Master's Dissertation

BIOINFORMATICS RESEARCH ON THE INVESTIGATION OF TARGET GENES FOR BREAST CANCER

바이오인포매틱스 기법을 활용한 유방암
타겟 유전자 탐색 연구

February 2015

Mikyung Je

Lab. of Computational Biology and Bioinformatics
Graduate School of Public Health
Seoul National University

ABSTRACT

Bioinformatics Research on the Investigation of Target Genes for Breast Cancer

Mikyung Je

Laboratory of Computational Biology and Bioinformatics
Graduated School of Public Health
Seoul National University

Cancer remains a worldwide problem, and so requires solutions. Research on cancer has been progressed extensively, but no perfect solution exists, despite the numerous diagnosis and treatment methods. Biomarkers are being used as a part of solutions against cancer, and research to discover new markers is being proceeded using many methods. In such research, bioinformatics methods can be used to find specific targets, with networks and gene ontology analysis relating to characteristics of cancer. Cancer has different features, and needs different approaches pathologically; thus, the present research was performed on breast cancer alone. First, database was constructed specializing in cancer and biomarkers for efficient management of data and for a firm foundation of analysis. Data required for the database construction were brought from many public repository such as NCBI, NCI, and UniProt. These data were processed by Html, Java, JavaScript, MySQL and JSP as several programming languages to fit their purposes. Second, cancer network was formed by utilizing the collected data. It was produced by mapping the cancer gene ID in protein interaction data, and inspected for power-law distribution of degree to see if the produced network has a scale-free nature. The cancer-

specific network produced showed such a distribution, so the existence of the hub was ascertained, and reflecting on the characteristics of the graph determined the rank of hub genes. Accordingly, by calculating the network parameters such as degree distribution, betweenness centrality, and stress centrality, the top 150 hub genes were obtained. The following analysis filtered for specific GO terms reflecting on the characteristics of breast cancer through GO enrichment analysis utilizing 2,902 genes forming cancer network and breast cancer–related genes. By adjusting these specific GO terms to 2,902 genes of the cancer network, a gene list was obtained that is involved in general cancer with has breast cancer-specific GO terms. In the last step of the analysis, target genes were obtained through Venn diagram analysis by PPI network with GO enrichment analysis and target candidates, and results for five clinical markers were referred to. As a result, 34 genes with similar traits as clinical markers in PPI network and GO were selected as targets, and they are anticipated to be utilized primarily in selection of the next breast cancer marker. In addition, if such methods could filter candidates in selecting markers, they would be experimentally and clinically useful.

Keywords: Bioinformatics, Breast cancer, Database, Gene ontology, Marker, PPI network, Target

Student ID: 2013-21873

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABL	Abelson murine leukemia viral oncogene homolog
APC	Adenomatous Polyposis Coli
BCR	Breakpoint Cluster Region
BLAST	Basic Local Alignment Search Tool
BP	Biological Process
CA125	Cancer Antigen 125
CC	Cellular Component
CDK	Cyclin Dependent Kinase
CEA	Carcinoembryonic Antigen
DCC	Deleted in Colorectal Cancer
DEP	Differentially Expressed Protein
E3 ligase	E3 ubiquitin ligase
EBV	Epstein-barr Virus
ECM	Extracellular Matrix
EDRN	Early Detection Research Network
EGF	Epidermal Growth Factor
EGFR	Epidermal Growth Factor Receptor
EGR-1	Early Growth Response protein 1
FASTA	Fast-All
FDA	Food and Drug Administration
GAP	GTPase-activating Protein
GDP	Guanosine Diphosphate
GEF	Guanine nucleotide-exchange Factor
GO	Gene Ontology
GTP	Guanosine Triphosphate
HBV	Hepatitis B Virus
HDAC	Histone Deacetylation
HE4	Human Epididymis protein 4
HIV	Human Immunodeficiency Virus
HPV	Human Paillomavirus
HTML	Hypertext Markup Language
HUGO	Human Genome Organisation
JDBC	Java Database Connectivity

JSP	Java Server Page
MDM2	Mouse Double Minute 2 homolog
MF	Molecular Function
MKK4	Mitogen-activated protein Kinase Kinase 4
MYC	Myelocytomatosis
MySQL	My Sequel
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
NF1	Neurofibromatosis type 1
NM23	Nucleoside diphosphate kinase A, also known as NME1
PDGF	Platelet-derived Growth Factor
PPI	Protein-protein Interaction
R point	Restriction point
RAS	Rat Sarcoma
RAS-GRF	Ras-specific Guanine nucleotide-releasing Factor
RB	Retinoblastoma protein
RTK	Receptor-tyrosine Kinase
SH2	Src Homology 2
SOS	Son Of Sevenless
SRC	Sarcoma
SSV	Simian Sarcoma Virus
Tcf-4	Transcription Factor 4
TP53	Tumor Protein 53
WT1	Wilms Tumor 1

CHAPTER I.

INTRODUCTION

1.1 Overview of cancer

Cancer is a major problem worldwide and accounts for one-quarter of deaths in the U.S. (Siegel et al., 2013). The situation in South Korea is no different: according to the National Cancer Information Center, the death rate from cancer declined 2.7% from 2002 to 2010, yet the incidence rate of cancer was 3.3% annually from 1999 to 2010 (Jung et al., 2013). Though the major cancers are different among races or nations, the core mechanism through which cancer cells form and lead to death are similar. Cancer is a disease where the expression and biochemical functions of genes are inhibited by genetic dysregulations from DNA recombination and point mutations. Therefore, we need to understand the effect and mechanisms of the variation and other regulatory factors of genes (Bishop, 1987). Cancer is a genetic disease in which the accumulation of genetic damage, activation of oncogenes and deactivation of tumor suppressor genes causes normal cells to become malignant cells (Anderson et al., 1992; Vogelstein and Kinzler, 2004). Therefore, it is important that the concepts, general features, and roles of proto-oncogenes, oncogenes, and tumor suppressor genes are understood in carcinogenesis. Importantly, genes that accumulate mutations and genes that help cancer cells metastasize need to be covered (Yokota, 2000). In this chapter, the general characteristics of cancer, the roll of cell cycle in cancer, the oncoviruses, and cancer research in epigenetics will be described.

1.1.1 General characteristics of cancer

Two major differences differentiate cancer from other genetic diseases. One is that cancer is often caused by mutation in somatic cells, and the other is that cancer involves the accumulation of mutations (Vogelstein and Kinzler, 1993). Cancer cells have numerous different features compared to normal cells, including the potential for infinite proliferation and cell divisions, and that can be metastasize to other locations in the body. These features may not be applicable to all cancers, but they are general features that can be applied to many cases (Hanahan and Weinberg, 2000). See Table 1.1. The first feature of cancer cells is self-proliferation that does not rely on growth factor signaling. ECM (Extracellular Matrix) refers to the structure surrounding cells within the body to which they attach; changes in the material and structure affect the processes of metastasis and division (Watt, 1986). Cells grow by growth factor signaling mediated by cell-cell communication. However, if a mutation occurs to a gene or receptor coding a growth regulator, cancer cells are likely to form (Lukashev and Werb, 1998). The second feature of cancer cells is that they are not regulated by the signaling of tumor suppressor genes such as TP53. If a dysregulation occurs to differentiation and proliferation signaling in cases such as DNA damage, p53 mediated the repair mechanism is activated to start repair of the cells. Failure to repair can occur, causing p53 to induce apoptosis of the cells to prevent other mutations (Yonish-Rouach et al., 1991). However, because p53 is deactivated by mutation in tumor cells, they are unable to cause apoptosis; this leads to transition into a malignant tumor by increasing genetic instability, accumulation of mutations, and the possibility of genetic rearrangements. The loss of apoptotic function by mutated p53 is known to contribute significantly to the survival of tumor cells (Lane, 1992). The third feature of cancer cells is the evasion of the normal apoptosis process. Apoptosis performs a core role in the elimination of cells

through normal embryonic development and in healthy adults. Additionally, apoptosis occurs naturally in untreated malignant tumors, and seems to have an effect on the degeneration of certain types of tumors. It shows that apoptosis could affect both the biological degeneration and atrophy of various tissues and organs, and that this mechanism can be induced by harmful changes within the body (Kerr et al., 1972). Multicellular animals control the number of cells by maintaining a balance between cell proliferation and apoptosis; tumor cells are resistant to apoptosis, the resistance is causing various diseases including cancer (Thompson, 1995). Therefore, the apoptosis mechanism without protectional functions such as cell removal is the main cause of disease (Vaux and Korsmeyer, 1999). The fourth feature of cancer cells is infinite replication. A specialized functional structure called the telomere, which has short repeating DNA sequences, exists in the terminal part of eukaryotic chromosomes. Telomeres are fundamental to the duplication and stabilization of chromosomes (Blackburn and Szostak, 1984). The enzyme telomerase interferes in the length maintenance and duplication of the telomere (Morin, 1989). Normal somatic cells lose telomeric DNA sequences through incomplete duplication or other erosions of chromosome terminals, and at a certain point, go through the aging process due to a lack of telomerase activity. However, in immortalized cells such as cancer cells, the telomere is maintained by activation of telomerase to evade the aging process (Harley, 1991). Telomerase activation during immortalization is confirmed by the result of an experiment showing that telomerase is suppressed in most normal cells in tissues, but reactivated in most cells in malignant tumors (Kim et al., 1994). Therefore, for cancer cells to be immortal, the length of telomeres must be maintained, and the reactivation of the telomerase is crucial in this process (Bryan et al., 1995). The fifth feature of cancer cells is the ability to induce angiogenesis. Blood vessels are composed of endothelial cells connected together to network and maintain the flow of blood throughout

tissues. When an organism matures, new vessels are only formed through angiogenesis, a process controlled by pathological situations such as the curing of wounds and growth of tumors. The growth of new blood vessels is particularly fundamental in the growth and metastasis of tumors (Hanahan and Folkman, 1996): if the proper construction of vasculature is interrupted, the tumor cells either undergo necrosis (Brem et al., 1976) or apoptosis (Parangi et al., 1996). If angiogenesis is insufficient, a tumor is not supplied with the nutrition necessary for survival, and is unable to grow to a certain extent or undergo metastasis (Carmeliet and Jain, 2000). The sixth feature of cancer cells is metastasis. Unlike normal cells that do not invade non-resident tissues in the body, cancer cells are able to invade and spread to distant tissues from a primary tumor source. While most facts regarding the development of metastasis for cancer cells are uncertain, the basic process can be understood. Metastasis is composed of consecutive processes such as formation of primary tumors, proliferation / angiogenesis, separation / invasion, circulation, extravasation, and colonization (Fidler, 2003). Metastasis accounts for 90% of deaths caused by cancer occurs when tumor cells adjust to the delicate environment of tissues far from the primary tumor; these adjustments include being able to selectively choose features advantageous for overcoming resistance to the environment, gained through heterogeneity of cancer cells that have lost control of genetic stability (Gupta and Massague, 2006).

1.1.2 Cell cycle and cancer

The general features of cancer mentioned above describe the uncontrolled proliferation cells. They gain the ability for unregulated cell proliferation through damage to the genes controlling their life cycle (Sherr, 1996). Therefore, the relationship between cancer and the cell cycle is important. The cell cycle refers to the process of cell duplication through

growth and division; it comprises multiple phases: G1, S, G2, and M. At a certain point in the cell cycle, protein complexes of cyclin and CDKs (Cyclin Dependent Kinases) exist that have important roles in the various different processes of the cell cycle (Morgan, 1997). First, the cyclin D / CDK 4, 6 complex controls the cell cycle during G1, and at R point (Restriction point) at the end of G1, this complex decides whether to continue the cell cycle or shift to interphase (G0). The cyclin E / CDK 2 complex serves an important role in shifting from G1 to S phase if the R point has been passed. Afterwards, cyclin A forms a complex with CDKs 1 and 2 to control completion of S phase and G2 phase. Lastly, if M phase starts, the cyclin B / CDC 2 (CDK 1) complex induces mitosis (Pardee, 1974; Nurse, 2000; Hochegger et al., 2008). A checkpoint exists as a biochemical pathway to control mechanisms such as DNA replication and separation of chromosomes by regulating the timing of each phase. The G1 checkpoint prevents the replication of damaged DNA in S phase by halting the cell cycle. The G2 checkpoint serves to inspect the replication of S phase, and the M phase checkpoint ensures that adherence of chromosomes to the spindle fibers during mitosis was completed normally. If there are any problems with checkpoint functions, genetic stability is inhibited, thus increasing the possibility of mutations that contribute to cancer (Hartwell and Weinert, 1989; Hartwell and Kastan, 1994; Elledge, 1996). This process is shown in Figure 1.1.

1.1.3 Virus and cancer

Reviewing the general properties of cancer, another important relationship exists between viruses and cancer. Oncoviruses with DNA or RNA molecules manifest continuously in transformed cells to maintain this altered feature in offspring cells (Martin, 1970). The diversity of proteins coded by tumor viruses can induce malignant mechanisms in the nucleus,

cytoplasm, and cell membrane (Bishop, 1985). Additionally, tumor viruses indirectly induce immunosuppressive reactions, and can directly transform protein expression in host cells to contribute to the development of tumors within the body (Hausen, 1991). Tumor viruses include HPV (Human Papillomavirus), HBV (Hepatitis B Virus), EBV (Epstein-Barr Virus), and HIV (Human Immunodeficiency Virus); we will briefly take an in-depth look into EBV and HIV next. EBV was first discovered by electron microscopy of cells cultured from Burkitt lymphoma (Epstein et al., 1964). A gene encoded by EBV induces the transformation of B cells by altering gene transcription and activating the cells' signaling pathway structurally. Additionally, EBV utilizes the biological features of normal B cell division to avoid elimination by the host immune system. Transplant patients could be vulnerable to EBV infection by proliferation of transformed B cells. Likewise, EBV infection affects the etiology of various other lymphocytes and epithelial in malignant tumors (Young and Rickinson, 2004). It was also revealed that HIV, another important tumor virus, tends to increase the risk of cervical legions (Chirenje, 2004), and Hodgkin's lymphoma in people with HIV / AIDS (Bigger et al., 2006).

1.1.4 Cancer in epigenetics

Taking a look at recent cancer research trends, there is an abundance of research from the perspective of epigenetics. This starts from the recognition that mutations of various genetic information contained within cells cannot be completely explained at the level of the DNA sequence. Cancer epigenetics mainly deals with the mechanics of DNA methylation, genomic imprinting, and histone modification and their effects on cancer (Finberg and Tycko, 2004). DNA methylation (5-methylcytosine) in the promoter of mammals inhibits transcription and plays an important role in gene activation (Jones and

Laird, 1999). Promotor-associated CpG islands in special cases such as X-chromosome inactivation and gene silencing suppresses expression. From this, we are able to appreciate that DNA methylation plays an important role in the control of gene expression, and this mechanism is important in cancer as well (Birs, 2002; Jones and Baylin, 2002). Hypomethylation and hypermethylation occur in cancer-related DNA methylation. The former affects the overexpression of oncogenes and genetic stability, while the latter affects the suppression of expression of tumor suppressor genes to contribute to the formation of tumors (Ehrlich, 2002). Histone deacetylation and histone methylation contribute to abnormal gene silencing mechanisms of specific tumor suppressor genes in tumor cells with aberrant DNA methylation (Bachman et al., 2003; Esteller, 2007). Altogether, gene silencing is controlled by the interactions between DNA methylation, histone modification, and nucleosomal remodeling, with anomalies in this process causing diseases such as cancer (Jones and Baylin, 2007).

1.2 Oncogenes and proto-oncogenes

Oncogenes and proto-oncogenes are genes involved in the process of carcinogenesis. In this section, we will first cover the general features of oncogenes, and specifically the most well-known oncogenes: the protein kinase and Ras gene families. For normally functioning genes to undergo carcinogenesis, accumulated mutation needs to occur to transform them into oncogenes. A simple diagram is shown in Figure 1.2. Normal genes with the potential to transform into oncogenes are called proto-oncogenes, and the proteins that they code for normally perform functions related to the control of cell growth and the cell cycle. The mechanism of activation for a proto-oncogene to become an oncogene can be divided into three parts: mutation, gene amplification, and gene rearrangement. These three mechanisms bring about changes in the structure or expressed level of the oncogenes, and are activated into oncogenes by increasing the survival rate of cells and providing a growth advantage. The activation of oncogenes through mutation in the coding region brings about changes such as an increase in the protein's activation, while gene amplification leads to the overexpression of the normal protein. Additionally, if rearrangement such as inversion or translocation of the chromosome occurs, normal proteins could be overexpressed or abnormal recombined proteins could induce the activation of oncogenes, ultimately leading to carcinogenesis. The mechanisms of action for oncogenes are shown in Figure 1.3 (Carlo, 2008; Alberts et al., 2009).

