found in a non-redundant protein database, as well as search parameters and taxonomic information, are stored in a relational database system. User interfaces to this database allow searches for proteins containing specific combinations of domains in defined taxa [99].

3.1.6 InterPro

InterPro is a database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to new protein

sequences in order to functionally characterize them. The contents of InterPro consist of diagnostic signatures and the proteins that they significantly match. The signatures consist of models (simple types, such as regular expressions or more complex ones, such as Hidden Markov models) which describe protein families, domains or sites. Models are built from the amino acid sequences of known families or domains and they are subsequently used to search unknown sequences (such as those arising from novel genome sequencing) in order to classify them [100]. Each of the member databases of InterPro contributes towards a different niche, from very high-level, structure-based classifications (SUPERFAMILY and CATH-Gene3D) through to quite specific sub-family classifications (PRINTS and PANTHER).

3.2 Fast Adaptive Shrinkage Threshold Algorithm (FASTA)

The FASTA algorithm is a heuristic method for string comparison. Lipman and Pearson developed it in 1985 and improved further in 1988 [101] [102]. FASTA is a pair sequence alignment tool that takes input as nucleotide or protein sequences and compares them to existing databases. It is a text-based format that can be read and written using a text editor or word processor. A

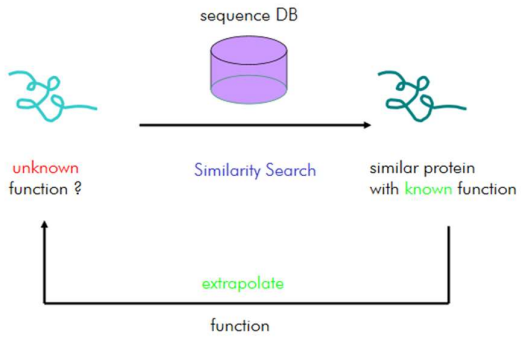

Figure 3.2.a. Flow diagram Similarity Searches on Sequence Database.

sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column. For DNA and proteins, it is represented in one letter IUPAC nucleotide codes and amino acid codes. It finds the local similarity between the sequences and calculates the statistical significance of matches (Figure 3.2.a). It can be also used to find the functional and evolutionary relationship between the sequences [103].

FASTA program uses word hits to identify potential matches before attempting the optimized search that takes more time. The speed and sensitivity are controlled by the ktup parameter, which specifies the word size. The number of background hits decreases by increasing the ktup. It initially checks for the segment that includes several nearby hits. This program is much more sensitive than BLAST programs, which is reflected in the length of time required to produce results. FASTA produces local alignment scores to compare the query sequence with each sequence in the database.

3.2.1 Working procedure of FASTA algorithm

- Nucleotide or protein sequence is taken as input.

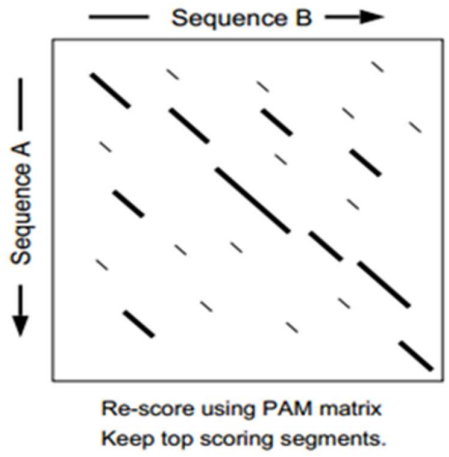

Figure 3.2.1.a. Diagonal line representation in dotplot.

- The speed and sensitivity are controlled by the parameter called ktup, which specifies the size of the word. This program uses the word hits to identify potential matches between the query sequence and database sequence. Lesser the ktup value, more sensitive the search. By default, ktup is 2 for proteins and ktup is 4 or 6 for nucleotides, initially, it checks for the segment's containing several nearby hits.
- Then it finds similar local regions based on matches and mismatches (scoring) and isolates the highest matches from the background hits. The scoring matrix used is BLOSUM50 for the sequence of proteins and the nucleotide sequence identity matrix [104]. In dotplot between two sequences, local regions are represented as diagonal lines (Figure 3.2.1.a).
- It finds and saves the best local regions.
- The local regions are rescanned and scored with a suitable scoring matrix.
- Take the subregions with the maximum score from the local regions (Figure 3.2.1.b). From that, the highest score of the subregion will be referred as

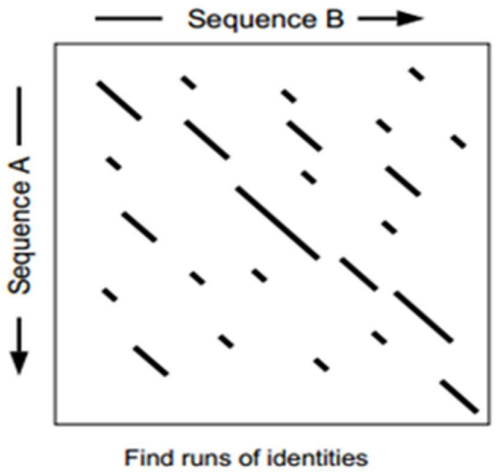

Figure 3.2.1.b. Rescoring using PAM matrix to keep high scoring segment.

init1.

- Subsequences (subregions) are searched through the library sequences to determine the similarity. From these sequences which are having less than the cutoff value will be eliminated (Figure 3.2.1.c).
- Checks whether gaps are required to fill the sequence similarity search. An initial similarity score is used to rank the library sequence (initn).
- It uses the Smith-Waterman algorithm to calculate an optimal score for whole alignment (Figure 3.2.1.d).

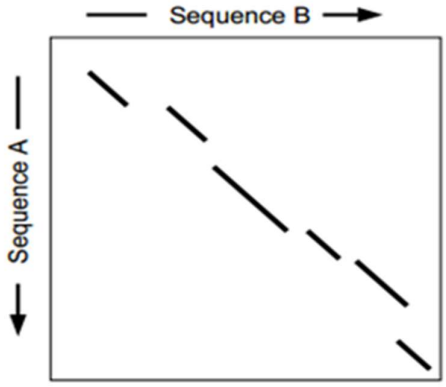

Figure 3.2.1.c. Using threshold elimination of sequences unlikely to be a part of the alignment that includes highest scoring segment.

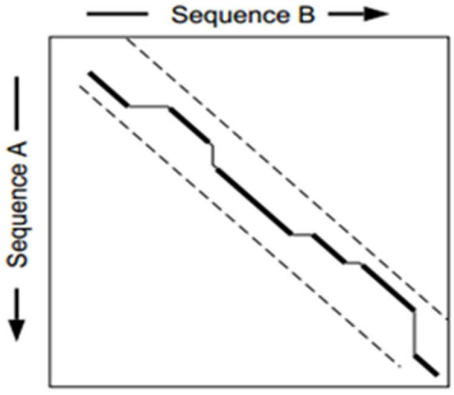

Figure 3.2.1.d Using Smith-Waterman algorithm, optimization of the alignment in a narrow band that encompresses the top scoring segment.

3.2.2. Advantage of FASTA over traditional Dynamic Programming algorithms

With the Dynamic Programming algorithm, one gets an alignment in a time that is proportional to the product of the lengths of the two sequences being compared. Therefore, the computation time increases linearly with the size of the database when searching for a whole database. Current databases are too slow to calculate a complete dynamic programming alignment for each database sequence (unless implemented in specialized parallel hardware). Therefore, the number of searches currently carried out on entire genomes requires faster procedures. By prune the search space by using fast approximate methods to select the sequences of the database, FASTA becomes 50-100 times faster than traditional Dynamic Programming algorithms.

3.3 GAP

Gap considers all possible alignments and gap positions between two sequences and creates a global alignment that maximizes the number of matched residues and minimizes the number and size of gaps [105]. A scoring matrix is used to assign values for symbol matches. In addition, a gap creation penalty *and a* gap extension penalty are required to limit the insertion of gaps into the alignment. Gap uses the alignment method of Needleman and Wunsch [106].

Figure 3.3.a. Sequence after inserting a Gap.

Proteins often contain regions where residues have been inserted or deleted during evolution. In Figure 3.3.a, clearly shows that it can be improved by inserting a

gap. Two alignments with identical number of gaps but very different gap distribution provide a different score. It should be preferable one large gap to several small ones because gap starting penalty is very high.

Figure 3.3.b. Comparison gap distribution of a sequence.

With a match score of 1 and a mismatch score of 0 and a gap opening penalty of 10 and extension penalty of 1, the sequence similarity is scored 2 types of gap distribution in Figure 3.3.b. For the first type, score is -3 and the second type scored -43. The raw score is the sum of the amino acid substitution scores and gap penalties depend on the scoring system. Different alignments should not be compared based only on the raw score but the normalized score is required.

Chapter 4

Centrality Analysis in PPI Networks

Centrality is a fundamental concept in network analysis, typically linked to a network element's topological significance. As such, it is also a well-studied topic, starting with Bavelas contributions earlier in 1948, Leavitt (1951), Sabidussi (1966), and Freeman (1979), and reaching out to more recent contributions (e.g., Borgatti and Everett (2006), Kosch¨utzki et al. (2005), Boldi and Vigna (2013) for a comprehensive review of centrality indices).

Basic centrality measures that have been proposed over the years can be categorized based on their local or global considerations. For example, node degree centrality, which represents the number of nodes adjacent to a given node i, is a local metric as it only considers the neighborhood of the node at hand. On the other hand, node betweenness centrality, a metric that can be defined as the fraction of the shortest paths from any two nodes in the network that use a node i as an intermediary, is global [107].

4.1. Analyzing centrality in the context of PPI networks

A PPI network represents protein interoperability and how protein coordinates to perform certain functions. A group of similar proteins, cumulatively known as the protein complex, is basically associated with daily cellular activity. Although the protein genomic sequences are already known, it remains a challenge to predict their molecular function. As a result, unknown protein functions were predicted using experimental and computational methods. Computational methods are based on the fact that each protein complex member works in coordination with the others to achieve certain biological activity. So, although the functions of the individual members of a known protein complex are unclear, their characteristics can be easily predicted. For example, a protein complex consists of proteins such as SET3, HST1, SIF2 and HOS2. It is known that this complex is associated with cellular transport activity. Despite the fact that SIF2 is uncharacterized as one of

its member proteins, it is still possible to hypothesize that its function would be related to transportation activity as its parent complex participates in the same activity. Identifying suitable complexes from PPI networks can thus be a major source of biologist assistance. A number of methods for complex detection have been proposed over the years. These methods are based on PPI network topological properties [108].

A combination of topological and functional information is used in some of these methods to detect quality complexes. In the context of PPIN, one can study centrality measures in order to analyze PPI networks from a topological point of view. Centrality was widely used to determine the network significance of nodes. They can also be used in the network to study disease gene organization. Some researchers believe that disease genes tend to have more partners in interaction than other genes. This contradiction calls for a careful analysis of nodes in the network. Various centrality measures have been proposed in graph theory literature. Some centrality measures that use a variety of criteria criterion are discussed in Table 4.1.a. The existence of so many centrality measures highlights the inability of a particular measure to accurately rank the nodes in a graph all the time. Each measure is based on some intuitions and works towards fulfilling them. However, there is no one measure which alone can decide a node's all relevant properties. In order to assist the biologist to choose among these measures.

In the year 2008, Arzucan Özgür and his team evaluated an approach for prostate cancer and showed that degree and eigenvector centrality metrics achieve highly accurate results (95% of the top 20 genes are actually related to the disease), whereas closeness and betweenness centrality metrics introduce genes that are currently unknown to be related to the disease [109]. They were able to extract genes, which are not marked as being related to prostate cancer by the curated PGDB even though there are recent articles that confirm the association of these genes with the disease. The approach can be used to extract known gene-disease

associations from the literature, as well as to infer unknown gene-disease associations which are good candidates for experimental analysis. So, in this thesis,

Measure	Approach	Formula	Application	Limitation
Degree centrality	It is the number of other nodes with which a given node is connected.	$C_{degC}(v_i) = \lvert v_j \rvert.$	Estimating the importance of a protein in a network to predict consequences after its removal from the network	Do not consider the global structure of a network
Betweenness centrality	Quantifies a node's ability to monitor information flow between other vertices.	$C_{betwC_{v_i}} = \sum_{v_a \neq v_b \neq v_i} \frac{\alpha_{v_a v_b}(v_i)}{\alpha_{v_a v_b}}$ where $\alpha_{v_a v_b}$ is the number of shortest paths from node v_a to node v_b and $\alpha_{v_a v_b}(v_i)$ is the number of those paths that pass through v_i.	Estimate biological significance when applied to regulatory networks of mammals This measure has also been used to suggest modular property of the yeast interactome.	A great proportion of nodes do not lie on the route of any shortest path, i such cases, they get a score of 0.
Eigenvector centrality	Assigns a higher rank to nodes that are connected to more important neighbors. It ensures that a node affects all its neighbors in a similar way.	If A represents the PPI network in adjacency matrix format, then equation for eigenvector is $\lambda E_v = A E_v$ where λ is the eigenvalue and E_v is the eigenvector. The eigenvector of a node $v_i \in V$ is given by $C_{EvC_{v_i}} = \frac{1}{\lambda} \sum_{v_u} a(v_u, v_i) E_u$ where $a(v_u, v_i)$ is the entry in the u^{th} row and i^{th} column of the adjacency matrix representation of the network and $\lambda \neq 0$ is a constant	Used in biological networks to identify gene-disease associations and to discover unknown gene-disease associations for further analysis.	Repeated reflection of centrality from central nodes to its neighbors result in accumulation of large centrality near hub nodes in the network.
Closeness centrality	Measures the significance of a node based on its degree of closeness to other nodes in the network. This measure assigns higher values to nodes which can communicate quickly with other nodes in the network.	Closeness centrality of node v_i is given as $C_{cC_{v_i}} = \frac{1}{\sum_{v_j} d(v_i, v_j)}$ where $d(v_i, v_j)$ is the shortest distance from node v_i to node v_j.	Used for ranking pathways and for identifying core metabolite molecules in metabolic networks.	Cannot be applied to networks with disconnected components.
Radiality	Based on the reachability of a node to all other nodes in the network.	It uses the reverse distance matrix for its calculation which is defined as $RvD_{v_i v_j} = diam(G) + 1 - Dist_{v_i v_j}$ where $Dist_{v_i v_j}$ is the distance between nodes v_i and v_j in the network and $diam(G)$ is the length of the shortest path between the most distant nodes, v_x and v_y in the whole network. Radiality of a node v_i is then given as $C_{R_{v_i}} = \frac{\sum_{v_u \neq v_i} RvD_{v_u v_i}}{(n-1)}$ where $RvD_{v_u v_i}$ is the reverse distance between node v_u and v_i and n is the total number of nodes.	Used to interpret the possibility of certain proteins to be highly relevant to a subset of other proteins while being completely dissimilar to another subset of proteins in the network.	
PageRank centrality	It is a variation of the eigenvector centrality measure. It considers both the number and quality of links to decide the score of a node.	Pagerank centrality of node v_i is given as $C_{PgR_{v_i}} = \sum_{v_u \in B_u} \frac{PgR(v_u)}{L(v_u)}$ where B_u is the set of all nodes linking to node v_i and $L(v_u)$ is the number of links from node v_u.	Used to decide upon the criticality of proteins in a regulatory pathway.	Gives more importance to hub nodes.

Table 4.1.a. Centrality measures and their approaches, methods, applications and limitation

Eigenvector centrality is selected to be used for finding the most influential proteins over the networks after some preprocessing.

4.2. Eigenvector centrality using power iteration

A natural extension of degree centrality is eigenvector centrality. In-degree centrality awards one centrality point for every link a node receives. But not all vertices are equivalent: some are more relevant than others, and, reasonably, endorsements from important nodes count more [110] [111].

Eigenvector centrality differs from in-degree centrality: a node receiving many links does not necessarily have a high eigenvector centrality (it might be that all linkers have low or null eigenvector centrality). Moreover, a node with high eigenvector centrality is not necessarily highly linked (the node might have few but important linkers).

Eigenvector centrality, regarded as a ranking measure, is a remarkably old method. Early pioneers of this technique are Wassily W. Leontief and John R. Seeley. Google's PageRank and the Katz centrality are variants of the eigenvector centrality [112].

For a given graph G: =(V, E) with |V| vertices let A = $(a_{v,t})$ be the adjacency matrix, i.e. $(a_{v,t})$ = 1 if vertex v is linked to vertex t, and $(a_{v,t})$ = 0

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t$$

otherwise. The relative centrality score of vertex v can be defined as:

Where M (v) is a set of the neighbors of v and λ is a constant. With a small rearrangement this can be rewritten in vector notation as the eigenvector equation:

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x}$$

In general, there will be many different eigenvalues λ for which a non-zero eigenvector solution exists. However, the additional requirement that all the entries in the eigenvector be non-negative implies (by the Perron–Frobenius theorem) that

only the greatest eigenvalue results in the desired centrality measure [113]. The v^{th} component of the related eigenvector then gives the relative centrality score of the vertex v in the network. The eigenvector is only defined up to a common factor, so only the ratios of the centralities of the vertices are well defined [114]. To define an absolute score one must normalize the Eigen vector e.g. such that the sum over all vertices is 1 or the total number of vertices n. Power iteration is one of many eigenvalue algorithms that may be used to find this dominant eigenvector. Furthermore, this can be generalized so that the entries in A can be real numbers representing connection strengths, as in a stochastic matrix.

Below there is a step by step evolution to find eigenvector centrality of a toy graph (Figure 4.2.a.)

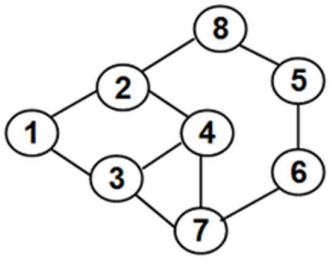

Figure 4.2.a. A simple toy graph (Unweighted)

Iteration 1,

For this graph, consider each node value to be 1 for this graph and create an adjacency matrix (Figure 4.2.b) by taking edge weight 1 if there is an edge between two nodes, otherwise considered to be 0.

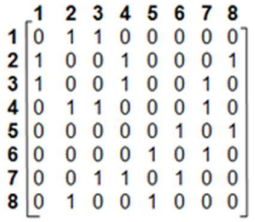

Figure 4.2.b. Adjacency matrix for toy graph

Now, multiply Eigenvector (dominant) with adjacency matrix (Figure 4.2.c) and using result matrix normalize value is

$$\sqrt{2^2 + 3^2 + 3^2 + 3^2 + 2^2 + 2^2 + 3^2 + 2^2} = 7.21$$

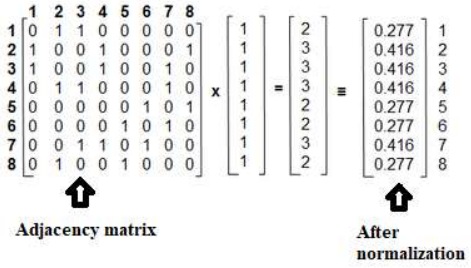

Figure 4.2.c. Result after first iteration (Matrix multiplication and Normalization)

Iteration 2,

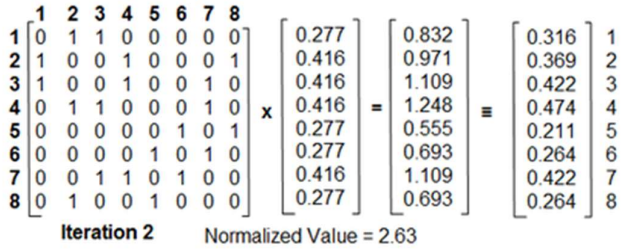

Figure 4.2.d. Second Iteration. Normalized value 2.63

Iteration 3 (Figure 4.2.e),

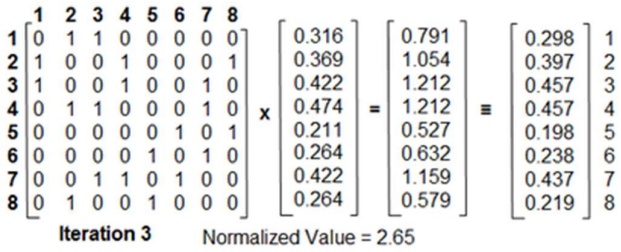

Figure 4.2.e Third iteration. Normalized value = 2.65

Iteration 4 (Figure 4.2.f),

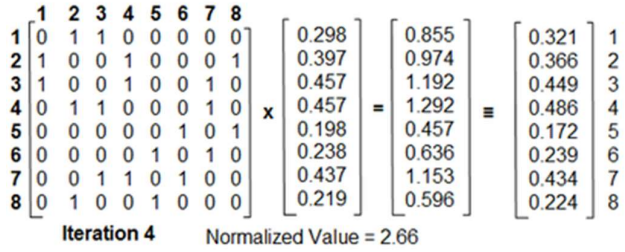

Figure 4.2.f. Fourth iteration. Normalized value = 2.66

Iteration 5 (Figure 4.2.g),

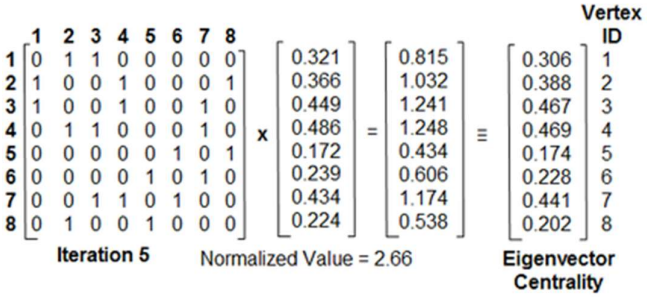

Figure 4.2.g. Fifth iteration. Normalized Value is 2.66

For this thesis, there is a weight in every interacting edge because of protein-protein interaction nature. So, the weighted adjacency matrix of a weighted graph is used.