1.2.1 Features of oncogenes

Proto-oncogenes function normally until they are activated as oncogenes. Therefore, we must discuss how normal proto-oncogenes function before transformation and activation. First, we will explore oncogenes

mutated from growth factors. V-sis is a viral oncogene extracted from SSV (Simian Sarcoma Virus). Comparing the sequence of the p28sis protein created by this gene and PDGF (Platelet-derived Growth Factor) showed they have high homology. This revealed that v-sis is a result from the recombination of PDGF coding gene and virus (Doolittle et al., 1983). The signal secreted by PDGF activated in cancer cells affects the proliferation of the clone before expression of the tumor. Additionally, if consistent genetic changes are made, PDGF contributes to the formation of the tumor by creating an unstable cell environment (Andrae et al., 2008). Growth factor genes such as PDGF may lead to various diseases such as cancer (Aaronson, 1991). Second, oncogenes can be transformed growth factor receptors. As previously mentioned, cancer cells are able to grow without relying on growth factor signals. Many growth factors and growth factor receptors are involved, and we will more closely explore EGFR (Epidermal Growth Factor Receptor). EGFR is a receptor for EGF (Epidermal Growth Factor) that is a tyrosine kinase among the EGF receptor family (ErbB family), including ErbB-2, ErbB-3, and ErbB-4 (Salomon et al., 1995). The ErbB receptors and associated ligands control intercellular functions such as growth, differentiation, survival and angiogenesis that take part in the mechanism of tumor progression (Normanno et al., 2006). Third, oncogenes can be mutated from the signal transducer. To send the mitosis signal from the growth factor receptor to the nucleus of cells, a complicated mechanism is required called signal transduction. Signal transducers functioning in this pathway can be separated into two groups of non-receptor protein kinases and GTP (Guanosine Triphosphate)-binding proteins. Non-receptor protein kinases are divided into subgroups of tyrosine kinases and serine / threonine kinases, while GTP-binding proteins are categorized into H-ras, K-ras, N-ras and Gsp, Gip according to their chemical features. If signal transducers are mutated into oncogenes, they induce irregular intercellular activities, which can cause

uncontrolled cell proliferation (Sevik, 2012). For instance, Src (Sarcoma), a tyrosine kinase, controls functions of cells such as adherence, invasion and movement; these functions can contribute to the process and metastasis of the tumor (Yeatman, 2004). Fourth, oncogenes can arise from mutations of transcription factors. APC (Adenomatous Polyposis Coli) is a tumor suppressor gene that combines with cadherin to affect cell adherence, and combines with Tcf-4 (Transcription factor 4) to induce the decomposition of β-catenin, which controls gene transcription (Munemitus et al., 1995; Morin et al., 1997). In most colorectal cancers, deactivation of APC by mutation is found, inducing the abnormal accumulation of β-catenin. This increases the transcriptional activation of genes, and ultimately contributes to tumor formation. In the abnormal APC pathway, c-Myc (cellular-Myc (Myelocytomatosis)), a transcription factor control gene, is overexpressed and plays an important role in the detection of abnormal activation of the APC pathway (He et al., 1998). Table 1.2 has categorized proto-oncogenes by their normally performed functions with examples. Two pivotal oncogenes mentioned above (protein kinase and Ras) will be explored in more detail.

1.2.2 Representative oncogene: protein kinase

We will first review the protein kinase gene family, the core proteins of which participate in the signal transduction pathway and are found in mutated states in various diseases such as cancer (Schlessinger, 2000). The cell growth and mutation mechanisms used by mammals to interpret signal transduction are complicated, but some genes are known that control this process. Among them is the protein kinase gene family, one of the biggest gene families in eukaryotes. They play an important role in cell cycle control and signal transduction through phosphorylation. For instance, tyrosine protein phosphorylation controls the bonding of related proteins within cells by

producing specific combinations of parts of the protein including the SH2 (Src Homology 2) domain. The interaction within molecules of the phosphorylated tyrosine and the SH2 domain may also interfere with control of enzyme activation (Cantley et al., 1991). An example of a protein kinase is the Abl gene: Abl produces a non-receptor tyrosine kinase within the nucleus. Abl (Abelson murine leukemia viral oncogene homolog) exists on chromosome 9 and is activated as an oncogene through the formation of the Bcr-Abl fusion by gene translocation with BCR (Breakpoint Cluster Region) on chromosome 22, leading to strong phosphorylation activation (Jensen and Hunter, 2001). Normally Abl induces the apoptosis of cells when their DNA is damaged; since the Bcr-Abl fusion exists within the cytoplasm, not the nucleus, the cell is able to evade DNA damage-directed apoptosis by this mechanisms. This mutation is common in acute leukemia (Vigneri and Wang, 2001). The protein kinase gene family exists in the cell membrane as either a transmembrane receptor or a signal transducer, or in the nucleus to interfere with control mechanisms of processes such as transcription, cell cycle, apoptosis, and differentiation. It controls signal transduction in eukaryotic cells and interferes with gene expression (Manning et al., 2002).

1.2.3 Representative oncogene: Ras

The next crucial oncogene is the Ras (Rat sarcoma) family. The Ras family is activated as oncogenes through point mutations arising in specific codons; these mutations are discovered in various tumors such as colorectal cancer, lung cancer, and pancreatic cancer. This gene family codes similar proteins and include functional components of H-ras, K-ras and N-ras. Ras protein plays a pivotal role in stimulation and signal transduction of factors associated with cell growth and differentiation (Barbacis, 1987; Bos, 1989). Ras is also important in the signal transduction system: it is activated by

connecting to GTP and deactivated by attaching to GDP (Guanosine Diphosphate). Stimulation of Ras occurs via signaling mediated by RTK (Receptor-tyrosine Kinase), and this stimulation or signal activates GEF (Guanine nucleotide-exchange factors) such as SOS (Son Of Sevenliss) 1/2 and Ras-GRF (Ras-specific Guanine nucleotide-releasing Factor). Activated GEF stimulates Ras-GDF to induce them to form Ras-GTP. The activation signal is terminated when Ras-GTP is hydrolyzed into Ras-GDP through the GTP hydrolysis catalyst function of Ras-GAP (GTPase-activating Proteins). Suppressor genes such as SPROUTY and the ACK protein family delay the activation of SOS1/2 proteins to inhibit Ras activation mediated by RTK (Malumbres and Barbacid, 2003). The activation of Ras is controlled by complex processes; if mutation occurs to a specific codon of this gene (12, 61, and rarely 13 amino acids), the GTPase function of the Ras protein either decreases or is exterminated. Then, activated Ras-GTP does not transform into deactivated Ras-GDP and the mutated phenotype is continuous activation (Scheffzek et al., 1997). Ras interferes with various signaling pathways and is a pivotal oncogene necessary for the generation and maintenance of tumor cells (Chin et al, 1999).

1.3 Tumor suppressor genes (TSGs)

The tumor suppressor gene is another important factor participating in the carcinogenic mechanism by functioning as an antiproliferative signal transducer biochemically; biologically, they play an important role in the proliferation, differentiation, and apoptosis of cells. If mutation of a tumor suppressor gene occurs, this disables functions such as maintaining the number of cells, leading to diseases such as cancer (Weinberg, 1991). In this chapter, we will take a look at the general features of tumor suppressor genes and two the representative tumor suppressor genes (Rb and TP53). Tumor suppressor genes show different behavior from oncogenes. Oncogenes appear to be dominant, able to form tumors only if either of the alleles mutate; tumor suppressor genes appear to be recessive, able to form tumors only when both the alleles mutate. Even if one tumor suppressor gene is damaged, the other is able to produce normal protein. This is known as the 'two-hit hypothesis', which explains that if there is a genetic mutation, the tumor can only be induced through two mutations (Knudson, 1971). However, this hypothesis does not always apply to tumor formation by all tumor suppressor genes. Mutation in TP53 could function as a dominant negative, meaning that the protein produced by a mutated TP53 could inhibit the function of a normal protein from a non-mutated allele (Baker et al., 1990). Though there are exceptions, the two-hit hypothesis is important in explaining tumor formation involving tumor suppressor genes. Figure 1.4 and 1.5, respectively, explain the two-hit hypothesis and the differences in activation mechanisms of oncogenes and tumor suppressor genes.

1.3.1 Features of TSGs

The maintenance of the number of cells plays an important role in survival, and tumor suppressor gene are controlled by proliferation, differentiation, and apoptosis of cells. Tumor suppressor genes are considered important in cancer because they are involved in functions such as cell cycle control, DNA damage detection and repair, degradation of protein, cell differentiation, and tumor angiogenesis (Sherr, 2004). Tumor suppressor genes can be categorized according to their functions: Table 1.3 shows examples of tumor suppressor genes and the functions of their respective proteins. First, tumor suppressor genes related to the cell cycle. $p16^{INK4a}$ combines with CDK4 to inhibit the activation of the CDK4 / cyclin D complex; it can also combine with Rb to function as a negative regulator of proliferation in normal cells. $p16^{INK4a}$ produces a negative feedback loop to control Rb with the suppression of proliferation (Serrano et al., 1993). Suppression of proliferation occurs when $p16^{INK4a}$ is overexpressed, inhibiting the phosphorylation of Rb and the function of the CDK4 / cyclin D complex to stop the cell cycle at the end of G1 phase (Lukas et al., 1995). To prevent indiscrete proliferation of cells, the functions of tumor suppressor genes $p16^{INK4a}$ and Rb are important. However, in many cases these two genes are deactivated by mutations in tumor cells and unable to suppress overproliferation of cells. Second, tumor suppressor genes that interfere with signal transduction. The genetic product of NF1 (Neurofibromatosis type 1), one of the genes causing neurofibromatosis type 1, causes GAP activation in the Ras signal transduction pathway. Protein produced by NF1 is fundamental for the accurate negative regulation of Ras of cells and enables the suppression of cell proliferation. This phenomenon can also be observed in the abnormal activation of Ras in tumor cells of malignant neurofibromatosis (Basu et al., 1992). Third, tumor suppressor genes can be involved in

transcription. WT1 (Wilms Tumor 1) is a causal gene of the Wilms tumor and codes four zinc finger proteins that combine with the DNA of specific sequences such as EGR-1 (Early Growth Response protein 1). If WT1 does combine with specific DNA, it functions as a transcriptional suppressor in the NH_2-terminus abundant with glutamine and proline. In tumor cells, the DNA binding domain of WT1 is deactivated, and could cause transcriptional suppression by WT1 (Madden et al., 1991). Fourth, tumor suppressor genes involved in cell adherence and metastasis. Problems with cell adherence affect carcinogenesis and the biological features of cancer. For instance, the genetic change shown in the cell adherence system with E-cadherin could cause several steps in carcinogenesis. Genes involved in cell adherence prevent the spread of cancer cells without cancer inhibition and play an important role in suppressing metastasis (Yoshida et al., 2000; Hirohashi and Kanai, 2003). Some genes with metastasis suppressing functions, called metastasis suppressors, generally code proteins involved in cell adherence and cell movement. The first metastasis suppressor gene to be discovered is NM23 (Nucleoside diphosphate kinase A, also known as NME1). MKK4 (Mitogen-activated protein Kinase Kinase 4), another metastasis suppressor, induces the death of cells to inhibit the last part of metastasis: colony formation (Steeg, 2003). Fifth, tumor suppressor genes involved in DNA repair. Abnormality in the repair mechanism increases the mutation rate of genes, speeding up the development of cancer due to the accumulation of mutations (Loeb, 1991). A flaw in the repair mechanism contributes to the development of cancer by increasing the mutation rate within cells, potentially deactivating tumor suppressor genes and activating oncogenes by inducing a high mutation rate in other tumor suppressor genes and oncogenes, thus contributing to carcinogenesis (Markowitz, 2000).

1.3.2 Representative TSG: Rb

Next, we examine two genes that are important tumor suppressor genes. The first gene is Rb (Retinoblastoma protein, also known as pRB, RB1), the causal gene of retinoblastoma. The product of this gene shows functional abnormalities in several cancers (Murphree and Benedict, 1984). Normal Rb is phosphorylated by CDK in the process of shifting from G1 to S phase, functioning as a transcriptional regulator. From the perspective of phosphorylation, this process is carried out such that Rb, which used to exist in a hypophosphorylated state becomes hyperphosphorylated at the end of G1, maintaining that state for the rest of G1 to lose phosphate in M phase (Weinberg, 1995). In the control of cell cycle and metastasis, the interactions of Rb, the transcription factor E2F, and the HDAC (Histone Deacetylation) enzyme are very important, and the mechanism functions as follows. When there is a growth factor signal, hyperphosphorylated Rb activates metastasis through genes such as E2F and HDAC to enable cell cycle to shift from G1 to S phase. If there is no growth factor signal, Rb is hypophosphorylated, forming a complex structure with E2F and HDAC to inhibit the interaction with metastasis factors and the expression of E2F target genes, thereby suppressing the expression of genes necessary for cell cycle progression (Zhu, 2005). Also, the relationship between Rb and p53, another tumor suppressor gene, is important. The loss of function of Rb causes abnormal proliferation due to failure to control E2F and the activation of p53, which should suppress cell growth. Here, E2F directly binds with the MDM2 (Mouse Double Minute 2 homolog)-p53 complex to activate the tumor suppressing protein, $p14^{ARF'}$; this inhibits the degradation of p53. Through this mechanism, it is possible to understand the interaction of Rb and p53 in signal transduction as explained by induction of $p14^{ARF}$ and stabilization of p53 (Bates et al., 1998; Sherr and McCormick, 2002).

1.3.3 Representative TSG: TP53

Another important tumor suppressor gene is the aforementioned gene TP53 (Tumor Protein 53). The gene encodes for the phosphoprotein p53. Mutation of this gene is most frequently observed in almost cancers. The common genetic mutation is a missense mutation in cancer cells, producing defective protein (Cho et al., 1994). Rarely, null mutations do occur and research results indicate they cause a predisposition to cancer (Donhober et al., 1992). In the formation of a tumor, expression of p53 affects the phenotype of cancer, and the point at which p53 shifts somatic cells to malignant tumors depends on the tumor type. For instance, most colorectal cancers start from APC mutation to undergo genetic changes in Ras, DCC (Deleted in Colorectal Cancer), and lastly, p53 (Levine, 1997). In normal cases, the amount of p53 expressed in the cell is low and controlled by the degradation of protein by the E3 ligase (E3 ubiquitin ligase) known as MDM2, the negative regulator of p53 (Kubbutat et al., 1997). The mechanisms of p53 activation discovered in various tumors is subdivided into upstream, where p53 is activated, and downstream, where activated p53 affects the cells. First, there are three upstream pathways that can activate p53: 1) DNA damage such as ionized radiation, 2) abnormal growth signals such as expression of activated oncogenes like Ras or Myc, and 3) through other causes such as chemical carcinogens. The upstream activations are activated through stress on the cells, and inhibit the degradation of p53. However, in the downstream pathway, p53 expresses target genes such as tumor suppressor genes that inhibit cell cycle, induce cell death, maintain genetic stability and block tumor angiogenesis (Vogelstein et al., 2000).

1.4 Cancer and biomarkers

The demand for quick and accurate diagnosis is an important problem worldwide. Biomarkers could be used to address this problem, and if applied accurately for each cancer discovered, they could serve various purposes including diagnosis, prognosis, and the assessment of treatment effects. The aim of this research is to perform bioinformatics research as part of the search for useful biomarkers. Before providing a detailed description of the research, the general concepts of biomarkers are described. There are several definitions of biomarkers. The definition from the Biomarkers Definition Working Group is 'Indicators having characteristics that can be objectively measured and evaluated for pathogenic processes, biological processes, and drug response by the treatment' (Biomarker Definition Working Group, 2001). Biomarkers with these features could be applied in various fields such as detection of disease and monitoring of health status. Biomarker applications in cancer could be include quantifying the probability of carcinogenesis, early diagnosis, disease progression, and observation of treatment efficacy. Substances used as biomarkers have been expanded to more varied purposes, and can be generally classified as DNA, mRNA, protein, and metabolites. In uncovering these biomarkers, among the several methods that can be used, we use methods in which genome, proteome, and the metabolome are utilized. There are advantages and disadvantages for each method, and approach can be applied in multiple methods (Llyin et al., 2004). Figure 1.6 depicts the aforementioned methods. An examples of biomarkers used in various cancer diagnoses are found in Table 1.4 (Roche Diagnostics Global Tumor Marker Workshop, 2013). In this section, CA125 and CEA are explained in more detail. CA125 (Cancer Antigen 125) has been utilized as an important marker in diagnosis of ovarian cancer, yet in cases of women before / after menopause this biomarker shows low sensitivity and specificity. To address

this limitation, multiple biomarker analysis was performed and it was discovered that using CA125 in combination with HE4 (Human Epididymis protein 4) increased both sensitivity and specificity increases in women before / after menopause. The use of CA125 with HE4 enhanced the accuracy of diagnosis, yielding more accurate results in stage 1 ovarian cancer (Moore et al., 2008). CEA (Carcinoembryonic Antigen), another important marker, is an antigen first discovered in colorectal cancer, and likely expressed in various cancers including stomach, breast, thyroid, pancreatic, liver, gastric, renal, ovarian, and cervical cancer. Due to lack of cancer specificity, CEA cannot be used as a marker for diagnosis or screening tests, but it can be used as a useful marker for predictions before surgery and assessment of treatment effectiveness after surgery (American Cancer Society, 2012). Furthermore, biomarkers detected with new technological advances will play an important role in the diagnosis and treatment of numerous diseases besides cancer. These technologies may include the development of mathematical algorithms to analyze many factors associated with the diagnosis of disease (Bhatt et al., 2010) and biomarker panels using multiple markers.

1.5 Objective of study

Research into more accurate and specific biomarkers is important for solving the problem of cancer. This study was performed to provide useful information to suggest potential biomarkers for bioinformatics and the systems biology aspect. Two major research objectives were set, and the entire process was as follows.

First, a database that was specialized for cancer and biomarkers was constructed for efficient data collection and management. The constructed database was designed for academic utilization, and cancer-related genes (oncogene, proto-oncogene, tumor suppressor gene) and protein biomarkers were manipulated for research purposes from large public data storage.

Second, PPI (Protein–Protein Interaction) and GO (Gene Ontology) analysis were performed based on this database and other public databases. It is based on the hypothesis that biological pathways, molecular functions, and positions with interactive roles among the factors in the network of specific targets are important in cancer. Here, the target refers to potential markers for breast cancer.

In particular, the target candidates utilized in the target selection step are ones that are thought to have possibilities as biomarkers via experimentation. There candidates were utilized to select the highest-possibility targets as markers. In addition, through this analysis, the general characteristics of markers were expected to be discovered. For the research, bioinformatics analysis was performed, based on the constructed database, and it is described in detail in Chapter 2.

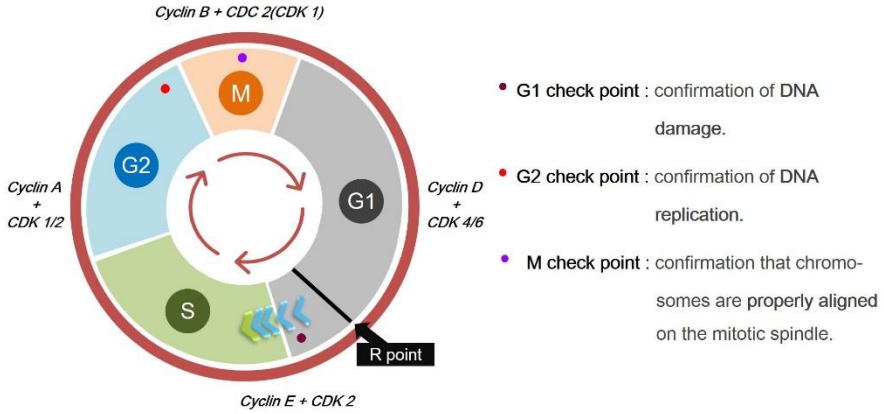

Figure 1.1 Cell cycle checkpoint and Cyclin/CDK complex. Cell cycle is composed of G1, S, G2, and M phases. The checkpoints located at the end of each phase of the cell cycle ensure transition to the next phase. Cell cycle is controlled by cyclins complexed with CDKs (Cyclin Dependent Kinases) that act at specific times (Weinberg, 2006).

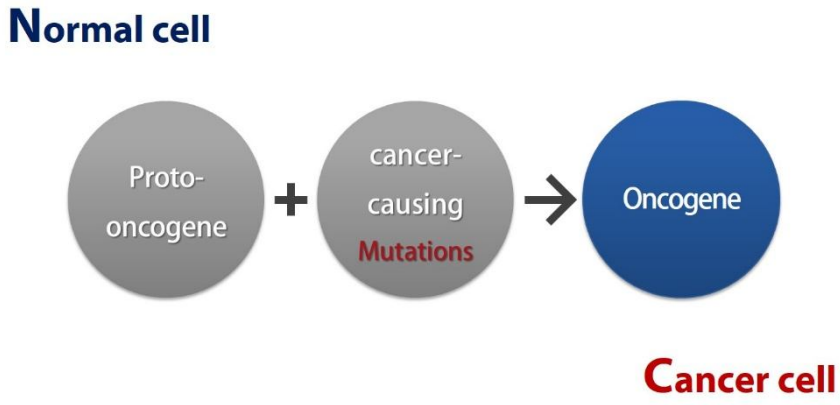

Figure 1.2 A brief schematic of cancer cell formation. Proto-oncogene is transformed into an oncogene by a mutation such as gene deletion, amplification or rearrangement.

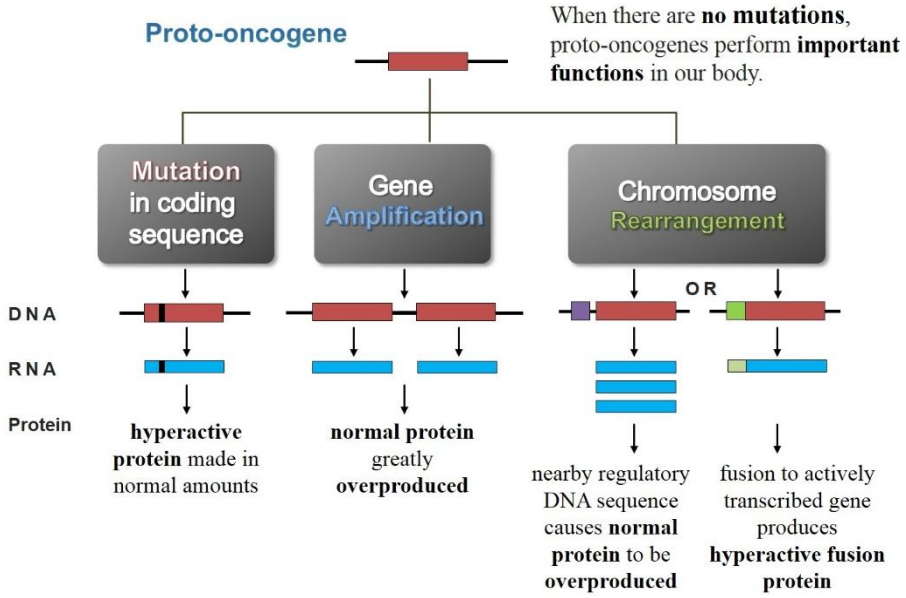

Figure 1.3 Mutations transform proto-oncogenes into oncogenes. To be an activated oncogene, a proto-oncogene needs to be mutated by the mechanisms depicted. Cancer is formed by accumulation of these mutations (Alberts et al., 2009).

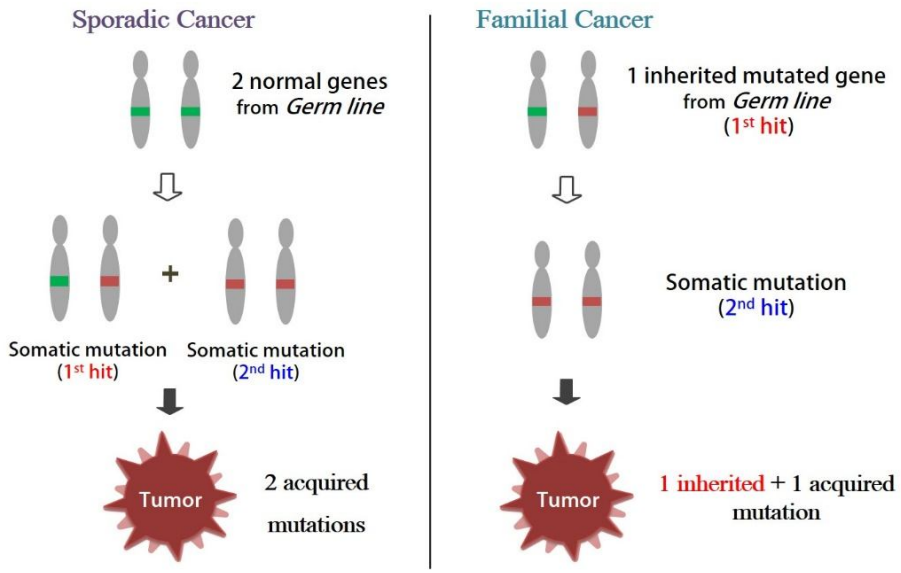

Figure 1.4 Two-hit hypothesis in tumor suppressor genes. This hypothesis explains the mechanism of inactivation of tumor suppressor genes. If a genetic mutation is inherited, it can increase the risk of cancer (Knudson, 1971).

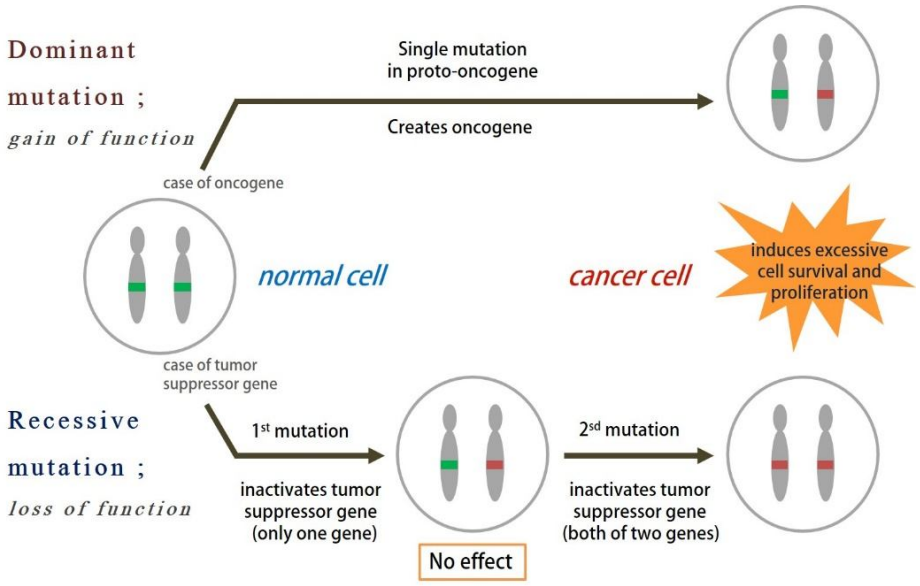

Figure 1.5 Differences in mutation mechanisms of oncogenes and tumor suppressor genes. Oncogenes and tumor suppressor genes yield the same result, such as unlimited cell division, through different paths. Oncogenes are related dominant mutations, whereas tumor suppressor genes are related recessive mutations (Alberts et al., 2009).

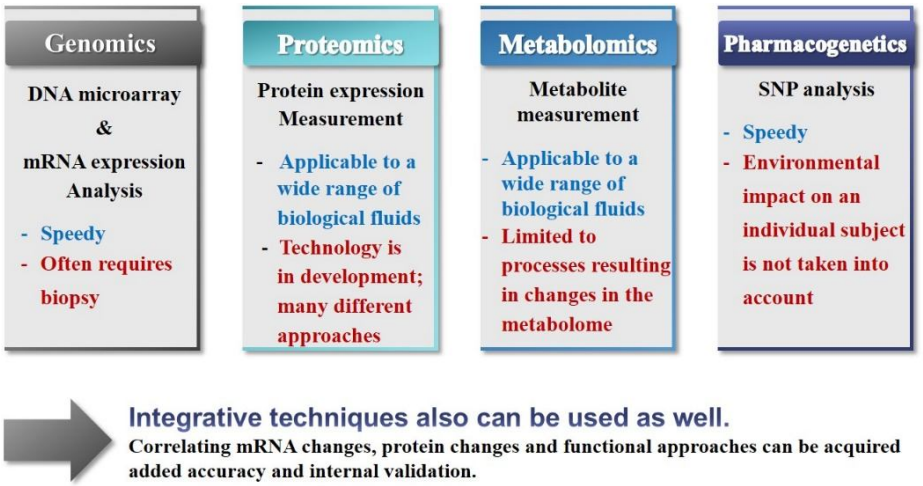

Figure 1.6 A type of biological material that can be used as a biomarker and the technical method for identifying the biomarker. A biological material that serves as a biomarker, and the associated techniques for identifying biomarkers based on advantages (blue letters) & disadvantages (red letters) (Llyin et al., 2004).

Table 1.1 Comparison of features of a normal cell and a cancer cell

Features	Normal cell	Cancer cell
Growth signaling	Growth factor signal required	Proliferative signal sustained
Growth suppressor signaling	Receives growth suppressor signaling	Evades growth suppressors signaling
cell death	If necessary, removed by apoptosis	Resists cell death
Replication	Cell division is limited	Immortal reproduction is possible
Angiogenesis	Generally maintains without creating new blood vessels	Induces angiogenesis
Metastasis	Cell location is generally fixed	Actively invades and metastasizes

Through the comparison between cancer cell and normal cell, it can be understood the characteristics of cancer cell (Hanahan and Weinberg, 2011).

Table 1.2 Classification of oncogenes according to the function of proto-oncogenes

Classification	Function	Genes		
Growth factor	Induces proliferation and have extracellular signal functions	INT2, KGF(HST), SIS		
Growth factor receptor	Enables passage of information from the cell membrane, regulates cell growth	ErbB, FMS, KIT, MET, RET, ROS, TRK		
Signal transducer	Transmits growth factor receptor signals to the nucleus	GTP-binding protein	Monomeric	H-ras, K-ras, N-ras
			Heterotrimeric	GIP, GSP
		Non-receptor protein kinase	Tyrosine kinase	ABL, LCK,SRC
			Serine/ threonine kinase	MOS, PIM1, RAF
Transcription factor	Regulates the expression of genes in the final step of signal transduction	MYC family, ErbA, ETS, FOS, JUN, MYB		
Other	Related to apoptosis	BCL2		

Oncogenes are classified as a function and biochemical properties of normal genes that is proto -oncogene (Bast et al., 2000; Sevik, 2012).

Table 1.3 Representative tumor suppressor genes in humans

Gene	Function of protein	Familial cancer syndrome	Sporadic cancer
RUNX3	TF co-factor	-	gastric carcinoma
HRPT2	chromatin protein	parathyroid tumors, jaw fibromas	parathyroid tumors
FH	fumarate hydratase	familial leiomyomatosis	-
FHIT	diadenosine triphosphate hydrolase	-	many types
RASSF1A	multiple functions	-	many types
TGFBR2	TFG-β receptor	HNPCC	colon, gastric, pancreatic carcinomas
VHL	ubiquitylation of HIF	von Hippel-Lindau syndrome	renal cell carcinoma
hCDC4	ubiquitin ligase	-	endometrial carcinoma
APC	β-catenin degradation	familial adenomatous polyposis coli	colorectal, pancreatic, and stomach carcinomas; prostate carcinoma
NKX3.1	homeobox TF	-	prostate carcinoma
$p16^{INK4A}$*	CDK inhibitor	familial melanoma	many types
$p14^{ARF}$	p53 stabilizer	-	all types
PTC	receptor for hedgehog GF	nevoid basal cell carcinoma syndrome	medulloblastomas
TSC1	inhibitor of mTOR	tuberous sclerosis	-
BMPR1	BMP receptor	juvenile polyposis	-
PTEM†	PIP_3 phosphatase	Cowden's disease, breast and gastrointestinal carcinomas	glioblastoma; prostate, breast, and thyroid carcinomas
WT1	TF	wilms tumor	wilms tumor
MEN1	histone modification, transcriptional repressor	multiple endocrine neoplasia	-
BWS/	$p57^{Kip2}$ CDK	Beckwith-	-

CDKN1C	inhibitor	wieddemann syndrome	
SDHD	mitochondrial protein	familial paraganglioma	pheochromocytoma
RB	transcriptional repression; control of E2Fs	retinoblastoma, osteosarcoma	retinoblastoma; sarcomas; bladder, breast, esophageal, and lung carcinomas
TSC2	inhibitor of mTOR	tuberous sclerosis	-
CBP	TF co-activator	Rubinstein-Taybi syndrome	AML
CYLD	deubiquitinating enzyme	cylindromatosis	-
CDH1	cell-cell adhesion	familial gastric carcinoma	invasive cancers
TP53	TF	Li-Fraumeni syndrome	many types
NF1	Ras-GAP	neurofibromatosis type 1	colon carcinoma, astrocytoma carcinoma
BECN1	autophagy	-	breast, ovarian, prostate
PRKAR1A	subunit of PKA	multiple endocrine neoplasia	multiple endocrinamas
DPC4‡	TFG-β TF	juvenile polyposis	pancreatic and colon carcinomas
LKB1/ STK11	serine/threonine kinase	peutz-jegher syndrome	hamartomatous colonic polyps
RUNX1	TF	familial platelet disorder	AML
SNF5	chromosome remodeling	rhabdoid predisposition syndrome	malignant rhabdoid tumors
NF2	cytoskeleton-membrane linkage	neurofibroma-predisposition syndrome	schwannoma, meningioma; ependymoma

Expression of proteins of tumor suppressor genes have specific functions. They are related to various familial and sporadic cancer (Weinberg, 2006).
*also known as MTS1, CDKN2, and p16
†also called MMAC or TEP1
‡also known as MADH4 and SMAD4

Table 1.4 Biomarkers used in the diagnosis of cancer

Cancer		First marker	Second marker
Colorectal		CEA	CA19-9
Pancreatic		CA19-9	CEA
Stomach		CA72-4	CEA, CA19-9
Esophagus		CEA, SCC	-
Liver		AFP	-
Gall Bladder		CA19-9	-
Breast		CEA, CA15-3	-
Ovary		AFP, CA125, HE4	CA72-4
Cervix uteri		SCC	CEA
Lung	SCLC*	NSE, ProGRP	CYFRA21-1
	NSCLC†	CYFRA21-1	CEA, SCC
Germ cell		AFP, HCG+b	-
Prostate		PSA	-
Bladder		-	CYFRA21-1
Thyroid		SCC, TG	CEA
C-Cell		Calcitonin	CEA
ENT‡		-	CEA, CYFRA21-1
mal.Melanoma		S100	-

The types of the biomarkers used for diagnosis of cancer are shown (Roche Diagnostics Global Tumor Marker Workshop, 2013).
*Small Cell Lung Cancer
†Non Small Cell Lung Cancer
‡Ear, Nose and Throat

CHAPTER II.