Chapter 5

Methodology

Seven protein biomarkers are selected as initial biomarkers for prostate cancer after a brief literature survey and previous research papers. In chapter 2, the UniProt database was also used to find out protein UniProt Id. The protein families (Table 5.a) were found using the protein UniProt ID. For humans and mouse, all proteins belonging to the same family were found out from each and every family.

Table 5.a. UniProt ID and respective protein families

UniProt ID	**Protein Family**
P02792	PF00210
P10809	PF00118
P30101	PF00085
Q9NR45	PF03102
P07355	PF00191
Q01469	PF00061
P31949	PF00036, PF01023

In Pfam database, an automatically generated full alignment, which contains all detectable protein sequences belonging to a family, as defined by profile HMM searches of primary sequence database, is used to search all protein sequences of the same family by FASTA method (described in Chapter 3). Only human and mouse proteins need to be selected from the families for further progress. From all the protein sequences of every species only human and mouse proteins are selected separately for each and every family. A Flow chart describing the entire process is provided in Figure 5.a.

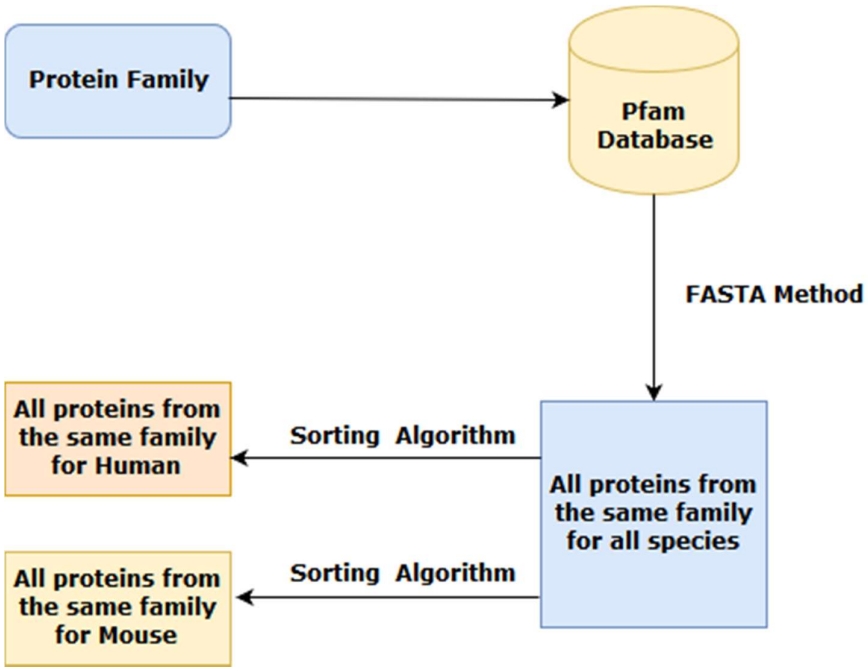

Figure 5.a. Flow chart to find species specific all proteins of a family

PPIN needs to be found for each and every protein family of the protein of a particular species, which has been founded from the Pfam database (SMART and InterPro is used when there is no family description of a protein in Pfam), for the purpose of this thesis work. For this intent, StringDB is used. Figure 5.b is a PPI

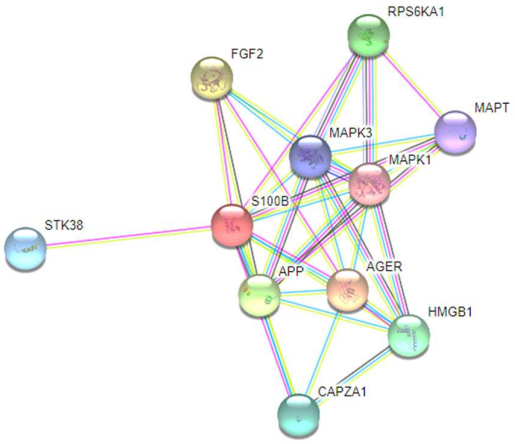

Figure 5.b. PPI Network of protein S100B.

network of a protein's current interactions which has been drawn by this approach.

Weighted PPIN is found from StringDB for each protein where weight is the combined interaction score between two proteins. The combined score is computed by combining the probabilities from the different evidence channels (neighbourhood on chromosome, homology, phylogenetic cooccurrence, automated text-mining, gene fusion, database annotated, experimentally determined) and corrected for the probability of randomly observing an interaction. To each channel, a 'prior' has been added to account for the probability that two randomly picked proteins are interacting [97]. Before combing the channels the 'prior' has to be removed and then added back again to the combined score. All current interactions with their combined interaction scores are shown in Table 5.b for protein S100B.

Node1	Node2	Combined score
RELA	NFKB1	0.993
S100B	GFAP	0.981
MAPK3	MAPK1	0.978
MAPK3	NFKB1	0.976
NFKB1	MAPK1	0.975
AGER	HMGB1	0.964
HMGB1	NFKB1	0.961
S100B	APP	0.96
HMGB1	MAPK1	0.952
AGER	S100B	0.952
RELA	MAPK1	0.951
RELA	MAPK3	0.95
RELA	HMGB1	0.943
APP	MAPK3	0.942
HMGB1	MAPK3	0.941
HMGB1	S100B	0.938
APP	MAPK1	0.936
APP	NFKB1	0.934
S100B	CAPZA1	0.933
S100B	NFKB1	0.929
S100B	MAPK3	0.927
S100B	MAPK1	0.927
AGER	NFKB1	0.923
AGER	APP	0.92
HMGB1	CAPZA1	0.92
RELA	APP	0.919
HMGB1	APP	0.917

RELA	S100B	0.912
S100B	NCOR1	0.91
NCOR1	GFAP	0.91
AGER	MAPK3	0.909
AGER	MAPK1	0.904
APP	CAPZA1	0.9
RELA	AGER	0.9
AGER	CAPZA1	0.9
APP	GFAP	0.714
NCOR1	NFKB1	0.561
RELA	NCOR1	0.501
MAPK3	GFAP	0.498
APP	NCOR1	0.424
GFAP	MAPK1	0.41

Table 5.b. S100B current PPIN with combined scoring of different interaction

Group was formed separately from the protein network of mouse protein families and human protein families. After that, each grouped protein network was merged and the final network for each protein family was formed (8 for Human and 8 for Mouse). For example protein family PF01023 of Human, there is 8 protein networks are found by the previous process. Now 8 networks are merged and form a single network (Figure 5.c).

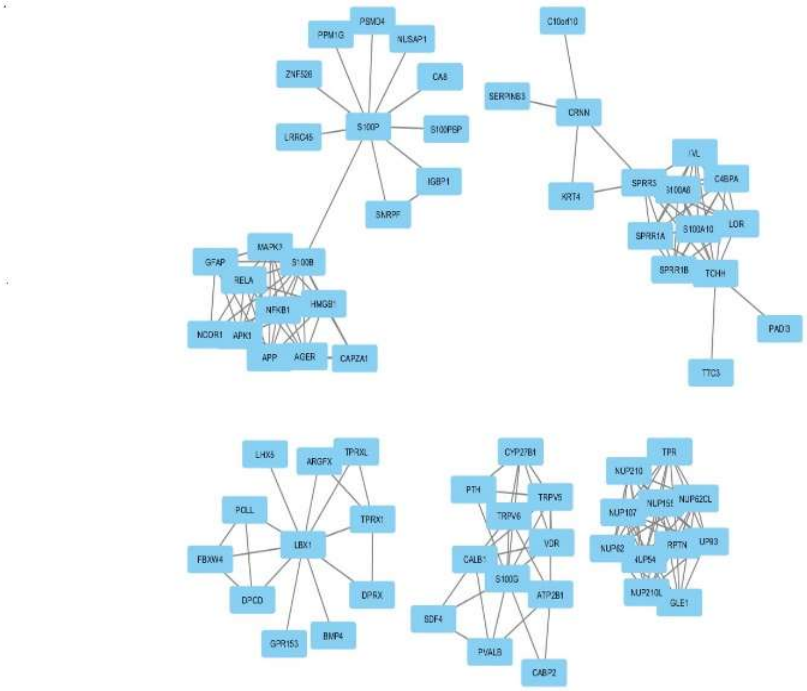

Figure 5.c. Marged PPIN for protein family PF01023 generated using Cytoscape from interaction data table 5.b. Created using Cytoscape

The eigenvector centrality power iteration method is used on each merged weighted PPIN to achieve a score for each node of each network. That network's most influential node is the family's most influential protein for thesis goal. Figure 5.d is a general flowchart of this approach.