MATERIALS AND METHODS

2.1 Database construction

A database was structured for the collection and management of necessary data for cancer-related research. The constructed database included information regarding cancer-related genes such as oncogenes, proto-oncogenes, tumor suppressor genes, and biomarker candidates of major cancers, as well as bioinformatics analysis tools and link sites for future usage. Additionally, to make information searching more convenient, various search functions were equipped, and a FASTA (Fast-All) file download function was added. It was created as a web-based database for convenient access to and manipulation of the data. For this, an 8C AMD Opteron-6128 2.0 Ghz, 8 Gb RAM, 50 Gb SATA 7200 rpm 3 Gbps HDD server was created based on the HPC cluster system, and Linux v2.6318 was used as the operating system. The information within the database came from several public databases, and it was then manipulated for research purposes using Java. The manipulated data was saved on a MySQL (My Sequel) server, and the information on the server was synchronized with the web using HTML (HyperText Markup Language), JDBC (Java Database Connectivity), and JSP (Java Server Page), and Tomcat v7.0 utilized as a web container. For the web-based database, the system development information is found in Table 2.1.

The database is largely composed of three parts. First, in the case of cancer-related genes, two databases were used to collect gene information.

After gathering information about oncogenes and tumor suppressor genes through the NCBI (National Center for Biotechnology Information) PubMed data mining results by Tumor Associated Gene (http://www.binfo.ncku.edu.tw/TAG/GeneDoc.php), the categorized information on the UniProt (http://www.uniprot.org) database was utilized to classify the proto-oncogenes (The UniProt Consortium, 2014). In addition, the information about produced proteins was obtained from the UniProt database, and parsing that utilized Java was performed to extract information in the required form. The reason that the data had to be extracted in a specific form is that the format of information and the content provided by each public database was different, and thus the constructed database sought to unify and compile large data. Data was formed as a text file, and detailed components were GENE, CHROMOSOME, LOCUS, PROTEIN, LENGTH, and FASTA format. To provide a FASTA file with gene and protein sequence information, Java was utilized. The produced FASTA file included protein name, organism, gene name, and the protein sequence information provided by each gene, starting with '>' mark. The protein sequence is enumerated by capital letters, and breaks into a new line after each 70. The parsing process and the resulting FASTA file can be reviewed in Figure 2.1.

The previously described text file composed of GENE, CHROMOSOME, LOCUS, PROTEIN, LENGTH, and FASTA format is saved in the table produced on the MySQL open database server. The column format saved in this case is text, and the gene name, chromosome location, and the protein sequence length information were set as 'varchar' type, a variable length string. In addition, the protein name was set as 'text' type, considering that the length tends to be long, and protein sequence was set as the 'longtext' type, as they contain very long information. The gene name was saved as a primary key for future search in the table-making process, and null values were not allowed. Table 2.2 summarizes the MySQL column

information regarding each produced table. Additionally, MySQL with the data saved, JSP as a web-programming language, and JDBC as a database program interface were synchronized to enable searching according to input query and to allow searching of the gene table with only the gene name or keywords. The JSP language refers to the Java Server Page; it is able to control the content or shape of the webpage, and can call up the Java program activated by a web server such as Tomcat. JDBC (Java Database Connectivity) functions as a link between databases to run SQL queries within the program coded by Java. Such a search function is necessary as it provides the convenience of enabling the download, and opening and/or saving information regarding the produced protein of each gene and the FASTA file of selected genes directly onto the user's computer. Additionally, even if the gene's classification information is unknown, it is still searchable, and the classification information is provided in the results page.

Next, cancer protein biomarkers and differentially expressed protein are explained. As the database has been designed for biomarkers, the marker information for various types of cancers required attention. The related marker data was collected utilizing EDRN (Early Detection Research Network) (http://edrn.nci.nih.gov/) under NCI (National Cancer Institute, http://www.cancer.gov/). Cancer biomarkers were categorized according to the material: Gene, genome, protein, proteome, and epigenome. The provided biomarker included ones that were not applied clinically, yet are meaningful markers, as suggested by researchers. Therefore, there is the advantage that it acquires information that has not been provided by other public databases. The information provided by EDRN biomarkers has been classified according to the applied cancer to be shown in Table 2.3. The detailed data provided by one marker is composed of the gene name, protein name, other aliases, applied cancer, and a short description. In this research, such information was made into a table to make its utilization easier. Marker materials were

restrained to protein, and sought to provide various information including the molecular function and biological process of each marker from the GO (Gene Ontology, http://www.geneontology.org/) database (The Gene Ontology Consortium 2000), and the sequence data of the UniProt was added to provide various information. In addition to the data explained thus far, the marker panel information provided by EDRN and the protein domain data of the classified markers by cancer type from Pfam (http://pfam.xfam.org/) database were parsed in Java, and then posted in the constructed database (Finn et al., 2014). The information regarding the data parsing that utilized Java, HTML, MySQL, JavaScript, JSP, and Tomcat was carried out in the same way as the cancer-related genes previously mentioned, and were omitted. Table 2.4 shows the MySQL column information of the produced table regarding cancer biomarkers. For the diversity of target candidates, DEFs (Differentially Expressed Proteins) through assay experiments of the dbDEPC (http://lifecenter.sgst.cn/dbdepc/index.do) database are also provided, as well as EDRN biomarkers. Only validated data were included (He et al., 2012), and the data processing using computer languages was the same.

The last part is a page related to bioinformatics research, composing a link site out of BLAST (Basic Local Alignment Search Tool) and other various analysis tools (Altschul et al., 1990). In the BLAST page, the standalone BLAST and the wwwblast provided by NCBI (http://www.ncbi.nlm.nih.gov/) were used, composed to allow the utilization of sequence data such as cancer-related genes and the biomarkers of each cancer. Additionally, various bioinformatics research–related sites used for analyzing this research were linked to enhance the convenience of future usage. The detailed content of the database results are dealt with in Chapter 3.

2.2 Data collection

2.2.1 Cancer-specific network

In searching for breast cancer targets, various data and bioinformatics tools were utilized. In data collection, the constructed database for this research and several public databases were used, and the outline of the research process is shown in Figure 2.2. Protein interaction data and cancer genes data were necessary to form a cancer-specific network, and all data was gathered from public databases. First, protein interaction data was brought from the HIPPIE (Human Integrated Protein-Protein Interaction rEference) (http://cbdm.mdc-berlin.de/tools/hippie/index.php) (Release 1.5: Dec 21, 2012), which integrates public database data regarding protein interactions, and provides experimental evidence of the interactions of each protein with scored information. It includes the HPRD (Keshava Prasad et al., 2009), BioGRID (Stark et al., 2011), IntAct (Aranda et al., 2010), MINT (Ceol et al., 2009), DIP (Salwinski et al., 2004), and BIND (Bader et al., 2003) public databases, for 13,949 proteins and 122,708 interaction data in total, which were used to generate a cancer-specific network (Schaefer et al., 2012).

The cancer genes data for the cancer network were brought from four databases, and in total, 3,399 data were used. The four databases are the Cancer Gene Census (http://cancer.sanger.ac.uk/) (Futreal et al., 2004), Bushman Lab (Cancer Gene List, http://www.bushmanlab.org/links/genelists, accessed July 13, 2014), NCG 4.0 (http://ncg.kcl.ac.uk/) (An et al., 2014), and the Tumor Associated Gene database, as explained in the database construction process. Additionally, for verification of future research results, five clinical markers: CEACAM5, ERBB2, ESR1, MUC1, and PGR were included in the data list with reference to the NCI (Tumor Markers, http://www.cancer.gov/cancertopics/factsheet/detection/tumor-markers, acces

sed July 25, 2014), FDA (Food and Drug Administration, Drugs, http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm, accessed July 25, 2014), and ROCHE (Roche Diagnostics Global Tumor Marker Workshop, 2013). Whether the clinical markers had an important position in the network and gene ontology was confirmed. So far, data collection was explained for creating a cancer-specific network. The Gene Symbol of the HUGO (Human Genome Organisation) transformed by WebGestalt (Wang et al., 2013) and the HIPPIE protein interaction data corresponding to the symbol were used as final input data.

2.2.2 Target candidates list

The possible target candidates list for breast cancer was built from two databases, and 210 data were used in the analysis. This includes the protein biomarker list of breast cancer from EDRN and experimentally validated data in a specific expressed protein list from dbDEPC. These data were utilized to select the most potent biomarker, and the IDs of the collected protein data were transformed into a Gene Symbol in previous methods for analysis in the next step.

2.3 Cancer-specific network analysis

To analyze the features of the produced network and define hub genes in detail, the degree, betweenness, and stress thought to be biologically meaningful or in cancer via previous research were decided by network analysis parameters (Jeong et al., 2001; Joy et al., 2005; Jonsson and Bates, 2006; Yu et al., 2007; Vashisht and Bagler, 2012). The results of the respective selected parameters were ranked, and higher hub genes were filtered. NetworkAnalyze of Cytoscape was used in this calculation (Assenov et al., 2008).

Degree distribution Degree is one of the fundamental factors in determining the features of specific nodes, and is expressed as connectivity. From this value, it is possible to learn how many direct connections a specific node has with other nodes, and obtain the likelihood that a randomly selected node has the exact degree of *k* by calculating *P(k)*, the degree distribution.

$$P(k) = n_k / N$$

As for a network with a total of *N* nodes, *P(k)* is obtained by dividing the number of nodes, *n(k)*, by the degree of *k* with the total number of nodes, *N* (Barabasi and Oltvai, 2004).

Betweenness centrality This reflects the extent of influence that a certain node has on other nodes (Yoon et al., 2006), and it is as follows.

$$C_B(n) = \sum_{s \neq n \neq t} \sigma_{st}(n) / \sigma_{st}$$

Here, *s* and *t* are values from other nodes, not the value of the specific node *n*, and σ_{st} refers to the number of shortest paths from *s* to *t*, and σ_{st} refers to the number of specific nodes *n* that are in the shortest paths (Brandes, 2001).

Stress centrality This refers to the number of shortest paths passing through specific node *n*, and if the number of shortest paths is high, that node has a high stress value. The equation is as follows:

$$C_S(n) = \sum_{s \neq n \neq t} \sigma_{st}(n)$$

The distribution of stress centrality provides information regarding the number of nodes with stress *s* with various values (Shimbel, 1953; Brandes, 2001).

Additionally, it was reviewed whether the produced network is scale-free nature, by seeing if the degree distribution, one of the parameters, follows the power-law distribution. The power-law is known to be appropriate for applying to network science, and following this distribution enables us to explain the existence of the hub. The equation is as follows:

$$P(k) \sim k^{-\gamma}$$

If *P(k)* which is the likelihood that the selected node has the exact degree of *k* follows power-law, it is called the scale-free network, and in most cases in these networks, the degree exponent γ has value 2–3 (Barabasi and Albert, 1999). To check this, the Origin (OriginLab, Northampton, MA) program was utilized.

2.4 Gene ontology analysis

If the hub genes in the cancer network were verified through network analysis, this time the significant GO terms regarding Biological Process (BP), Molecular Function (MF), and Cellular Component (CC) were determined through GO enrichment analysis; Gorilla (Eden et al., 2009) was utilized in the analysis. GO term selection was carried out in two steps: First, the 'two unranked lists of genes' mode was selected to obtain the GO terms specific to breast cancer. In here, the background set is the total genes in the cancer network, and the target set is the breast cancer–related genes. The breast cancer–related genes used in this process included the defect gene list that increases the risk of breast cancer from Cancer Research UK (Breast cancer genes, http://www.cancerresearchuk.org/cancer-help/type/breast- cancer/ about/risks/breast-cancer-genes, accessed July 25, 2014), the Top 20 breast cancer mutant gene provided by COSMIC (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/) (Forbes et al., 2011), the breast cancer–related gene list of the Genetics Home Reference (Breast cancer, http://ghr.nlm.nih.gov/condition/breast-cancer, accessed July 25, 2014), and in total, 38 genes were used. The second step was to process a 'single ranked list of genes' mode. It covered all genes included in the cancer network, and the genes with selected GO terms from breast cancer–related genes in the previous step were filtered to obtain the list.

The threshold value of the p value was set at < 0.001, and to treat the GO terms, the threshold value of B (the number of genes associated with significant GO terms) was set to one standard. Based on the prior research results in which the B value explains the signal (specific) and noise (non-specific) (Vashisht and Bagler, 2012), in this research, the threshold was set as ≤ 100, which is the most applicable section from BP and CC. The gene list of GO terms specified for breast cancer from genes forming the cancer

network was obtained from such gene ontology analysis, and these were considered as the specific factors of gene ontology, regardless of their location and roles within the network as a hub.

2.5 Target selection for breast cancer

Utilizing the results from the cancer network analysis and gene ontology analysis, and selecting the target that is most likely for breast cancer, Venny's Venn diagram (Oliveros, 2007) was utilized. First, the Venn diagram of the three groups was produced to see which parts differed among what the three groups shared in a group. The first group was the hub genes from the cancer-specific networks, the second group included genes with significant GO terms in breast cancer, and the third group was composed of would-be target candidates of breast cancer. The area formed by the hub genes, the GO genes, and the candidates group was decided as the area for targets that were specialized for breast cancer. Additionally, the area in which the previously mentioned breast cancer clinical markers CEACAM5, ERBB2, ESR1, MUC1, and PGR appear was checked for reference purposes. Therefore, new candidates that have potential as breast cancer markers with specificity of PPI and the GO terms can be selected. This process is explained briefly in Figure 2.3. So far, the research process has been explained step-by-step. A database specialized in cancer and biomarkers was structured for efficient data management and analysis. It is the point of this research to screen high potential targets by applying the cancer-specific network and the gene ontology analysis results to the target candidates based on the constructed database and other public databases. The detailed contents of the research results are explained in Chapter 3.