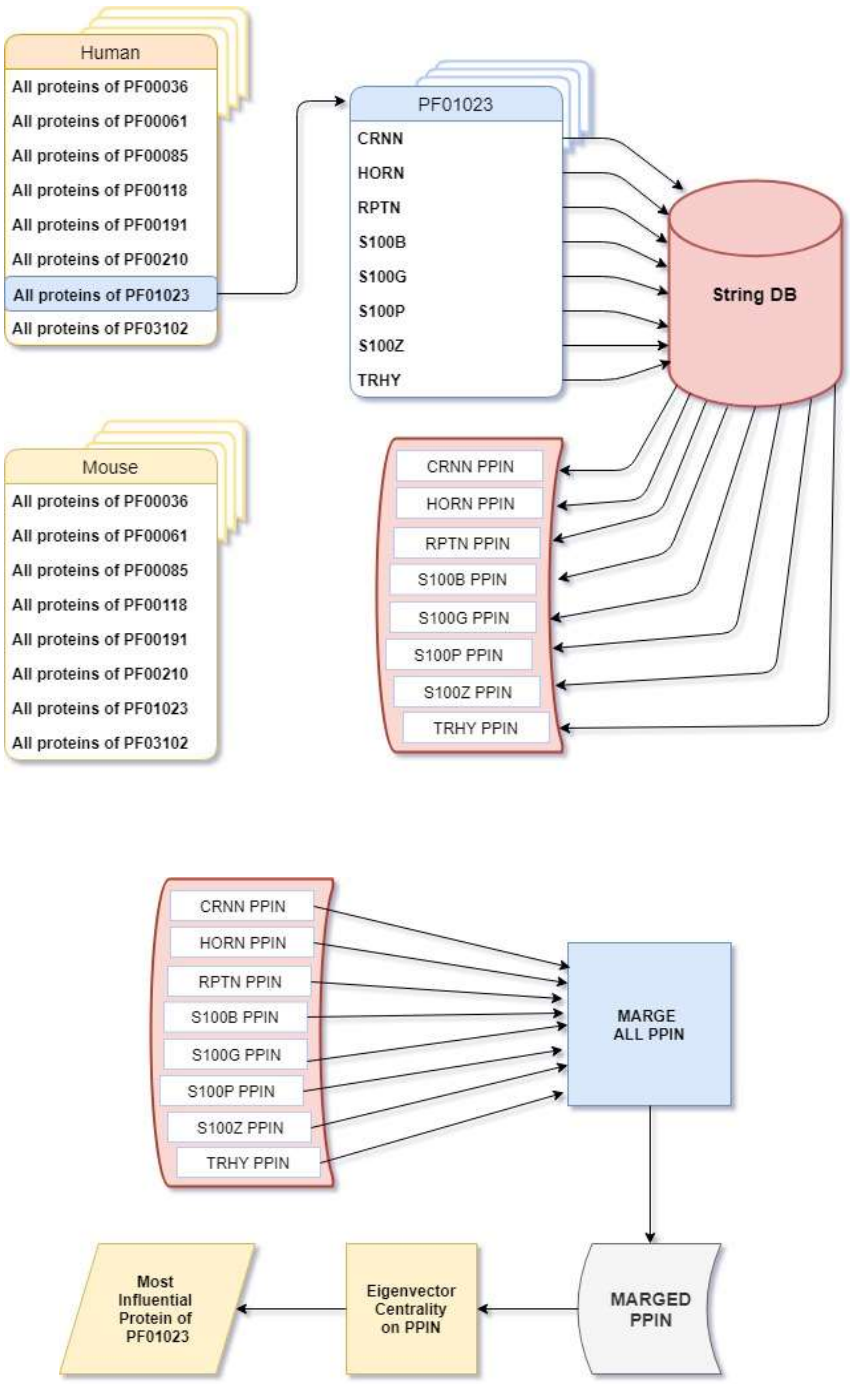

Figure 5.d Overall flowchart of processing a single family's all proteins to get most influential proteins of this family

Chapter 6

Result and Discussion

In this chapter, experimental results are discussed in order to find the most influential proteins of each family for each species. First, it has been found if there is any biological significance of influential proteins in cancer. Then validate the thesis goal approach.

By evaluating the thesis approach over initial biomarkers, 8 protein families are first found. As a result, two most influential nodes of merged PPINs of each family are chosen for each species to check whether or not they have any biological significance. Following the proposed methodology, a total of 32 proteins are found as a final result. These proteins are shown with the family and species in Table 6.a.

Table 6.a List of result proteins with their family and species

Protein Family	**Influential proteins for Human**	**Influential proteins for Mouse**
PF00036	ACTN2, TPM1	PPP3CB,PPP3CC
PF00061	PTGDS, PTGS2	NCOA1,NCOA2
PF00085	HSPA5, CALR	HSPA5,CALR
PF00118	CCT 2, CCT4	CCT2,CCT5
PF00191	CASP3, BCL2	CASP3,BCL2
PF00210	FTH1, CTSD	FTI1,LYZ2
PF01023	NUP93,NUP115	NUP 107,NUP115
PF03102	NANS, GNE	ST8SIA3,NCAM1

6.1. Validation

Now, there should be some proven biological significance for cancer to validate the goal of the thesis. This section, the biological significance of each distinct protein has been discussed with respect to humans. Mouse PPINs influential node proteins have been also

checked for humans because there are similarities of PPINs and their evolutionary relationship.

6.1.1. ACTN2

In the human protein atlas, ACTN2 is not classified as cancer-related genes and over the years only 10 publications on ACTN2 and cancer are tracked on PubMed.

6.1.2. TPM1

In the human protein atlas, TPM1 is classified as cancer related genes and over the years 78 publications on TPM1 and cancer are tracked on PubMed.

In 2019, Jin Wang et al. concluded that TPM1 functions as a tumor suppressor for RCC cell proliferation, angiogenesis, and metastasis, suggesting it is a potential therapeutic target for advanced RCC. (Carcinoma of the renal cell) [115].

Also Hao Pan, his team's study demonstrated that TPM1 could act as a tumor-suppressing gene in OSCC. By using real-time Real-Time Polymerase Chain Reaction, western blotting and immunohistochemistry, they detected TPM1 expression at both the mRNA and protein levels in OSCC cells and specimens from patients. They found that TPM1 expression levels were significantly higher in adjacent normal tissues than in OSCC lesions [116].

The tumor suppressor gene tropomyosin 1 (TPM1) is downregulated in several human cancer types [117-125].

6.1.3. PTGDS

In the human protein atlas, PTGDS is not classified as cancer related genes and over the years 25 publications on PTGDS and cancer are tracked on PubMed.

6.1.4. PTGS2

In the human protein atlas, PTGS2 is classified as cancer related genes and over the years 2085 publications on PTGS2 and cancer are tracked on PubMed.

Kosuke Mima *et al. concluded that MIR21* expression level in colorectal carcinoma is associated with worse clinical outcome, and this association is stronger in carcinomas expressing high-level PTGS2, suggesting complex roles of immunity and inflammation in tumor progression [126].

Freitas-Alves et al. research suggests that a potential contribution of *PTGS2* genotyping for additional prognostic evaluation of breast cancer outcomes, especially among obese patients [127].

Cebrián, A. et al also conclude that functional PTGS2 polymorphism-based models as novel predictive markers in metastatic renal cell carcinoma patients receiving first-line sunitinib [128].

6.1.5. HSPA5

In the human protein atlas, HSPA5 is not classified as cancer related genes and over the years 113 publications on HSPA5 and cancer are tracked on PubMed.

Chen et al. conclude that HSPA5 is overexpressed in advanced breast cancer and relates to poor survival outcome and metastatic event [129].

6.1.6. CCT

In the human protein atlas, CCT is not classified as cancer related genes and over the years 573 publications on CCT and cancer are tracked on PubMed.

Guest et al. identified two genes, TCP1 and CCT2, as being recurrently altered in breast cancer, necessary for growth/survival of breast cancer cells in vitro, and determinants of overall survival in breast cancer patients. They also show that expression of TCP1 is regulated by driver oncogene activation of PI3K signaling in breast cancer. Interestingly, the TCP1 and CCT2 genes both encode for components of a multi-protein chaperone complex in the cell known as the TCP1 Containing Ring Complex (TRiC). Their results demonstrate a role for the TRiC subunits TCP1 and CCT2, and potentially the entire TRiC complex, in breast

cancer and provide rationale for TRiC as a novel therapeutic target in breast cancer [130].

Khaled et al. found that these cancers expressed higher levels of CCT2 as compared to normal tissues. Small cell lung cancer (SCLC) stood out as having statistically significant difference in CCT2. Higher levels of CCT2 in tumors from lung cancer patients were also associated with decreased survival. These results indicate that in SCLC, changes in CCT levels could be used as a biomarker for diagnosis and that targeting CCT for inhibition with CT20p is a promising treatment approach for those cancers such as SCLC that currently lack targeted therapeutics [131].

To investigate the role of CCT in cancer progression, they examined protein levels of CCT subunits in liver, prostate, and lung cancer using human tissue microarray sand they found that these cancers expressed higher levels of CCT2 as compared to normal tissues. Small cell lung cancer (SCLC) stood out as having statistically significant difference in CCT2. Higher levels of CCT2 in tumors from lung cancer patients were also associated with decreased survival.

6.1.7. CASP3

In the human protein atlas, CASP3 is classified as cancer related genes and over the years 6564 publications on CASP3 and cancer are tracked on PubMed.

The current study demonstrates that high caspase-3 expression is significantly associated with adverse breast cancer-specific survival.

Devarajan et al. conclude that the tumor cells as well as the normal parenchyma surrounding the tumor lack caspase 3 expression in the majority of breast cancer patients. As one would expect, loss of expression or function of this key caspase can render breast cancer cells resistant to apoptosis in response to certain apoptotic stimuli including chemotherapeutic drugs and thus may affect the outcome and prognosis of the disease. These findings may have important clinical

implications in terms of using caspase-3 not only as a marker of disease but also as a therapeutic target for breast cancer [132].

In one study, researchers demonstrated using immunohistochemistry with caspase-3 antibodies that increased levels of cleaved caspase-3 significantly associated with a higher rate of cancer recurrence and decreased survival time in cancer patients (site) and their findings demonstrate that cleaved caspase-3 is well correlated to progression, aggressive behaviors in the studied cancer. Elevated cleaved caspase-3 is associated with shortened OS, pointing it as a potential predictive factor for the prognosis of four types of cancers, namely gastric cancer, ovarian cancer, cervical cancer, colorectal cancer, though our study has some limitations such as failing to describe its prognostic value in terms of tumor recurrence [133]. Another research group at the University Of Pittsburgh School Of Medicine found that loss of caspase-3 sensitized cancer cells to DNA-damaging therapeutic agents and they have shown that genetic ablation of caspase-3 in colon cancer cells increases sensitivity to DNA-damaging agents through Receptor-interacting protein 1 -dependent necrosis without compromising apoptosis. Therefore, pharmacological manipulation of caspase-3 may provide a novel approach to enhance the killing of chemoresistant cancer cells [134].

6.1.8. BCL2

In the human protein atlas, BCL2 is classified as cancer related genes and over the years only 15274 publications on BCL2 and cancer are tracked on PubMed.

Less than 5% of patients with chronic lymphocytic leukemia have a detectable rearrangement of the BCL-2 gene, although the vast majority of BCL-2 overexpress [135]. Increased expression of BCL-2 is also frequently found in acute myeloid leukemia and in almost all patients with acute lymphocytic leukemia [136] [137].

BCL2 itself seems to act as both an oncogene and a tumor suppressor gene in different tumor types [138].