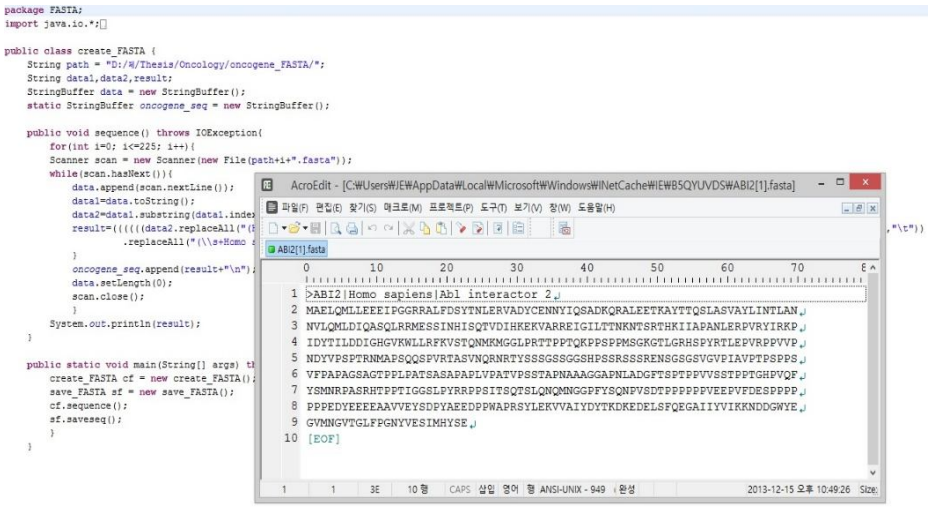

Figure 2.1 Java parsing process for creating FASTA files. Java s/w as programming language was utilized for data mining. File contents consist of '>' mark, protein name, organism, gene name and protein sequence information.

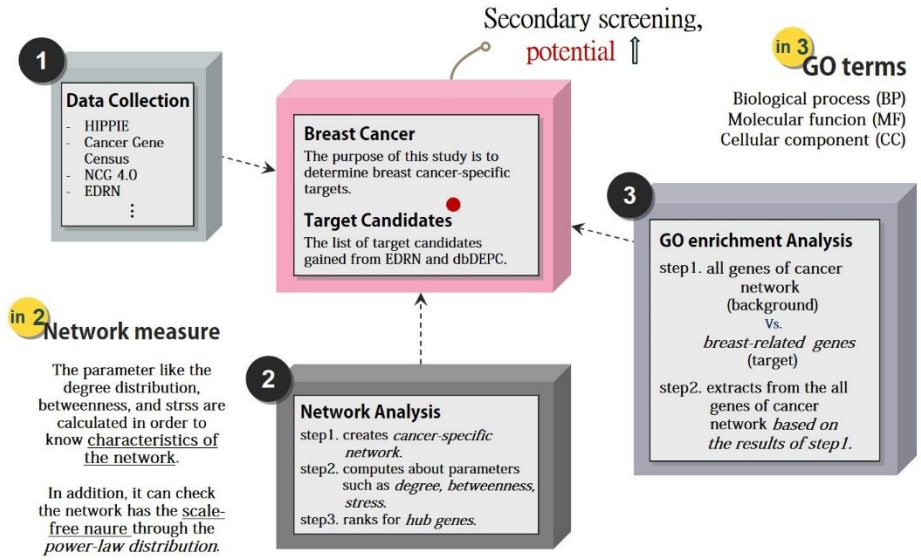

Figure 2.2 Workflow for the investigation of breast cancer-specific targets. The objective of this study is the specified breast cancer targets search. For this purpose, it tried to approach the problem in terms of bioinformatics and systems biology. It progressed by dividing into three detailed fields: Data collection, Network analysis, GO enrichment analysis.

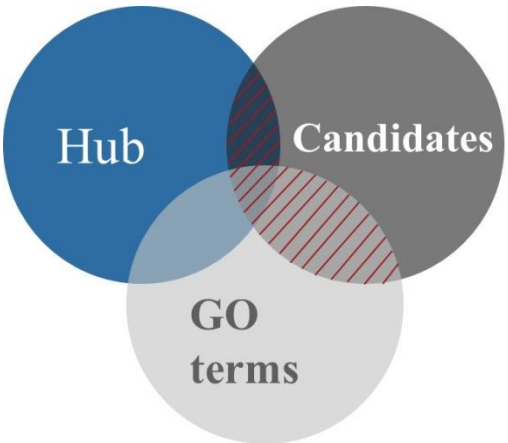

Figure 2.3 Target selection that utilizes the Venn diagram. To determine target area selection using a Venn diagram that was formed into three groups. Each group is composed of hub genes by the network, genes by GO terms, and candidates. The final targets are selected by area of the shaded.

Table 2.1 System development environment for webserver

Category	Description
System	HPC cluster system
CPU	Master node (24 core), 10 compute node (1 for 8 core)
Memory	Master node (16GB), 10 compute node (1 for 8 GB)
Operating system	Linux
Web server container	Apache Tomcat v7.0
DBMS	MySQL v5.5.28
Programming language	HTML, JAVA, JavaScript, JSP

Linux, Apache Tomcat, MySQL and other programing languages were used in order to build the web-based database.

Table 2.2 Schema of the cancer-related gene tables in MySQL

Database & Field	Data type	Null	Primary key
oncogene			
gene	varchar(20)	NO	PRI
chromosome	varchar(10)	YES	
location	varchar(20)	YES	
protein	text	YES	
protein_length	varchar(10)	YES	
proto_oncogene			
gene	varchar(20)	NO	PRI
chromosome	varchar(10)	YES	
location	varchar(20)	YES	
protein	text	YES	
protein_length	varchar(10)	YES	
tumor_suppressor_gene			
gene	varchar(20)	NO	PRI
chromosome	varchar(10)	YES	
location	varchar(20)	YES	
protein	text	YES	
protein_length	varchar(10)	YES	
oncogene_fasta			
protein	text	YES	
organism	varchar(20)	YES	
gene	varchar(20)	NO	PRI
sequence	longtext	YES	
tumor_fasta			
protein	text	YES	
organism	varchar(20)	YES	
gene	varchar(20)	NO	PRI
sequence	longtext	YES	

The tables show oncogene, proto_oncogene, tumor_suppressor_gene, oncogene_fasta, tumor_fasta. Each column of each table is constituted as follows.

Table 2.3 Description about cancer biomarkers from EDRN

Cancer	Type	Panel	No.
Bladder	Gene / Genomic		1 / 0
	Protein / Proteomic		1 / 0
Breast	Gene / Genomic		2 / 1
	Protein / Proteomic		113 / 0
Colon	Gene / Genomic		0 / 0
	Protein / Proteomic		10 / 2
	Epigenetic		1
Esophagus	Gene / Genomic	O (3)	0 / 8
	Protein / Proteomic		0 / 0
Head and Neck	Gene / Genomic		6 / 0
	Protein / Proteomic		2 / 0
Liver	Gene / Genomic		0 / 0
	Protein / Proteomic		9 / 0
Lung	Gene / Genomic	O (12)	34 / 20
	Protein / Proteomic		103 / 2
Ovary	Gene / Genomic	O (1)	1 / 1
	Protein / Proteomic		202 / 0
Pancreas	Gene / Genomic		0 / 0
	Protein / Proteomic		7 / 0
Prostate	Gene / Genomic	O (1)	347 / 8
	Protein / Proteomic		27 / 0

EDRN supplies biomarkers data about bladder, breast, colon, esophagus, head and neck, liver, lung, ovary, pancreas, and prostate cancer (Biomarkers, accessed May 31 2014).

Table 2.4 Schema of the biomarker tables in MySQL

Database & Field	Data type	Null	Primary key
Applies equally to all the tables.			
table 1			
cancer	varchar(20)	YES	
protein	text	YES	
organism	varchar(20)	YES	
gene	varchar(20)	NO	PRI
sequence	longtext	YES	
table 2			
gene	varchar(20)	NO	PRI
molecular_function	longtext	YES	
biological_process	longtext	YES	

The tables show bladder, breast, colon, esophagus, head and neck, liver, lung, ovary, pancreas, prostate cancer. each column of each table is constituted as follows.

CHAPTER III.

RESULTS

3.1 Cancer biomarker database: BWBD

This research performed bioinformatics analysis for investigation of breast cancer targets. A database to efficiently collect and manage the data was constructed. The constructed database was named BWBD (Bioinformatics With Biomarker Database), and the main page is shown in Figure 3.1. The details of the database are discussed below. The completed database can be found at http://lcbb.snu.ac.kr/BWBD/Main.html.

3.1.1 Cancer-related genes

The cancer-related genes page was composed with Allgene page, which show all the genes; in addition, the details pages in which oncogenes, proto-oncogenes, tumor suppressor genes and their specialized information were included. The search box for each page was separated for all the genes and the detailed gene categories. The user could use the search box to insert the first letter of the gene provided in the list of the genes from the database. Additionally, it is possible to search for each chromosome, too. Especially, if the user does not know that a gene belongs to a certain category, user can be provided with the category information of the gene from Allgene page. Each page provides information of cancer-related genes in table form, including oncogenes, proto-oncogenes, and tumor suppressor genes. Included are their short description, gene name, locus within chromosome, produced protein,

protein length, protein sequence, and FASTA file download field. The FASTA file produced as text contains the protein name, protein sequence, organism information, and is downloadable to use in future research. In addition, the domain information of the protein produced by each gene is provided for the expansion of information usage. The search box and the FASTA download are shown in Figure 3.2, and the respective detailed pages are seen in Figures 3.3, 3.4, 3.5.

3.1.2 Cancer biomarkers

The cancer biomarkers page provides information by categorizing the protein biomarker data from EDRN of NCI; cancers included are major cancers such as head and neck, breast, colon, liver, lung, ovarian, pancreatic, and prostate cancer. For each marker, the protein name, gene name, molecular function, biological process, and the sequence information of the marker is provided. The marker's molecular function and biological process were derived from the data of the GO database, and the sequence information was brought from the UniProt, protein public database. This page also enables search by gene name via a search form. Additionally, it shows the domain table of protein biomarkers to provide detailed family information, the sequence region of the domain, and E-value. If it is a cancer with marker panel information, it is posted to show a description of the combinations of markers and the components. The page with biomarkers is shown in Figure 3.6. In the case of DEPs, the information was brought from the dbDEPC database. This deals with information regarding UniProt ID, organism, description, up / down, and ratio, only if there is the biomarker data. This is shown in Figure 3.7.

3.1.3 BLAST (Basic Local Alignment Search Tool)

The BLAST page was created by connecting with the standalone BLAST and wwblast provided by NCBI, allowing homology analysis by using the local database specialized for the purpose of the user. Currently, oncogenes, proto-oncogenes, tumor suppressor genes, the marker protein of each cancer, human reference proteins, and primates proteins are listed. A proper BLAST program is selected for use in the search. When BLAST is used, if the query sequence is inserted in the search window, the user can be provided the search results from the local database according to homology. Though the BLAST tool was not used in this research, it is widely used for bioinformatics analysis. Therefore, it was included in the database design stage. The page could be seen in Figure 3.8. In addition, the link site page was produced, in which various bioinformatics research related sites such as GO, Cytoscape, and KEGG were linked to be used in other research. The next chapter will describe research results of utilizing various data and bioinformatics tools based on the constructed database.

3.2 Result of cancer-specific network analysis

3.2.1 Nature of cancer-specific network

A cancer network was created by mapping the cancer gene ID of 3,399 cancer genes with protein interaction data including 13,949 proteins, 122,755 interactions. The network is composed of 2,902 proteins and 17,383 interactions, and the biggest formed cluster which accounts for 87.9% of the cancer genes mapped in the network includes 2,551 proteins (node) and 17,381 interactions (edge). The major cluster is shown in Figure 3.10.

To see if the produced cancer-specific network has a hub, degree distribution, whether the basic parameter of the network follows power-law distribution, was observed. The conducted log-log graph for P(k), the likelihood that the selected node has the exact degree of k, showed a scale-free nature. 'Scale-free nature' means that there is no specific standard in describing the factors forming the network (Barabasi and Albert, 1999). Additionally, the degree exponent γ value was 1.36, not the general 2–3, the value required to deduce that the expression pattern of hub genes is relatively modest and numerous (Onnela et al., 2007). The results graph is shown in Figure 3.11. When there are 10 or less degrees, there are 2,067 nodes, 71.2% of the total; 100 or more, 42, 1.4%; 200 or more, 10, 0.3%; and 300 or more, 3, and 0.1% of the total. This is evidence that a small number of nodes have high degree values. From this, the existence of a hub as a feature of a scale-free network has been confirmed.

3.2.2 Top ranked hub genes

The previous section concerned the features of the cancer network. Based on the network features and network parameters of degree, betweenness, and stress, the rank of the hub gene was decided for the next analysis. As seen in Figure 3.11, the rank of the hub gene be considered according to the k value at the steep change of the P(k) through the vertical drop line on the graph. This is a point at which the log value of k is 1.65 or more. The nodes likely to be hub genes appear from this point. Therefore, to be used in the next ontology analysis, the top 150 genes of this point were filtered from three parameters, and a total of 166 hub genes were determined. The position of the top 150 hub genes of the cancer-specific network can be seen in Figure 3.12. Hub genes are the nodes marked yellow, and they are all included in the major cluster. TP53, RB1, and APC were included as tumor suppressor genes, and for oncogenes, genes such as EGFR, Src, and MYC were included. It was ascertained that the genes that are known to have important functions in cancer play their roles as hub genes in the cancer network created in this research. The list of 166 hub genes is given in Table 3.1, and they are divided by degree distribution.

3.3 Result of gene ontology analysis

3.3.1 GO terms of breast cancer–related genes

Enrichment analysis was performed to filter specific GO terms regarding breast cancer by gene ontology analysis. This is the process of obtaining GO terms regarding cancer and is more significant for breast cancer. All 2,902 genes of the cancer network were used as the background set, and 38 breast cancer–related genes were used as the target set. They were filtered with p-value <0.001 and B≤100 as the standard.

The top 10 of the 100 GO terms of the biological process in enrichment scores are protein K6-linked ubiquitination (GO:0085020), negative regulation of fatty acid metabolic process (GO:0045922), Schwann cell development (GO:0014044), response to indole-3-methanol (GO:0071680), cellular response to indole-3-methanol (GO:0071681), glial cell apoptotic process (GO:0034349), positive regulation of metaphase/anaphase transition of cell cycle (GO:1902101), positive regulation of mitotic metaphase/anaphase transition (GO:0045842), negative regulation of mammary gland epithelial cell proliferation (GO:0033600), and negative regulation of neuroblast proliferation (GO:0007406). As for molecular function, when filtered by p-value and B value, there are two GO terms of the standard: JUN kinase kinase activity (GO:0008545) and beta-catenin binding (GO:0008013). In the cellular process, nine GO terms were obtained: BRCA1-BARD1 complex (GO:0031436), Mre11 complex (GO:0030870), nuclear chromosome, telomeric region (GO:0000784), BRCA1-A complex (GO:0070531), cell-cell adherens junction (GO:0005913), chromosome, telomeric regions (GO:0000781), chromosomal region (GO:0098687), PML body (GO:0016605), and nuclear body (GO:0016604). Table 3.2 summarizes about GO terms by category.

3.3.2 Cancer network genes and breast cancer-specific GO terms

GO enrichment analysis was performed, in which the breast cancer-specific GO terms of each category were applied to all genes of the cancer network. This is the process to gain the gene list involved in cancer, especially more significant GO terms in breast cancer. In this instance, it was performed in a single ranked list of modes, filtered with p-value <0.001 and B≤100 as standard. Genes with specific GO terms of the biological process total 477 of 2,902 genes of the cancer network. This list includes ABL1, APC, BAD, EGFR, EP300, ERBB2/3/4, ESR1/2, FAS, FGFR1/2/3, FOS, JUN, KRAS, MUC1, MYC, NF1, and TP53. Molecular function showed 11 results: APC, AR, BTRC, CDH1, EP300, ESR1, GSK3B, HDAC6, PXN, SMAD3, and VCL. In the cellular component, 55 genes were identified in the results, including ATM, CDK2, CHEK1/2, MAX, PML, PTEN, RB1, SKIL, TP53, and XPO1. As a result, 499 genes were filtered as cancer-related genes with breast cancer-specific GO terms. For breast cancer target selection with 210 possible target candidates, 166 hub genes and 499 breast cancer-specific GO genes were used.

3.4 Result of target selection for breast cancer

To select the breast cancer target, the final research goal, the data obtained so far were expressed in a Venn diagram. Three groups were created: the hub genes group, specific GO genes group, and possible target candidates group. The result of the Venn diagram can be seen in Figure 3.12, and the area for target selection is shown as the hatched region. This field was set to filter targets that show breast cancer specificity in the PPI network and gene ontology by reflecting the hypothesis of this research. A total of 34 genes were selected, and the results are as follow.