Renner et al. provide evidence that the homozygous *BCL2*-938 CC genotype is strongly associated with reduced OS in prostate cancer patients [139].

6.1.9. FTH1

In the human protein atlas, FTH1 is classified as cancer related genes and over the years 40 publications on FTH1 and cancer are tracked on Pubmed.

An FTH pseudogene microRNA network regulates tumorigenesis in prostate cancer [140]. Also Huang et al. suggest that FTH1 expression is an effective prognostic and diagnostic biomarker for RCC [141].

6.1.10. CTSD

In the human protein atlas, CTSD is classified as cancer related genes and over the years 148 publications on CSTD and cancer are tracked on PubMed.

A clinical proteomics workflow research identified CTSD as an over-expressed protein in osteosarcomas and pulmonary metastases and could thus serve as a new biomarker for individualized treatment regimens for osteosarcoma patients, even at metastasis [142].

6.1.11. NUP93

In the human protein atlas, NUP93 is classified as cancer related genes and over the years 257 publications on NUP93 and cancer are tracked on PubMed.

Previously, Researcher detected NUP93 as a mutational cancer driver in 1 cancer type: Uterine corpus endometrioid carcinoma [143].

6.1.12. NUP115

In the human protein atlas, NUP115 is no data about cancer related genes.

6.1.13. PPP3CB

In the human protein atlas, PPP3CB is not classified as cancer related genes and over the years only 9 publications on PPP3CB and cancer are tracked on PubMed.

6.1.14. PPP3CC

In the human protein atlas, PPP3CC is not classified as cancer related genes and over the years only 6 publications on PPP3CC and cancer are tracked on PubMed.

6.1.15. NCOA1

In the human protein atlas, NCOA1 is classified as cancer related genes and over the years 243 publications on NCOA1 and cancer are tracked on PubMed.

It is well established that tumor growth requires angiogenesis for supplying oxygen and nutrients [144]. However, Wang et al. found that although NCOA1 promotes angiogenesis in breast tumors, overexpression or knockout of NCOA1 in mice does not significantly affect mammary tumor growth [145, 146]. NCOA1 also promotes breast cancer metathesis [147].

6.1.16. NCOA2

In the human protein atlas, NCOA2 is classified as cancer related genes and over the years only 173 publications on NCOA2 and cancer are tracked on PubMed.

NCOA2, which has been thought to be recruited as a coactivator, plays a corepressive role in AR of prostate cancer cells when treated with antiandrogens, suggesting its potential as a therapeutic target [148]. It also play a role in liver tumorigenesis [149].

6.1.17. FTL1

Previously, Ferritin light chain 1 was selected as initial biomarkers for cancer.

6.1.18. LYZ2

In the human protein atlas, LYZ2 is not classified as cancer related genes and over the years only 9 publications on LYZ2 and cancer are tracked on PubMed.

6.1.19. NANS

In the human protein atlas, NANS is not classified as cancer related genes and over the years no publications on NANS and cancer are tracked on PubMed.

6.1.20. GNE

In the human protein atlas, GNE is not classified as cancer related genes and over the years no publications on GNE and cancer are tracked on PubMed.

6.1.21. ST8SIA

In the human protein atlas, ST8SIA is not classified as cancer related genes and over the years only no publications on ST8SIA and cancer are tracked on PubMed.

6.1.22. NCAM1

In the human protein atlas, NCAM1 is not classified as cancer related genes and over the years 353 publications on NCAM1 and cancer are tracked on PubMed.

In the year 2019, Sasca et al. suggest NCAM1 as a biomarker to guide acute myeloid leukemia treatment [150].

There are only 22 distinct proteins from 32 proteins. From 22 distinct proteins, 13 proteins are biologically and clinically proven to be a biomarker for cancer diagnosis and no related data found for one.

Chapter 7

Conclusions

This thesis introduces a new approach to finding the most influential cancer proteins from processed PPINs using clinically proven prostate cancer biomarkers. Because of the huge and ever-growing protein database, it is not possible to manually check every biological significance of proteins to find biomarkers. The proposed method uses previously known biomarkers to find other influential proteins in processed PPINs that can be biomarkers. Instead of finding significant protein from the huge database, it is easy to check the biological significance of the influential proteins, which is a more faster process than traditional biomarkers selection.

7.1 Applications of the method

Biomarkers are the need of today to detect many lethal diseases including cancer. Cancer treatment is possible if it can be arrested tumor in the primary stage. The generalized form of this proposed method can be applied to many other disease networks to find influential proteins that boost the diagnosis process. To find new biomarkers instead of checking the entire protein database, finding this from some influential proteins is quick and efficient.

7.2 Limitations of present work

Not all influential proteins, as a result, are biomarkers or some of them may not have any biological significance for the disease. In this method, structural protein information is also overlooked. It should always be needed to validate the results of this method with biological significance and clinical evidence.

7.3 Future Scope

Proteins structural information can be added as additional information to achieve a more precise outcome. Using protein domain knowledge, it is possible to use the method to find a drug target. The more intelligent method like deep neural network, can be implemented on the network to find more accurately targeting influential proteins. Using this approach on different species, it is possible to track some evolutionary changes in the network and also to compare the structures of the influential proteins, it is possible to find diseases caused by structural changes and mutation-based diseases.

References

1. Carvalho Tito, Zhu Tian(2014, May 06) The Human Genome Project (1990-2003). Retrieved from embryo.asu.edu/pages/human-genome-project-1990-2003

2. Shruthi, BasavaradhyaSahukar & Vinodhkumar, Palani & Selvamani, Manickam. (2016). Proteomics: A new perspective for cancer. Advanced Biomedical Research. 5. 67. 10.4103/2277-9175.180636.

3. CRICK, F. (1970). Central Dogma of Molecular Biology. Nature, 227(5258), 561–563. https://doi.org/10.1038/227561a0

4. Watson JD, Crick FH (1953). "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid". Nature. 171 (4356): 737–8. Bibcode:1953Natur.171..737W. doi:10.1038/171737a0

5. Nirenberg MW, Matthaei JH (October 1961). "The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides". Proceedings of the National Academy of Sciences of the United States of America. 47 (10): 1588–602. Bibcode:1961PNAS...47.1588N. doi:10.1073/pnas.47.10.1588

6. Gutteridge A, Thornton JM (November 2005). "Understanding nature's catalytic toolkit". Trends in Biochemical Sciences. 30 (11): 622–29. doi:10.1016/j.tibs.2005.09.006

7. Dayhoff, M. O. (1974). "Computer analysis of protein sequences". Federation Proceedings. 33 (12): 2314–2316.

8. Bilal Aslam, Madiha Basit, Muhammad Atif Nisar, Mohsin Khurshid, Muhammad Hidayat Rasool, Proteomics: Technologies and Their Applications, Journal of Chromatographic Science, Volume 55, Issue 2, 1 February 2017, Pages 182–196,https://doi.org/10.1093/chromsci/bmw167

9. Sallam, R. M. (2015). Proteomics in Cancer Biomarkers Discovery: Challenges and Applications. Disease Markers, 2015, 1–12. https://doi.org/10.1155/2015/321370

10. Deribe YL, Pawson T, Dikic I (2010) Post-translational modifications in signal integration. Nat Struct Mol Biol 17: 666–672. 10.1038/nsmb.1842

11. Zhao S, Xu W, Jiang W, Yu W, Lin Y, et al. (2010) Regulation of cellular metabolism by protein lysine acetylation. Science 327: 1000–1004. 10.1126/science.1179689

12. De Las Rivas J, Fontanillo C (June 2010). "Protein-protein interactions essentials: key concepts to building and analyzing interactome networks". PLoS Computational Biology. 6 (6): e1000807. Bibcode:2010PLSCB...6E0807D. doi:10.1371/journal.pcbi.1000807

13. Gonzalez MW, Kann MG. Chapter 4: Protein interactions and disease. PLoS Comput Biol. ;8(12):e1002819. doi:10.1371/journal.pcbi.1002819

14. Vlaic, S., Conrad, T., Tokarski-Schnelle, C., Gustafsson, M., Dahmen, U., Guthke, R., & Schuster, S. (2018). ModuleDiscoverer: Identification of regulatory modules in protein-protein interaction networks. Scientific Reports, 8(1). https://doi.org/10.1038/s41598-017-18370-2

15. Sunjae Lee, Cheng Zhang, Muhammad Arif, Zhengtao Liu, Rui Benfeitas, Gholamreza Bidkhori, Sumit Deshmukh, Mohamed Al Shobky, Alen Lovric, Jan Boren, Jens Nielsen, Mathias Uhlen, Adil Mardinoglu, TCSBN: a database of tissue and cancer specific biological networks, Nucleic Acids Research, Volume 46, Issue D1, 4 January 2018, Pages D595–D600, https://doi.org/10.1093/nar/gkx994

16. Cho, W. C. (2007). Contribution of oncoproteomics to cancer biomarker discovery. Molecular Cancer, 6(1), 25. https://doi.org/10.1186/1476-4598-6-25

17. Banks RE, Dunn MJ, Hochstrasser DF, Sanchez JC, Blackstock W, Pappin DJ, et al. Proteomics: New perspectives, new biomedical opportunities. Lancet. 2000;356:1749–56.

18. Petersen, John R., and Amin A. Mohammad, eds. Clinical and Forensic Applications of Capillary Electrophoresis. New York: Humana P, 2001.

19. National Human Genome Research Institute (2010, July 23). Why Mouse Matters. Retrived from https://www.genome.gov/10001345/importance-of-mouse-genome.

20. Rao, V. S., Srinivas, K., Sujini, G. N., & Kumar, G. N. (2014). Protein-protein interaction detection: methods and analysis. International journal of proteomics, 2014, 147648. doi:10.1155/2014/147648

21. Adelmant, G., Garg, B. K., Tavares, M., Card, J. D., & Marto, J. A. (2019). Tandem Affinity Purification and Mass Spectrometry (TAP-MS) for the Analysis of Protein Complexes. Current Protocols in Protein Science, e84. https://doi.org/10.1002/cpps.84

22. Mayers, G. L., & van Oss, C. J. (1998). Affinity Chromatography. In Encyclopedia of Immunology (pp. 47–49). Elsevier. https://doi.org/10.1006

/rwei.1999.0012

23. Rosato, E. (Ed.). (2007). Circadian Rhythms. Methods in Molecular Biology. Humana Press. https://doi.org/10.1007/978-1-59745-257-1

24. S.H. Strogatz, “Exploring Complex Networks,” Nature, vol. 410, pp. 268-276, 2001

25. Protein-Protein Interaction Networks," in IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 2, pp. 261-277, Feb. 2014. doi: 10.1109/TKDE.2012.225

26. G.D. Bader and C.W. Hogue, “An Automated Method for Finding Molecular Complexes in Large Protein Interaction Networks,” BMC Bioinformatics, vol. 4, article 2, 2003.