Genes that function as a hub of the network were FLNB and HNRNPK, and genes with specific GO terms were ANXA1, CALR, CCND1, CDH1, CRYAB, CTBP1, EIF3E, ERBB3, FAS, G3BP1, HSPB1, JUNB, KIT, MAP2K1, MUC1, PFN1, PGR, PRDX1, PTPRC, RHOA, and SOX2. Genes that show their features both in the network and gene ontology were BRCA1, EP300, ERBB2, ESR1, FLNA, GRB2, HSP90AB1, NPM1, PCNA, PTPN11, and YWHAZ. Additionally, viewing the area with five clinical markers for result verification, MUC1 and PGR were found in the specific GO genes field, and ERBB2 and ESR1 were discovered in the common field of network with GO. The results for the sub-network of the selected targets and the genetic ontological features of the 34 targets utilizing Panther (http://www.pantherdb.org/) (Mi et al., 2005) are shown in Table 3.3 and Figure 3.12, and Table 3.4, respectively. Additionally, as a result of viewing how the targets are shown in the cancer map of KEGG (http://www.genome.jp/kegg/) (Kanehisa and Goto, 2000), 12 genes were shown in 15 cancer pathways, and shown in Table 3.5. Detailed descriptions are given in Chapter 4.

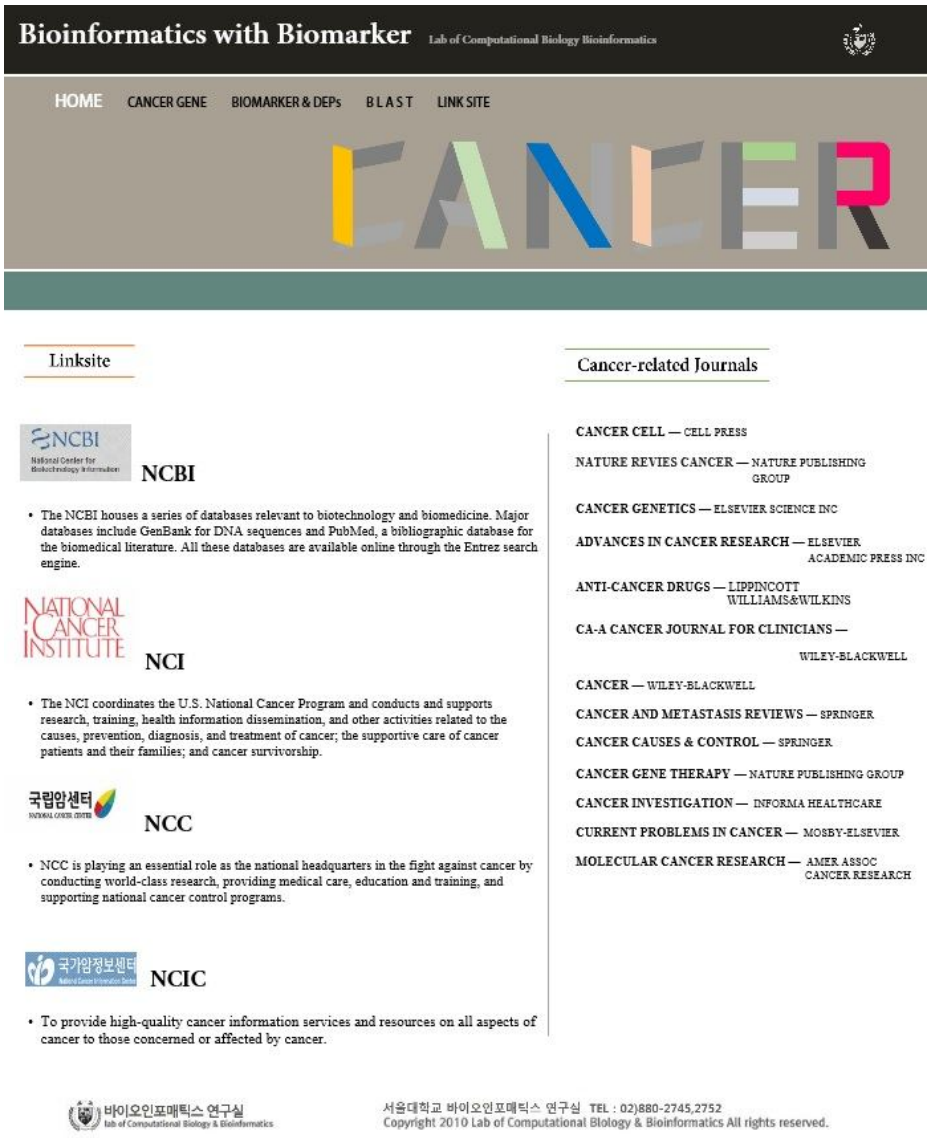

Figure 3.1 Main page of the database. This page consists of cancer-related link sites such as NCBI, NCI, NCC, NCIC, and cancer journals as CANCER CELL by Cell press, NATURE REVIEWS CANCER by Nature, etc. Submenu titles are CANCER GENE, BIOMARKER&DEPs, BLAST, and LINK SITE. All pages and links are activated.

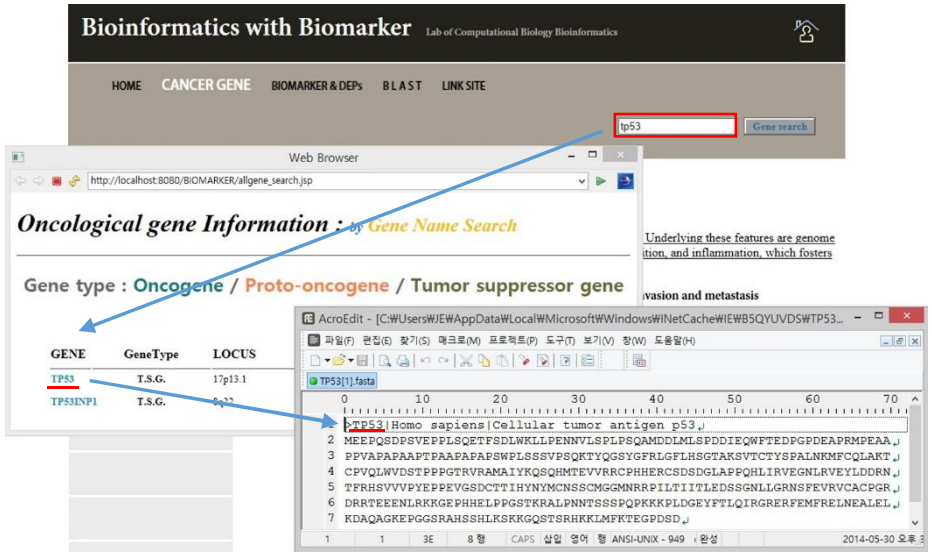

Figure 3.2 Search box and downloaded FASTA files. The search box of the Allgene page for cancer-related genes and each detail search box for oncogenes, proto-oncogenes, and tumor suppressor genes. In the search results table, users can download the selected files.

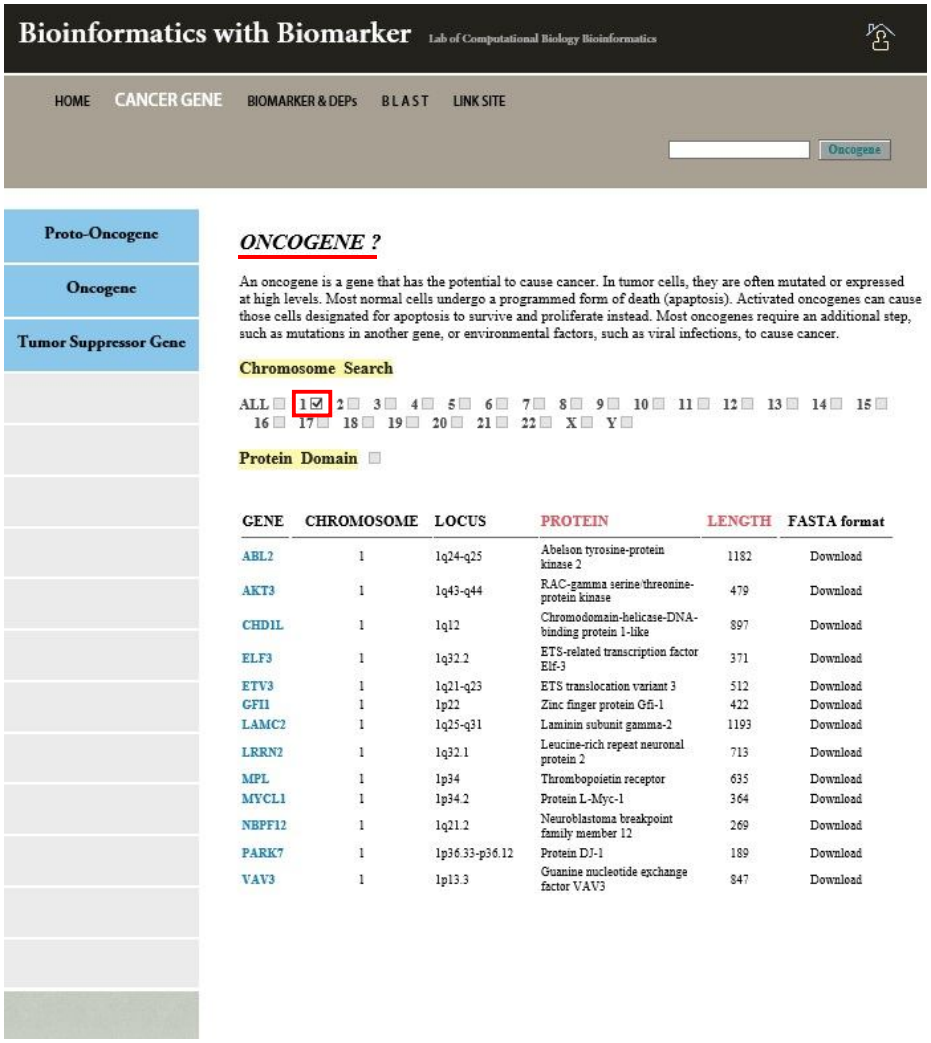

Figure 3.3 Oncogene page of database. Search box only for the oncogene, this page includes information on 123 oncogenes. The table is composed of gene name, chromosome, locus, produced protein, protein length, and field of download for protein sequence in FASTA file format.

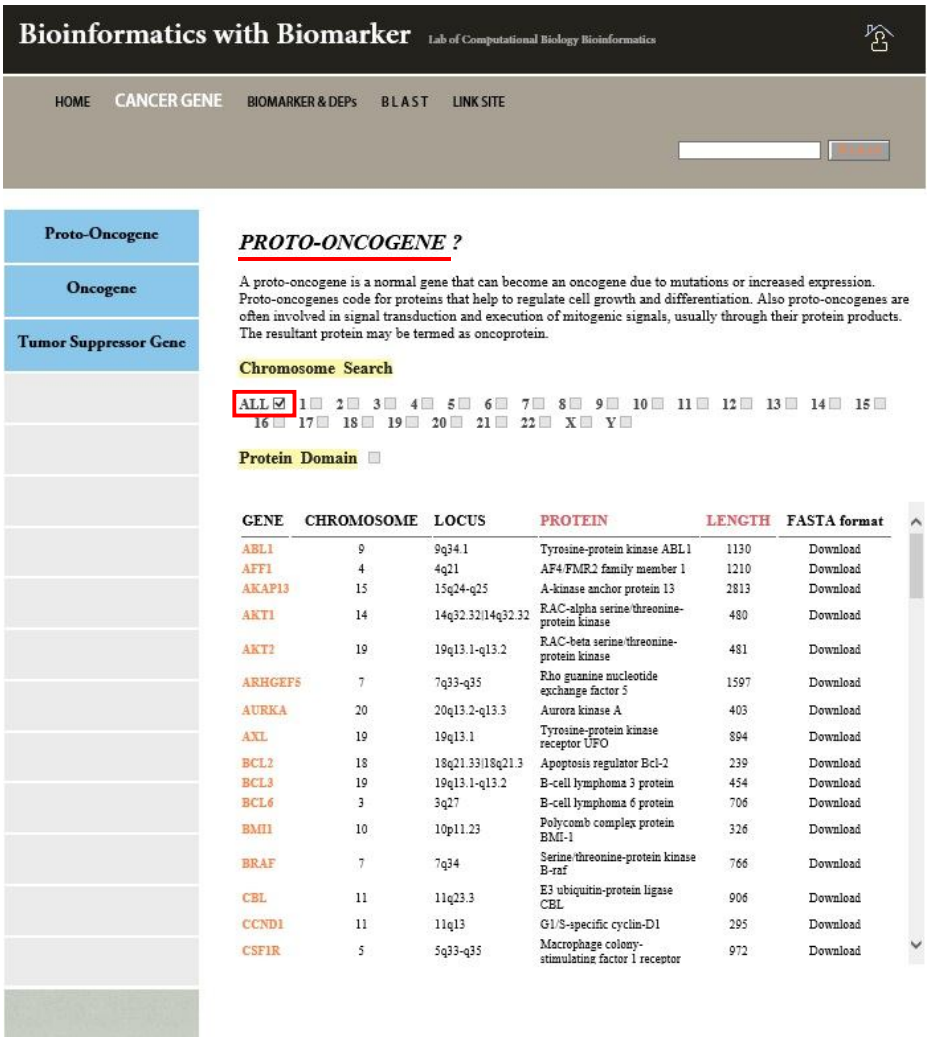

Figure 3.4 Proto-oncogene page of database. Search box only for the proto-oncogene, this page includes information on 105 proto-oncogenes. The table is composed of gene name, chromosome, locus, produced protein, protein length, and field of download for protein sequence in FASTA file format.

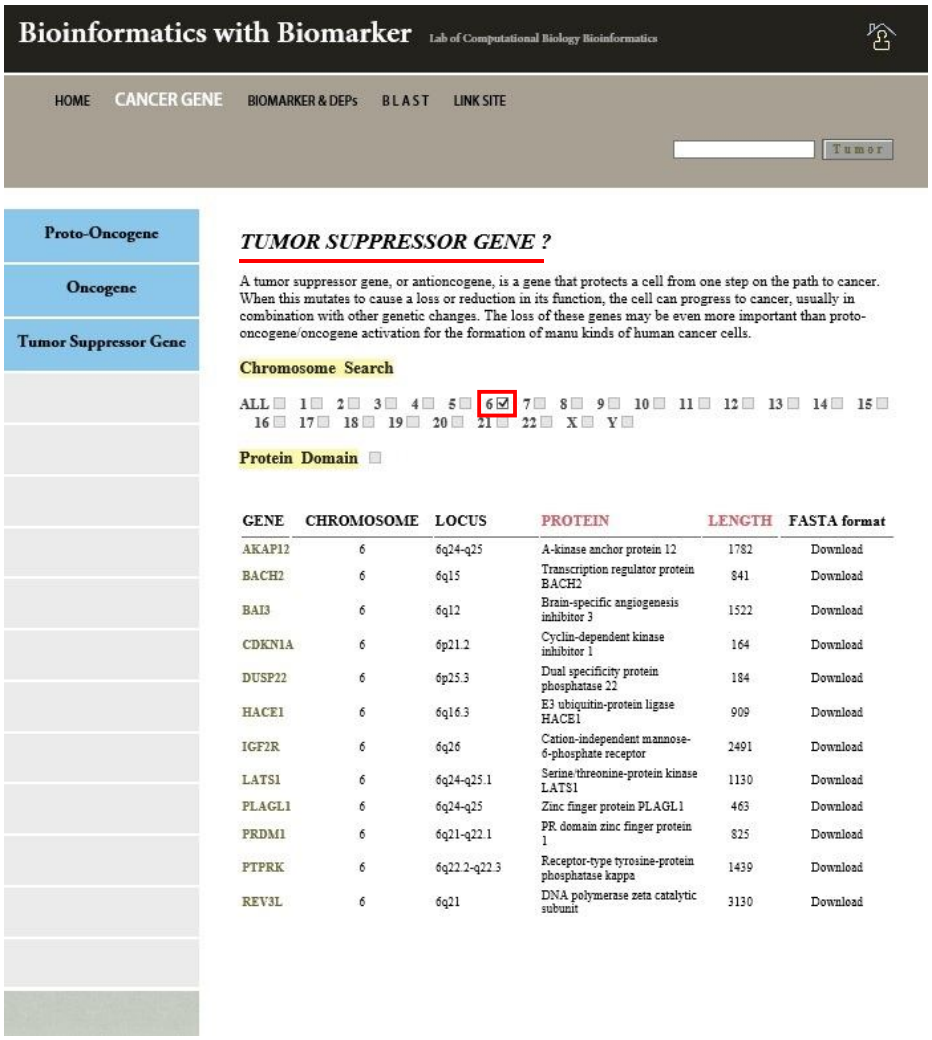

Figure 3.5 Tumor suppressor gene page of database. Search box only for the tumor suppressor gene, this page includes information on 244 tumor suppressor genes. The table is composed of gene name, chromosome, locus, produced protein, protein length, and field of download for protein sequence in FASTA file format.

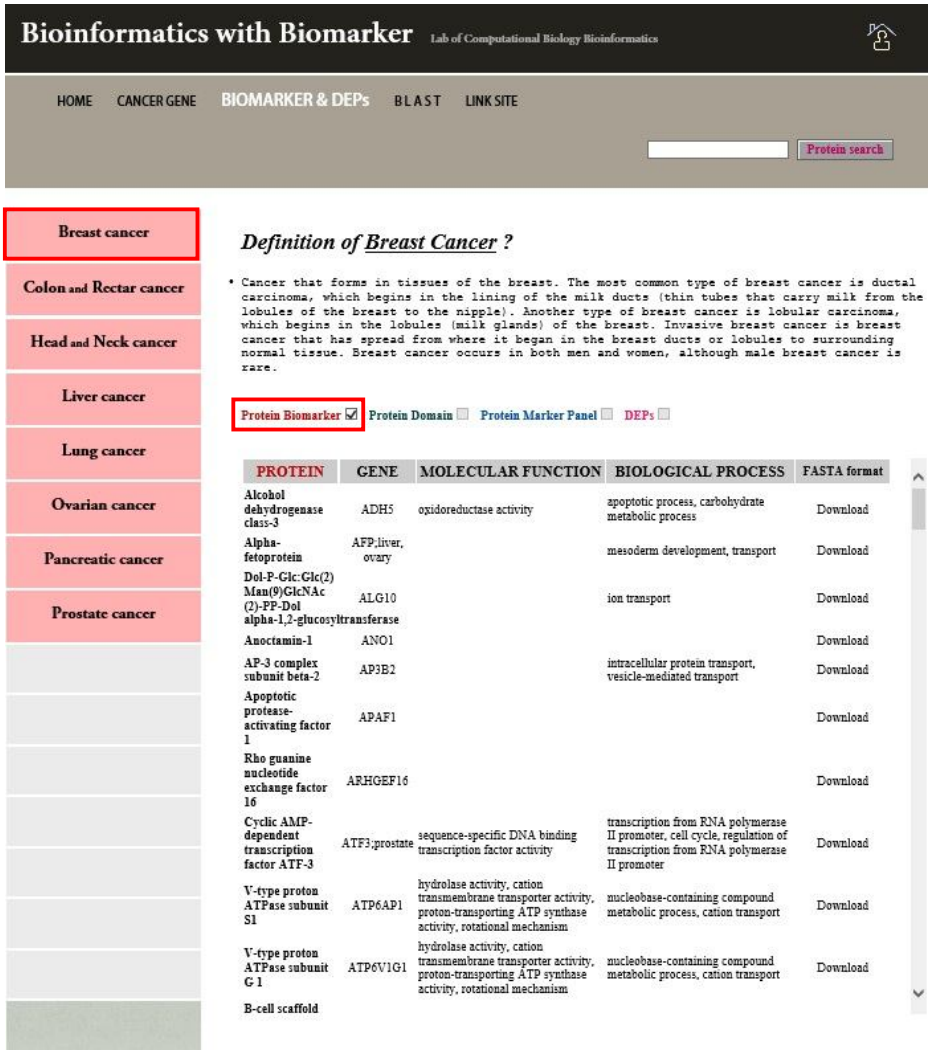

Figure 3.6 Biomarker page of database. This page provides information about protein biomarkers for breast cancer. Data for each biomarker provides protein name, gene name, molecular function, biological process, and FASTA format download.

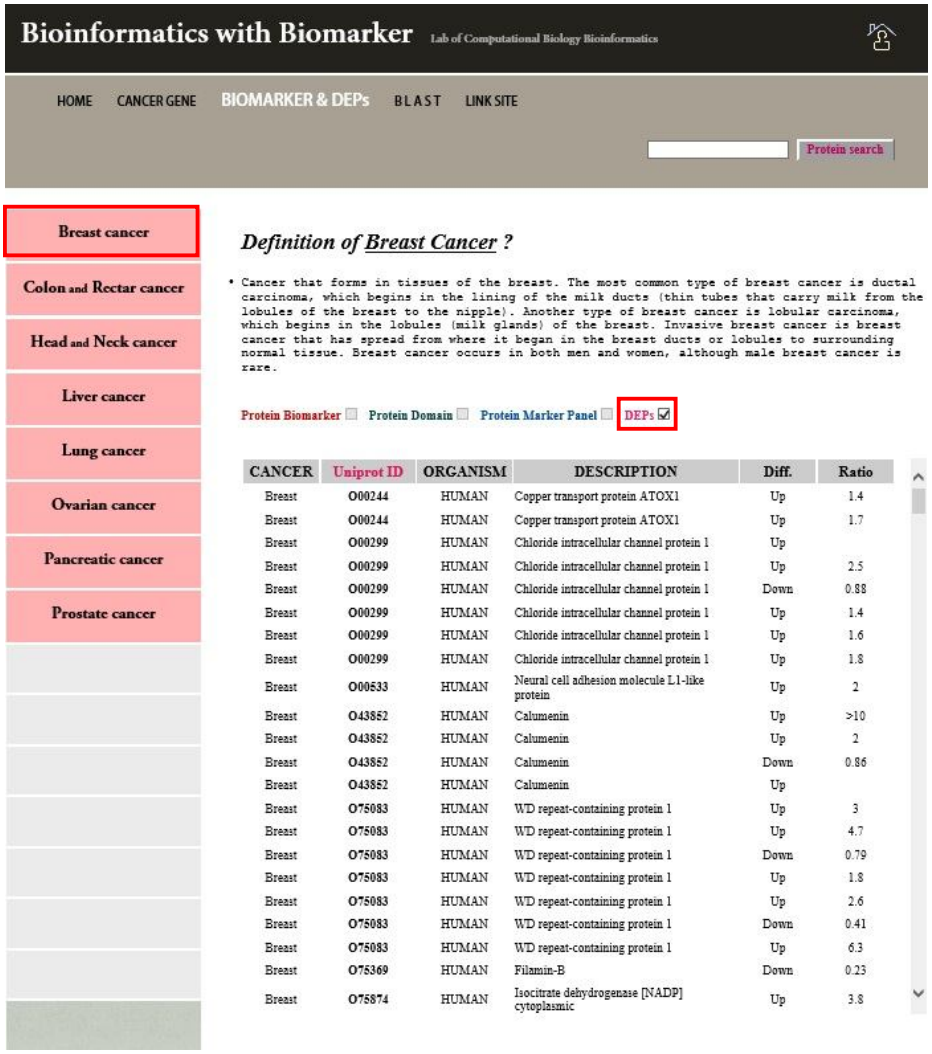

Figure 3.7 DEPs page of database. This page provides DEPs (Differentially Expressed Proteins) information for breast cancer. DEPs data was obtained only from assay results that are validated. The table is composed of cancer, UniProt ID, organism, description, diff., and ratio.

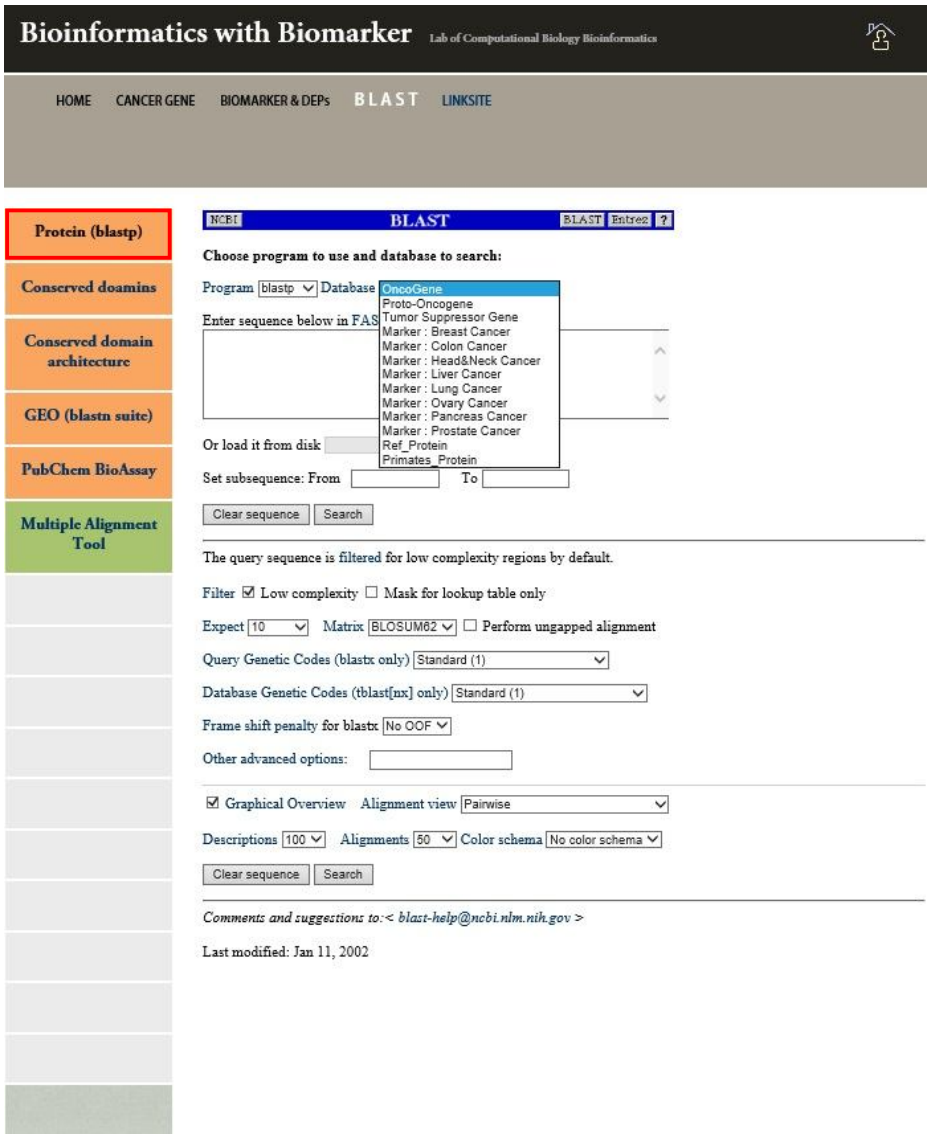

Figure 3.8 BLAST page of database. This page provides the BLAST tool for use with local databases. Lists consist of oncogene, proto-oncogene, tumor suppressor gene, each specific cancer protein marker, human reference protein, and primate protein.

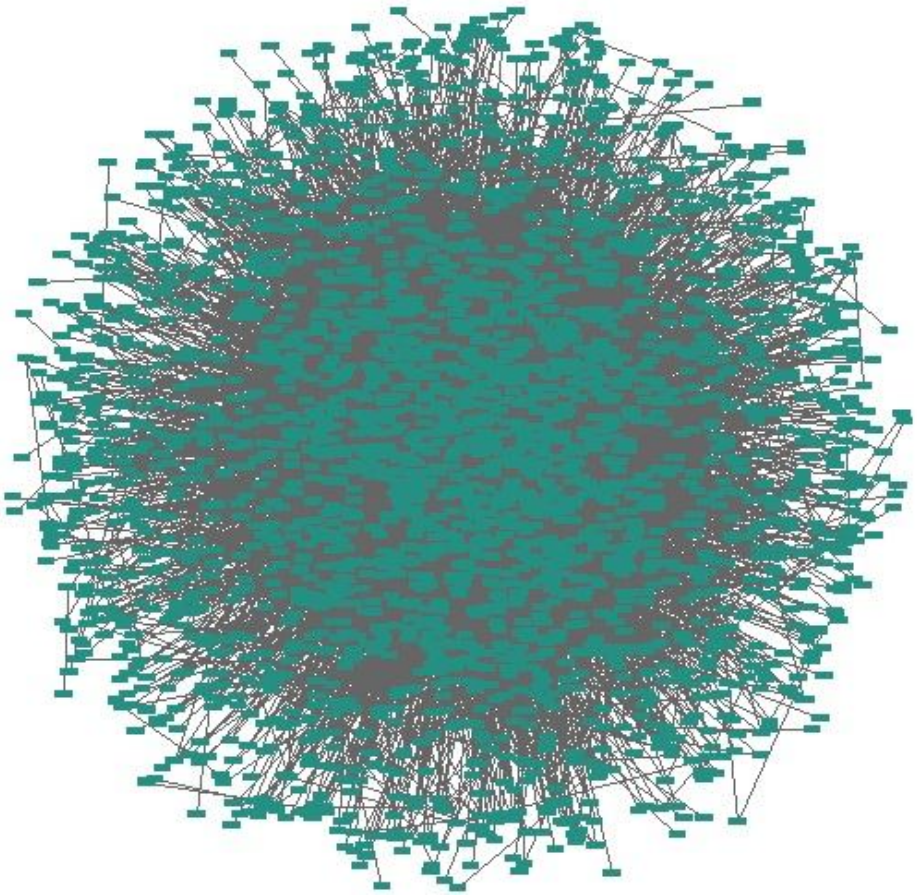

Figure 3.9 The major cluster that is generated as the cancer-specific network. This cluster consists of 2,551 nodes and 17,381 edges. The number of nodes corresponds to 88% of the total mapped cancer genes. The rest formed isolated nodes or very small clusters. The network was generated by Cytoscape.

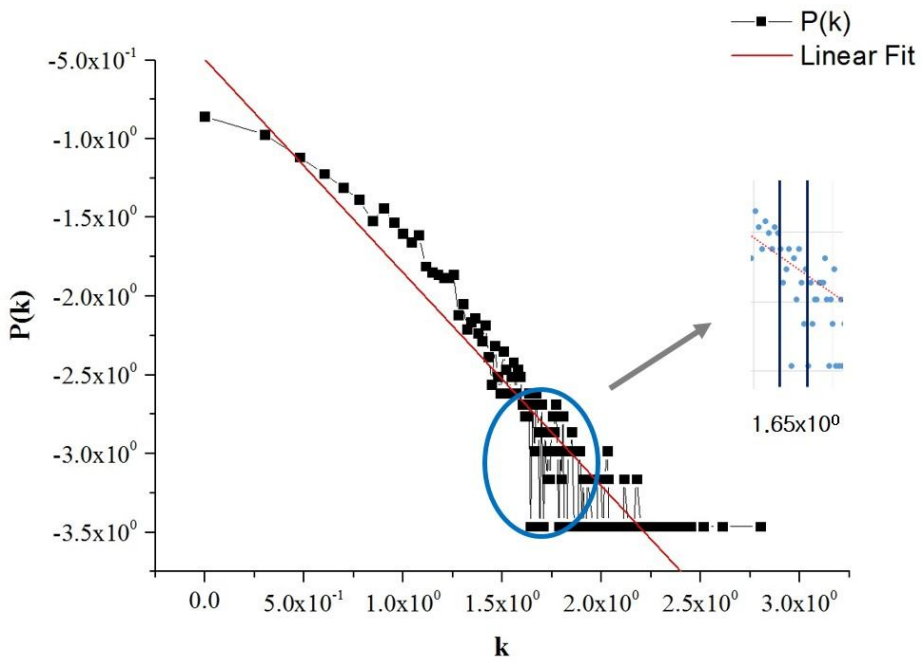

Figure 3.10 Degree distribution of cancer-specific network based on power-law and scale-free nature. The network degree exponent value was 1.36. The scale-free nature indicates the inequality of the degree distribution, such that nodes of a minority have a high degree in the network. The blue circle is the area for ranking hub genes.