27. K. Rhrissorrakrai and K.C. Gunsalus, “MINE: Module Identification in Networks,” BMC Bioinformatics, vol. 12, article 192, 2011.

28. B. Adamcsek, G. Palla, I.J. Farkas, I. Derenyi, and T. Vicsek, “CFinder: Locating Cliques and Overlapping Modules in Biological Networks,” Bioinformatics, vol. 22, no. 8, pp. 1021-1023, 2006.

29. P. Gergely, D. Imre, F. Illes, and V. Tamas, “Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society,” Nature, vol. 435, pp. 814-818, 2005.

30. M. Altaf-UI-adim et al., “Development and Implementation of an Algorithm for Detection of Protein Complexes in Large Interaction Networks,” BMC Bioinformatics, vol. 7, article 207, 2006.

31. X.L. Li, S.H. Tan, C.S. Foo, and S.K. Ng, “Interaction Graph Mining for Protein Complexes Using Local Clique Merging,” Genome Informatics, vol. 16, no. 2, pp. 260-269, 2005.

32. H.N. Chua, K. Ning, W.K. Sung, H.W. Leong, and L. Wong, “Using Indirect Protein-Protein Interactions for Protein Complex Prediction,” Proc. Ann. Int’l Conf. Computational Systems Bioinformatics, pp. 97-109, 2007.

33. M. Mete, F. Tang, X. Xu, and N. Yuruk, “A Structural Approach for Finding Functional Modules from Large Biological Networks,” BMC Bioinformatics, vol. 9, article S19, 2008.

34. A. Abdullah, S. Deris, S.Z.M. Hashim, and H.M. Jamil, “Graph Partitioning Method for Functional Module Detections of Protein Interaction Network,” Proc. Int’l Conf. Computer Technology and Development, pp. 230-234, 2009.

35. E. Ramadan, C. Osgood, and A. Pothen, “Discovering Overlapping Modules and Bridge Proteins in Proteomic Networks,” Proc. ACM Int’l Conf. Bioinformatics and Computational Biology (BCB ’10), pp. 366-369, 2010.

36. D. Efimov, N. Zaki, and J. Berengueres, “Detecting Protein Complexes from Noisy Protein Interaction Data,” Proc. 11th Int’l Workshop Data Mining in Bioinformatics, pp. 1-7, 2012.

37. E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.L. Barabasi, “Hierarchical Organization of Modularity in Metabolic Networks,” Science, vol. 297, pp. 1551-1555, 2002.

38. P. Holme, M. Huss, and H. Jeong, “Subnetwork Hierarchies of Biochemical Pathways,” Bioinformatics, vol. 19, pp. 532-538, 2003.

39. V. Arnau, S. Mars, and I. Marin, “Iterative Cluster Analysis of Protein Interaction Data,” Bioinformatics, vol. 21, no. 3, pp. 364378, 2005.

40. R. Aldecoa and I. Marin, “Jerarca: Efficient Analysis of Complex Networks Using Hierarchical Clustering,” PLoS ONE, vol. 5, no. 7, p. e11585, 2010.

41. Y.R. Cho, W. Hwang, and A.D. Zhang, “Efficient Modularization of Weighted Protein Interaction Networks Using k-Hop Graph Reduction,” Proc. Sixth IEEE Symp. Bioinformatics and Bioeng. (BIBM), pp. 289-298, 2006.

42. A.D. King, N. Przulj, and I. Jurisica, “Protein Complex Prediction via Cost-Based Clustering,” Bioinformatics, vol. 20, no. 17, pp. 30133020, 2004.

43. J. Vlasblom and S.J. Wodak, “Markov Clustering versus Affinity Propagation for the Partitioning of Protein Interaction Graphs,” BMC Bioinformatics, vol. 10, article 99, 2009.

44. R. Dunn, F. Dudbridge, and C.M. Sanderson, “The Use of EdgeBetweenness Clustering to Investigate Biological Function in Protein Interaction Networks,” BMC Bioinformatics, vol. 6, article 39, 2005.

45. A.J. Enright, S. Van Dongen, and C.A. Ouzounis, “An Efficient Algorithm for Large-Scale Detection of Protein Families,” Nucleic Acids Research, vol. 30, no. 7, pp. 1575-1584, 2002.

46. J.B. Pereira-Leal, A.J. Enright, and C.A. Ouzounis, “Detection of Functional Modules from Protein Interaction Networks,” Proteins, vol. 54, pp. 49-57, 2004.

47. W. Hwang, Y.R. Cho, A. Zhang, and M. Ramanathan, “A Novel Functional Module Detection Algorithm for Protein-Protein Interaction Networks,” Algorithms for Molecular Biology, vol. 1, p. 24, 2006.

48. W. Hwang, Y.R. Cho, A.D. Zhang, and M. Ramanathan, “CASCADE: A Novel Quasi All Paths-Based Network Analysis Algorithm for Clustering Biological Interactions,” BMC Bioinformatics, vol. 9, article 64, 2008.

49. Y.R. Cho, W. Hwang, M. Ramanathan, and A.D. Zhang, “Semantic Integration to Identify Overlapping Functional Modules in Protein Interaction Networks,” BMC Bioinformatics, vol. 8, article 265, 2007.

50. J. Feng, R. Jiang, and T. Jiang, “A Max-Flow Based Approach to the Identification of Protein Complexes Using Protein Interaction and Microarray Data,” Computational Systems Bioinformatics, vol. 7, pp. 51-62, 2008.

51. C. Kamp and K. Christensen, “Spectral Analysis of Protein-Protein Interactions in Drosophila Melanogaster,” Physical Rev., vol. 71, no. 4, p. 041911, 2005.

52. T.Z. Sen and R.L. Jernigan, “Functional Clustering of Yeast Proteins from the Protein Protein Interaction Network,” BMC Bioinformatics, vol. 7, article 355, 2006.

53. G. Qin and L. Gao, “Spectral Clustering for Detecting Protein Complexes in Protein Protein Interaction (PPI) Networks,” Math. and Computer Modeling, vol. 52, pp. 20662074, 2010.

54. K. Inoue, W. Li, and H. Kurata, “Diffusion Model Based Spectral Clustering for Protein Protein Interaction Networks,” PLoS ONE, vol. 5, no. 9, p. e12623, 2010.

55. Y.J. Qi, F. Balem, C. Faloutsos, J. Klein-Seetharaman, and Z. BarJoseph, “Protein Complex Identification by Supervised Graph Local Clustering,” Bioinformatics, vol. 24, no. 13, pp. 250-268, 2008.

56. H.C. Leung, S.M. Yiu, Q. Xiang, and F.Y. Chin, “Predicting Protein Complexes from PPI Data: A Core-Attachment Approach,” J. Computational Biology, vol. 16, no. 2, pp. 133-144, 2009.

57. M. Wu, X.L. Li, C.K. Kwoh, and S.K. Ng, “A Core-Attachment Based Method to Detect Protein Complexes in PPI Networks,” BMC Bioinformatics, vol. 10, article 169, 2009.

58. X.K. Ma and L. Gao, “Predicting Protein Complexes in Protein Interaction Networks Using a Core-Attachment Algorithm Based on Graph Communicability,” Information Sciences, vol. 189, pp. 233-254, 2012.

59. J. Sallim, R. Abdullah, and A.T. Khader, “ACOPIN: An ACO Algorithm with TSP Approach for Clustering Proteins from Protein Interaction Network,” Proc. Second UKSIM European Symp. Computer Modeling and Simulation, pp. 203-208, 2008. 53. S.

60. Wu, X.J. Lei, and J.F. Tian, “Clustering PPI Network Based on Functional Flow Model through Artificial Bee Colony Algorithm,” Proc. Seventh Int’l Conf. Natural Computation, pp. 92-96, 2011.

61. J.Z. Ji, Z.J. Liu, A.D. Zhang, L. Jiao, and C.N. Liu, “Improved Ant Colony Optimization for Detecting Functional Modules in Protein-Protein

Interaction Networks," Proc. Int'l Conf. Information Computing and Applications (ICICA '12), pp. 404-413, 2012.

62. J.Z. Ji, Z.J. Liu, A.D. Zhang, L. Jiao, and C.N. Liu, "Ant Colony Optimization with MultiAgent Evolutionary for Detecting Functional Modules in Protein-Protein Interaction Networks," Proc. Int'l Conf. Information Computing and Applications (ICICA '12), pp. 445-453, 2012.

63. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. (2001). Clinical Pharmacology & Therapeutics, 69(3), 89–95. https://doi.org/10.1067/mcp.2001.113989

64. WHO International Programme on Chemical Safety Biomarkers in Risk Assessment: Validity and Validation. 2001. Retrieved from http://www.inchem.org/documents/ehc/ehc/ehc222.htm.

65. WHO International Programme on Chemical Safety Biomarkers and Risk Assessment: Concepts and Principles. 1993. Retrieved from http://www.inchem.org/documents/ehc/ehc/ehc155.htm.

66. Taitt H. E. (2018). Global Trends and Prostate Cancer: A Review of Incidence, Detection, and Mortality as Influenced by Race, Ethnicity, and Geographic Location. American journal of men's health, 12(6), 1807–1823. doi:10.1177/1557988318798279

67. Issaq HJ, Waybright TJ, Veenstra TD. Cancer biomarker discovery: opportunities and pitfalls in analytical methods. Electrophoresis. 2011;32(9):967–975. doi: 10.1002/elps.201000588.

68. Lebo RV, Kan YW, Cheung MC, Jain SK, Drysdale J (December 1985). "Human ferritin light chain gene sequences mapped to several sorted chromosomes". Hum. Genet. 71(4): 325–8. doi:10.1007/BF00388458. PMID 3000916.

69. Gasparini P, Calvano S, Memeo E, Bisceglia L, Zelante L (Apr 1997). "Assignment of ferritin L gene (FTL) to human chromosome band 19q13.3 by in situ hybridization". Ann. Genet. 40 (4): 227–8. PMID 9526618.

70. Jump up to:a b "FTL ferritin, light polypeptide". National Center for Biotechnology Information. 5 July 2009. Retrieved 20 July2009.

71. Zhang Y, Zhang YL, Feng C, Wu YT, Liu AX, Sheng JZ, Cai J, Huang HF (October 2008). "Comparative proteomic analysis of human placenta derived from assisted reproductive technology". Proteomics. 8 (20): 4344–56. doi:10.1002/pmic.200800294. PMID 18792929.

72. Su Q, Lei T, Zhang M. Association of ferritin with prostate cancer. J BUON. 2017;22(3):766–770.

73. Wang, X., An, P., Zeng, J., Liu, X., Wang, B., Fang, X., … Min, J. (2017). Serum ferritin in combination with prostate-specific antigen improves predictive accuracy for prostate cancer. Oncotarget, 8(11). https://doi.org/10.18632/oncotarget.14977

74. Johansson, B., Pourian, M. R., Chuan, Y.-C., Byman, I., Bergh, A., Pang, S.-T., … Pousette, Å. (2006). Proteomic comparison of prostate cancer cell lines LNCaP-FGC and LNCaP-r reveals heatshock protein 60 as a marker for prostate malignancy. The Prostate, 66(12), 1235–1244. https://doi.org/10.1002/pros.20453

75. Ciocca, D. R., & Calderwood, S. K. (2005). Heat shock proteins in cancer: diagnostic, prognostic, predictive, and treatment implications. Cell stress & chaperones, 10(2), 86–103. doi:10.1379/csc-99r.1

76. National Center for Biotechnology Information (2010, Jan 06). HSPD1 heat shock protein family D (Hsp60) member 1. Retrieved from https://www.ncbi.nlm.nih.gov/gtr/genes/3329/

77. Lee, E., & Lee, D. H. (2017). Emerging roles of protein disulfide isomerase in cancer. BMB Reports, 50(8), 401–410. https://doi.org/10.5483/bmbrep.2017.50.8.107

78. Xu, S., Sankar, S., & Neamati, N. (2014). Protein disulfide isomerase: a promising target for cancer therapy. Drug Discovery Today, 19(3), 222–240. https://doi.org/10.1016/j.drudis.2013.10.017

79. Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., … Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. Cancer Cell, 1(2), 203–209. https://doi.org/10.1016/s1535-6108(02)00030-2

80. Steinberg GD, Carter BS, Beaty TH, Childs B, Walsh PC. Family history and the risk of prostate cancer. Prostate. 1990;17:337–47.

81. Lawrence SM, Huddleston KA, Pitts LR, Nguyen N, Lee YC, Vann WF, Coleman TA, Betenbaugh MJ (Jul 2000). "Cloning and expression of the human N-acetylneuraminic acid phosphate synthase gene with 2-keto-3-deoxy-D-glycero- D-galacto-nononic acid biosynthetic ability". J Biol Chem. 275 (23): 17869 77. doi:10.1074/jbc.M000217200. PMID 10749855.

82. "Entrez Gene: NANS N-acetylneuraminic acid synthase (sialic acid synthase)"

83. Zhang, C., Yan, L., Song, H., Ma, Z., Chen, D., Yang, F., … Xu, Z. (2019). Elevated Serum Sialic Acid Levels Predict Prostate Cancer As Well As Bone Metastases. Journal of Cancer, 10(2), 449–457. https://doi.org/10.7150/jca.27700

84. Christensen, M., Hï¿½gdall, C., Jochumsen, K., & Hïgdall, E. (2017). Annexin A2 and cancer: A systematic review. International Journal of Oncology. https://doi.org/10.3892/ijo.2017.4197

85. Christensen, M.V., Høgdall, C.K., Jochumsen, K.M., & Høgdall, E.V. (2018). Annexin A2 and cancer: A systematic review. International Journal of Oncology, 52, 5-18. https://doi.org/10.3892/ijo.2017.4197

86. Loss of Annexin II Heavy and Light Chains in Prostate Cancer and Its Precursors, Albert Chetcuti, Sienna H. Margan, Peter Russell, Stephen Mann,

Douglas S. Millar, Susan J.Clark, John Rogers, David J. Handelsman and Qihan Dong Cancer Res September 1 2001 (61) (17) 6331-6334

87. Senga, Shogo & Kawaguchi, Koichiro & Kobayashi, Narumi & Ando, Akira & Fujii, Hiroshi. (2018). A novel fatty acid-binding protein 5-estrogen-related receptor α signaling pathway promotes cell growth and energy metabolism in prostate cancer cells. Oncotarget. 9. 10.18632/oncotarget.25878.

88. Tölle, A., Suhail, S., Jung, M., Jung, K., & Stephan, C. (2011). Fatty acid binding proteins (FABPs) in prostate, bladder and kidney cancer cell lines and the use of IL-FABP as survival predictor in patients with renal cell carcinoma. BMC cancer, 11, 302. doi:10.1186/1471-2407-11-302

89. Donato R. S100: a multigenic family of calcium-modulated proteins of the EF-hand type with intracellular and extracellular functional roles. Int J Biochem Cell Biol 2001; 33: 637- 668.

90. Heizmann CW, Fritz G and Schafer BW. S100 proteins: structure, functions and pathology. Front Biosci 2002; 7: d1356-1368.

91. Donato R. Perspectives in S-100 protein biology. Review article. Cell Calcium 1991; 12: 713- 726.

92. Donato R. Functional roles of S100 proteins, calcium-binding proteins of the EF-hand type. Biochim Biophys Acta 1999; 1450: 191-231.

93. Donato R. Intracellular and extracellular roles of S100 proteins. Microsc Res Tech 2003; 60: 540-551.

94. Cancemi, P., Buttacavoli, M., Di Cara, G., Albanese, N. N., Bivona, S., Pucci-Minafra, I., & Feo, S. (2018). A multiomics analysis of S100 protein family in breast cancer. Oncotarget, 9(49), 29064–29081. doi:10.18632/oncotarget.25561

95. Peri, S., Navarro, J. D., Kristiansen, T. Z., Amanchy, R., Surendranath, V., Muthusamy, B., … Pandey, A. (2004). Human protein reference database as a

discovery resource for proteomics. Nucleic acids research, 32(Database issue), D497–D501. doi:10.1093/nar/gkh070

96. Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., … Punta, M. (2014). Pfam: the protein families database. Nucleic acids research, 42(Database issue), D222–D230. doi:10.1093/nar/gkt1223

97. Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., … von Mering, C. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic acids research, 43(Database issue), D447–D452. doi:10.1093/nar/gku1003

98. UniProt Consortium (2008). The universal protein resource (UniProt). Nucleic acids research, 36(Database issue), D190–D195. doi:10.1093/nar/gkm895

99. Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P., & Bork, P. (2000). SMART: a web-based tool for the study of genetically mobile domains. Nucleic acids research, 28(1), 231–234. doi:10.1093/nar/28.1.231

100. Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., … Yeats, C. (2009). InterPro: the integrative protein signature database. Nucleic acids research, 37(Database issue), D211–D215. doi:10.1093/nar/gkn785

101. Lipman, DJ; Pearson, WR (1985). "Rapid and sensitive protein similarity searches". Science. 227 (4693): 1435–41. doi:10.1126/science.2983426. PMID 2983426.

102. Pearson, WR; Lipman, DJ (1988). "Improved tools for biological sequence comparison". Proceedings of the National Academy of Sciences of the United States of America. 85 (8): 2444–8. doi:10.1073/pnas.85.8.2444. PMC 280013. PMID 3162770.

103. David W. Mount: Bioinformatics Sequence and Genome Analysis, Edition 1, Cold Spring Harbor Laboratory Press, 2001, pp. 295–297.

104. Henikoff, S.; Henikoff, J.G. (1992). "Amino Acid Substitution Matrices from Protein Blocks". PNAS. 89 (22): 10915–10919. doi:10.1073/pnas.89.22.10915

105. Wrabl JO, Grishin NV (1 January 2004). "Gaps in structurally similar proteins: towards improvement of multiple sequence alignment". Proteins. 54 (1): 71–87. doi:10.1002/prot.10508. PMID 14705025.

106. Needleman, Saul B. & Wunsch, Christian D. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". Journal of Molecular Biology. 48 (3): 443–53. doi:10.1016/0022-2836(70)90057-4. PMID 5420325.

107. Raghavan Unnithan, S. K., Kannan, B., & Jathavedan, M. (2014). Betweenness Centrality in Some Classes of Graphs. International Journal of Combinatorics, 2014, 112. https://doi.org/10.1155/2014/241723

108. Shao, M., Zhou, S., & Guan, J. (2014). Revisiting topological properties of protein-protein interaction networks from the perspective of dataset evolution. In 2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE. https://doi.org/10.1109/bibm.2014.6999297

109. Ozgür, A., Vu, T., Erkan, G., & Radev, D. R. (2008). Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics (Oxford, England), 24(13), i277–i285. doi:10.1093/bioinformatics/btn182

110. M. E. J. Newman. "The mathematics of networks" (PDF). Retrieved 2006-11-09.

111. Christian F. A. Negre, Uriel N. Morzan, Heidi P. Hendrickson, Rhitankar Pal, George P. Lisi, J. Patrick Loria, Ivan Rivalta, Junming Ho, Victor S. Batista. (2018). "Eigenvector centrality for characterization of protein allosteric pathways". Proceedings of the National Academy of Sciences. 115 (52): E12201--E12208. doi:10.1073/pnas.1810452115.

112. David Austin. "How Google Finds Your Needle in the Web's Haystack". AMS

113. M. E. J. Newman. "The mathematics of networks" (PDF). Retrieved 2006-11-09.

114. Engeln-Müllges, G., & Uhlig, F. (1996). Eigenvalues and Eigenvectors of Matrices. In Numerical Algorithms with C (pp. 155–178). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-61074-5_7

115. Wang J, Tang C, Yang C, Zheng Q, Hou Y. Tropomyosin-1 Functions as a Tumor Suppressor with Respect to Cell Proliferation, Angiogenesis and Metastasis in Renal Cell Carcinoma. J Cancer 2019; 10(10):2220-2228. doi:10.7150/jca.28261. Available from http://www.jcancer.org/v10p2220.htm

116. Pan H, Gu L, Liu B, et al. Tropomyosin-1 acts as a potential tumor suppressor in human oral squamous cell carcinoma. PLoS One. 2017;12(2):e0168900. Published 2017 Feb 9. doi:10.1371/journal.pone.0168900

117. Wang J, Guan J, Lu Z, Jin J, Cai Y, Wang C, et al. Clinical and tumor significance of tropomyosin-1 expression levels in renal cell carcinoma. Oncology reports. 2015;33(3):1326–34. 10.3892/or.2015.3733

118. Li DQ, Wang L, Fei F, Hou YF, Luo JM, Zeng R, et al. Identification of breast cancer metastasis-associated proteins in an isogenic tumor metastasis model using two-dimensional gel electrophoresis and liquid chromatography-ion trap-mass spectrometry. Proteomics. 2006;6(11):3352–68. 10.1002/pmic.200500617

119. Langer W, Sohler F, Leder G, Beckmann G, Seidel H, Grone J, et al. Exon array analysis using re-defined probe sets results in reliable identification of alternatively spliced genes in non-small cell lung cancer. BMC genomics. 2010;11:676 10.1186/1471-2164-11-676

120. Chen C, Zhang LG, Liu J, Han H, Chen N, Yao AL, et al. Bioinformatics analysis of differentially expressed proteins in prostate cancer based on proteomics data. OncoTargets and therapy. 2016;9:1545–57. 10.2147

/OTT.S98807

121. Khori V, Amani Shalamzari S, Isanejad A, Alizadeh AM, Alizadeh S, Khodayari S, et al. Effects of exercise training together with tamoxifen in reducing mammary tumor burden in mice: Possible underlying pathway of miR-21. European journal of pharmacology. 2015;765:179–87. 10.1016/j.ejphar.

2015.08.031

122. Hu J, Ho AL, Yuan L, Hu B, Hua S, Hwang SS, et al. From the Cover: Neutralization of terminal differentiation in gliomagenesis. Proceedings of the National Academy of Sciences of the United States of America. 2013;110

(36):14520–7. 10.1073/pnas.1308610110

123. Dube S, Yalamanchili S, Lachant J, Abbott L, Benz P, Mitschow C, et al. Expression of Tropomyosin 1 Gene Isoforms in Human Breast Cancer Cell Lines. International journal of breast cancer. 2015;2015:859427 10.1155/2015/859427

124. Ali S, Almhanna K, Chen W, Philip PA, Sarkar FH. Differentially expressed miRNAs in the plasma may provide a molecular signature for aggressive pancreatic cancer. American journal of translational research. 2010;3(1):28–47.

125. He QY, Chen J, Kung HF, Yuen AP, Chiu JF. Identification of tumor-associated proteins in oral tongue squamous cell carcinoma by proteomics. Proteomics. 2004;4(1):271–8. 10.1002/pmic.200300550

126. Kosuke Mima, Reiko Nishihara, Juhong Yang, Ruoxu Dou, Yohei Masugi, Yan Shi, Annacarolina da Silva, Yin Cao, Mingyang Song, Jonathan Nowak, Mancang Gu, Wanwan Li, Teppei Morikawa, Xuehong Zhang, Kana

Wu, Hideo Baba, Edward L. Giovannucci, Jeffrey A. Meyerhardt, Andrew T. Chan, Charles S. Fuchs, Zhi Rong Qian and Shuji Ogino Clin Cancer Res August 1 2016 (22) (15) 3841-3848; DOI: 10.1158/1078-0432.CCR-15-2173

127. Freitas-Alves, D., Vieira-Monteiro, H., Piranda, D., Sobral-Leite, M., da Silva, T., Bergmann, A., Valença, S., Perini, J., & Vianna-Jorge, R. (2018). PTGS2 polymorphism rs689466 favors breast cancer recurrence in obese patients, Endocrine-Related Cancer, 25(3), 351-365. Retrieved May 25, 2019, from https://erc.bioscientifica.com/view/journals/erc/25/3/ERC-17-0374.xml

128. Cebrián, A. et al. Functional PTGS2 polymorphism-based models as novel predictive markers in metastatic renal cell carcinoma patients receiving first-line sunitinib. Sci. Rep. 7, 41371; doi: 10.1038/srep41371 (2017).

129. Chen, Hsin-An & Chang, Yi-Wen & Tseng, Chi-Feng & Chiu, Ching-Feng & Hong, Chih-Chen & Wang, Weu & Wang, Ming-Yang & Hsiao, Michael & Ma, Jui-Ti & Chen, Chung-Hsing & Jiang, Shih Sheng & Wu, Chih-Hsiung & Hung, Mien-Chie & Huang, Ming-Te & Su, Jen-Liang. (2014). E1A-Mediated Inhibition of HSPA5 Suppresses Cell Migration and Invasion in Triple-Negative Breast Cancer. Annals of surgical oncology. 22. 10.1245/s10434-014-4061-3.

130. Guest, S. T., Kratche, Z. R., Bollig-Fischer, A., Haddad, R., & Ethier, S. P. (2015). Two members of the TRiC chaperonin complex, CCT2 and TCP1 are essential for survival of breast cancer cells and are linked to driving oncogenes. Experimental Cell Research, 332(2), 223–235. https://doi.org/10.1016/

j.yexcr.2015.02.005

131. Carr AC, Khaled AS, Bassiouni R, et al. Targeting chaperonin containing TCP1 (CCT) as a molecular therapeutic for small cell lung cancer. Oncotarget. 2017;8(66):110273–110288. Published 2017 Nov 25. doi:10.18632/ oncotarget.22681

132. Devarajan E, Sahin AA, Chen JS, Krishnamurthy RR, Aggarwal N, Brun AM. et al. Down-regulation of caspase 3 in breast cancer: A possible mechanism

for chemoresistance. Oncogene. 2002; 21(57):8843–51. doi: 10.1038/sj.onc.1206044.

133. Hu Q, Peng J, Liu W, et al. Elevated cleaved caspase-3 is associated with shortened overall survival in several cancer types. Int J Clin Exp Pathol. 2014;7(8):5057–5070. Published 2014 Jul 15

134. Brown MF, Leibowitz BJ, Chen D, et al. Loss of caspase-3 sensitizes colon cancer cells to genotoxic stress via RIP1-dependent necrosis. Cell Death Dis. 2015;6(4):e1729. Published 2015 Apr 23. doi:10.1038/cddis.2015.104

135. Wei MC, Zong WX, Cheng EH, et al. Proapoptotic BAX and BAK: a requisite gateway to mitochondrial dysfunction and death. Science. 2001;292:727-730. PMID: 11326099

136. Hanada M, Delia D, Aiello A, Stadtmauer E, Reed JC. bcl-2 gene hypomethylation and high-level expression in B-cell chronic lymphocytic leukemia. Blood. 1993;82:1820-1828. PMID: 8104532

137. Gala JL, Vermylen C, Cornu G, et al. High expression of bcl-2 is the rule in acute lymphoblastic leukemia, except in Burkitt subtype at presentation, and is not correlated with the prognosis. Ann Hematol. 1994;69:17-24. PMID: 8061103

138. Searle CJ, Brock IW, Cross SS, Balasubramanian SP, Reed MW, Cox A. A BCL2 promoter polymorphism rs2279115 is not associated with BCL2 protein expression or patient survival in breast cancer patients. Springerplus. 2012;1:38. doi: 10.1186/2193-1801-1-38

139. Renner W, Langsenlehner U, Krenn-Pilko S, Eder P, Langsenlehner T. BCL2 genotypes and prostate cancer survival. BCL2-Genotypen und Überleben bei Prostatakrebs. Strahlenther Onkol. 2017;193(6):466–471. doi:10.1007/s00066-017-1126-9

140. Chan, J. J., Kwok, Z. H., Chew, X. H., Zhang, B., Liu, C., Soong, T. W., … Tay, Y. (2018). A FTH1 gene:pseudogene:microRNA network regulates

tumorigenesis in prostate cancer. Nucleic acids research, 46(4), 1998–2011. doi:10.1093/nar/gkx1248

141. Huang, H., Qiu, Y., Huang, G., Zhou, X., Zhou, X., & Luo, W. (2019). Value of Ferritin Heavy Chain (FTH1) Expression in Diagnosis and Prognosis of Renal Cell Carcinoma. Medical Science Monitor, 25, 3700–3715. https://doi.org/10.12659/msm.914162

142. Gemoll, T., Epping, F., Heinrich, L., Fritzsche, B., Roblick, U. J., Szymczak, S., … Habermann, J. K. (). Increased cathepsin D protein expression is a biomarker for osteosarcomas, pulmonary metastases and other bone malignancies. Oncotarget, 6(18), 16517–16526. doi:10.18632/oncotarget.4140

143. Rubio-Perez, C., Tamborero, D., Schroeder, MP., Antolín, AA., Deu-Pons,J., Perez-Llamas, C., Mestres, J., Gonzalez-Perez, A., Lopez-Bigas, N. In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals novel targeting opportunities. Cancer Cell 27 (2015), pp. 382-396

144. Folkman J. Tumor angiogenesis: therapeutic implications. The New England journal of medicine. 1971;285:1182–1186.

145. Qin L, Wu YL, Toneff MJ, Li D, Liao L, Gao X, Bane FT, Tien JC, Xu Y, Feng Z, Yang Z, Theissen SM, Li Y, Young L, Xu J. NCOA1 Directly Targets M-CSF1 Expression to Promote Breast Cancer Metastasis. Cancer research. 2014;74:3477–3488.

146. Wang S, Yuan Y, Liao L, Kuang SQ, Tien JC, O'Malley BW, Xu J. Disruption of the SRC-1 gene in mice suppresses breast cancer metastasis without affecting primary tumor formation. Proceedings of the National Academy of Sciences of the United States of America. 2009;106:151–156.

147. Qin L, Wu YL, Toneff MJ, et al. NCOA1 Directly Targets M-CSF1 Expression to Promote Breast Cancer Metastasis. Cancer Res. 2014;74(13):3477–3488. doi:10.1158/0008-5472.CAN-13-2639

148. Takeda K, Hara N, Nishiyama T, Tasaki M, Ishizaki F, Tomita Y. Corepressive function of nuclear receptor coactivator 2 in androgen receptor of prostate cancer cells treated with antiandrogen. BMC Cancer. 2016;16:332. Published 2016 May 25. doi:10.1186/s12885-016-2378-y

149. Kathryn A. O'Donnell, Vincent W. Keng, Brian York, Erin L. Reineke, Daekwan Seo, Danhua Fan, Kevin A. T. Silverstein, Christina T. Schrum, Wei Rose Xie, Loris Mularoni, Sarah J. Wheelan, Michael S. Torbenson, Bert W. O'Malley, David A. Largaespada, Jef D. Boeke Proceedings of the National Academy of Sciences May 2012, 109 (21) E1377-E1386; DOI: 10.1073/pnas.1115433109

150. Sasca, D., Szybinski, J., Schüler, A., Shah, V., Heidelberger, J., Haehnel, P. S., … Kindler, T. (2019). NCAM1 (CD56) promotes leukemogenesis and confers drug resistance in AML. Blood, 133(21), 2305–2319. https://doi.org/10.1182/blood-2018-12-889725