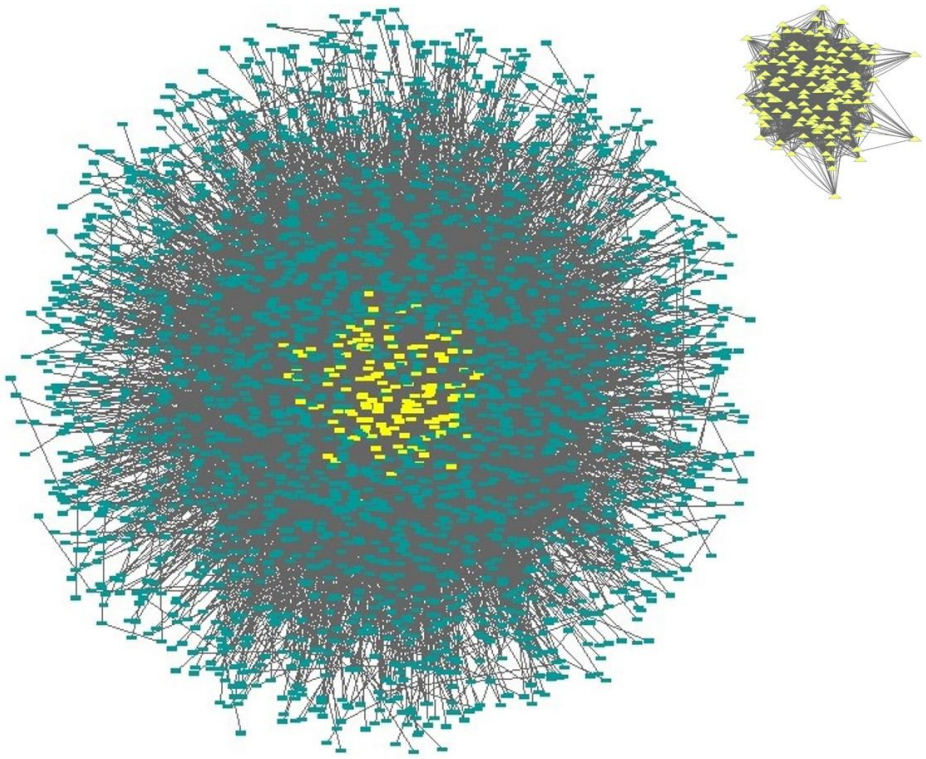

Figure 3.11 Top 150 hub genes from inside the major cluster. In the network, the location of the Top 150 hub genes is shown by the yellow node of the center, and these are all included in the major cluster. The small network on the upper right indicates the network of the hub genes. The network was generated by Cytoscape.

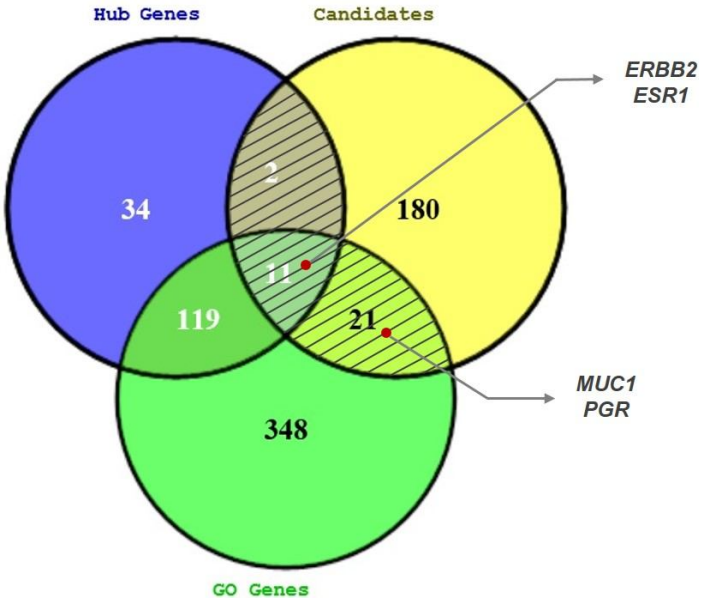

Figure 3.12 Specific target area consisting of 34 genes. The hatched region includes the targets that are reflected characteristics of ontology and the network. In total, 34 genes were selected as the final targets. ERBB2 and ESR1 are found in the GO area, and MUC1 and PGR were discovered in a common area of the network and GO.

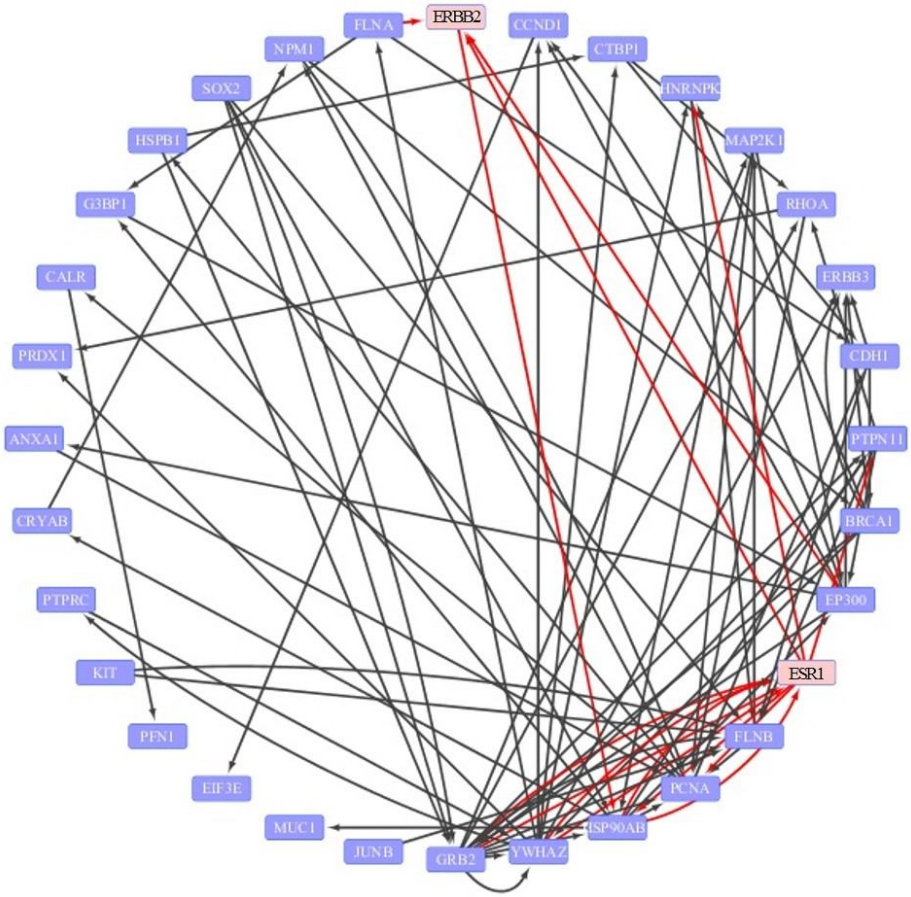

Figure 3.13 Sub-network for 34 target genes. This is a sub-network of only the target genes. Pink nodes indicate clinical markers ERBB2 and ESR1. Their interactions with other nodes are indicated by red lines, also the arrows indicate the direction for target nodes. Here, FAS and PGR genes are not included in the same cluster.

Table 3.1 Top 150 hub genes list is composed of 166 genes

No.	Genes
0 ~ 50 Hub genes	**ABL1**, **ACTB**, AKT1, ALB, APC, **AR**, ARF6, ATF2, ATF7IP, ATM, ATP1A1, BRCA1, BTRC, CASP8, CAV1, **CBL**, CDC42, **CDC5L**, CDK1, **CDK2**, CDK6, CDKN1A, CDKN2A, CHD3, CLTC, **COPS5**, **CREBBP**, **CRK**, **CSNK2A1**, CSNK2B, **CTNNB1**, **CUL1**, **CUL2**, **CUL3**, CUL4A, **CUL4B**, **CUL5**, DCUN1D1, DHX9, DLG4, E2F1, **EEF1A1**, **EGFR**, EIF1B, **ELAVL1**, ENO1, **EP300**, EPS15, ERBB2, **ESR1**
51~ 100 Hub genes	**ESR2**, EWSR1, FLNA, FLNB, **FLNC**, **FN1**, FOS, **FYN**, GNB2L1, **GRB2**, **GSK3B**, H2AFX, **HDAC2**, HDAC3, HDAC4, HDAC6, **HLA-B**, HNRNPK, **HSP90AA1**, **HSP90AB1**, HSPA5, HSPA9, JUN, KAT5, LCK, LYN, MAP3K1, MAP3K7, MAPK1, MAPK14, MAPK3, MAPK6, MAPK8, MCC, MCM2, **MDM2**, MLH1, MTOR, **MYC**, NCK1, NCOA3, NCOR1, NFKB1, NFKB2, NPM1, NUP98, PABPC1, PCNA, **PIK3R1**, PIN1
101 ~ 166 Hub genes	PLK1, PML, POT1, **PPP1CA**, PPP1CC, PPP2R1A, PPP2R2B, PRKAB1, PRKCD, PSMD2, PTN, PTPN11, PXN, RAC1, RAD21, RAF1, RAN, RAR, RB1, RBBP7, RELA, RPL5, RUVBL1, RUVBL2, **SHC1**, SIN3A, SIRT1, SKP1, **SMAD2**, **SMAD3**, SMAD4, SMARCA4, SMARCAD1, SMARCB1, SP1, **SRC**, SRRM1, STAT3, SYK, TCF3, TGFBR1, TK1, TNFRSF1A, TOP1, **TP53**, TP63, TP79, TPR, TSC22D1, TSG101, TUBB, UBE2D1, UBE2D2, UBE3A, UCHL5, USP7, VCL, **VHL**, XPO1, XRCC6, YBX1, YWHAB, YWHAE, **YWHAQ**, **YWHAZ**, ZBTB16

This network was obtained from the degree, betweenness, and stress parameters. The total 166 genes belong to the top rank 150. (red letter: degree 300 ↑ nodes; blue letter: degree 200 ↑ nodes; green letter: degree 100 ↑ nodes).

Table 3.2 Specific GO terms of breast cancer–related genes

GO terms	P-value	Enrichment	Representative Genes
Biological Process			
GO:0085020	1.57E-04	78.69	BARD1, BRCA1
GO:0045922	4.67E-04	52.46	AKT1, BRCA1
GO:0014044	4.67E-04	52.46	NF1, MED12
GO:0071680	9.27E-04	39.35	CDH1, BRCA1
GO:0071681	9.27E-04	39.35	CDH1, BRCA1
GO:0034349	9.27E-04	39.35	TP53, RB1
GO:1902101	9.27E-04	39.35	RB1, PTEN
GO:0045842	9.27E-04	39.35	RB1, PTEN
GO:0033600	9.27E-04	39.35	GATA3, BRCA2
GO:0007406	9.27E-04	39.35	TP53, NF1
...			
Molecular function			
GO:0008545	4.67E-04	52.46	MAP2K4, MAP3K1
GO:0008013	6.13E-04	9.84	AR, CDH1, MED12, APC
Cellular component			
GO:0031436	1.57E-04	78.69	BARD1, BRCA1
GO:0030870	4.67E-04	52.46	NBN, RAD50
GO:0000784	9.27E-04	39.35	NBN, RAD50
GO:0070531	9.27E-04	39.35	BARD1, BRCA1
GO:0000781	3.59E-05	19.67	CHEK2, ATM, NBN, RAD50
GO:0098687	4.65E-05	18.52	CHEK2, ATM, NBN, RAD50
GO:0005913	7.72E-04	15.74	CDH1, APC, MYH9
GO:0016605	4.04E-06	13.12	RAD51, CHEK2, TP53, RB1, PTEN, NBN
GO:0016604	9.90E-04	5.02	TP53, CHEK2, RAD51, RB1, PTEN, NBN

List of selected GO terms that appear specifically in breast cancer–related genes. The biological process is only shown for 10 of the total 100. Molecular function selected two GO terms, and cellular component selected nine GO terms. These have been sorted in descending order based on the score of the enrichment.

Table 3.3 34 target genes list and presence of a clinical marker in each area

Area	Target Genes	total No.	Clinical marker
Hub Genes			
	FLNB, HNRNPK	2	-
GO Genes			
	ANXA1, CALR, CCND1, CDH1, CRYAB, CTBP1, EIF3E, ERBB3, FAS, G3BP1, HSPB1, JUNB, KIT, MAP2K1, MUC1, PFN1, PGR, PRDX1, PTPRC, RHOA, SOX2	21	*MUC1, PGR*
Hub + GO Genes			
	BRCA1, EP300, ERBB2, ESR1, FLNA, GRB2, HSP90AB1, NPM1, PCNA, PTPN11, YWHAZ	11	*ERBB2, ESR1*

This is a detailed list of selected targets in each area. ERBB2 and ESR1 were found in the GO area, MUC1 and PGR were discovered in a common area of the network and GO.

Table 3.4 Gene ontology characteristics of the 34 target genes

GO terms	Description	%
Biological process		
GO:0008152	metabolic process	25.9
GO:0009987	cellular process	21.2
GO:0065007	biological regulation	12.9
GO:0032502	developmental process	9.4
GO:0002376	immune system process	7.1
GO:0050896	response to stimulus	5.9
GO:0006915	apoptotic process	4.7
GO:0051179	localization	3.5
GO:0032501	multicellular organismal process	3.5
GO:0022610	biological adhesion	2.4
GO:0071840	cellular component organization or biogenesis	2.4
GO:0000003	reproduction	1.2
Molecular function		
GO:0005488	binding	35.4
GO:0003824	catalytic activity	22.9
GO:0001071	nucleic acid binding transcription factor activity	12.5
GO:0004872	receptor activity	8.3
GO:0005198	structural molecular activity	8.3
GO:0030234	enzyme regulator activity	4.2
GO:0000988	protein binding transcription factor activity	4.2
GO:0016209	antioxidant activity	2.1
GO:0045182	translation regulator activity	2.1
Cellular component		
GO:0044464	cell part	50.0
GO:0043226	organelle	30.0
GO:0032991	macromolecular complex	10.0
GO:0016020	membrane	10.0

This GO terms belongs to the three categories of the selected targets. It has been sorted in descending order based on the percentage of GO terms in each category.

Table 3.5 A list of genes that are displayed on the cancer map of KEGG

KEGG ID	Pathway	Genes
hsa05206	MicroRNAs in cancer	BRCA1, CCND1, EP300, ERBB2, ERBB3, GRB2, HNRNPK, MAP2K1, RHOA
hsa05205	Proteoglycans in cancer	CCND1, ERBB2, ERBB3, FLNA, GRB2, MAP2K1, PTPN11, RHOA
hsa05200	Pathways in cancer	CCND1, EP300, ERBB2, GRB2, KIT, MAP2K1, RHOA
hsa05203	Viral carcinogenesis	CCND1, EP300, GRB2, HNRNPK, RHOA
hsa05215	Prostate cancer	CCND1, EP300, ERBB2, GRB2, MAP2K1
hsa05213	Endometrial cancer	CCND1, ERBB2, GRB2, MAP2K1
hsa05211	Renal cell carcinoma	EP300, GRB2, MAP2K1, PTPN11
hsa05223	Non-small cell lung cancer	CCND1, ERBB2, GRB2, MAP2K1
hsa05220	Chronic myeloid leukemia	CCND1, GRB2, MAP2K1, PTPN11
hsa05210	Colorectal cancer	CCND1, MAP2K1, RHOA
hsa05212	Pancreatic cancer	CCND1, ERBB2, MAP2K1
hsa05219	Bladder cancer	CCND1, ERBB2, MAP2K1
hsa05218	Melanoma	CCND1, MAP2K1
hsa05216	Thyroid cancer	CCND1, MAP2K1
hsa05222	Small cell lung cancer	CCND1

Twelve of the 34 genes are shown in the 15 different cancer maps of the KEGG database. Particularly, CCND1 and MAP2K1 showed up in nearly all pathways. CCND1 is involved in the cell cycle, and MAP2K1 is involved in mitosis.

CHAPTER IV.

DISCUSSION

So far, research in the investigation for target genes that are likely candidates as breast cancer markers was performed by utilizing the PPI network and gene ontology information. In this section, the selected targets, their sub-network, features of the gene ontology, and the result of the cancer pathway of KEGG are to be examined more closely.

In clinical markers for confirming the results, MUC1 and PGR had breast cancer-specific GO terms, and ERBB2 and ESR1 turned out to be hub genes in the network and had the breast cancer-specific GO terms at the same time. However, CEACAM5 was not confirmed here. In the case of CEACAM5, because this marker is shown in various cancers it decreases in cancer specificity; therefore, it is omitted from the results of this research. In fact, CEACAM5 is an antigen first discovered in colorectal cancer, and yet it is expressed highly in stomach, breast, thyroid, pancreatic, liver, gastric, renal, ovarian, and cervical cancer, and thus could be applied to various cancers.

Next, the sub-network of selected targets was discussed. This was produced by mapping the 34 target genes ID in the protein interaction data; as a result, a cluster was formed among genes without FAS and PGR. FAS and PGR showed features according to GO terms, and do not affect the overall results. This is seen in Figure 3.13. Subsequently, the result according to the percentage of the gene ontology was discussed. In viewing the GO terms with the highest percentage of each category, the metabolic process (GO:0008152) was the highest of the biological process at 25.9%, and had primary metabolic

process (GO:0044238) as sub GO terms. As for molecular function, binding (GO:0005488) was the highest at 35.4%, and had nucleic acid binding (GO:0003676), chromatin binding (GO:0003682), calcium ion binding (GO:0005509), and protein binding (GO:0005515) as sub GO terms. As for cellular component, the cell part (GO:0044464) was as high as 50%, and had intracellular (GO:0005622) as a sub GO term.

In addition, KEGG mapping was carried out on selected targets. As a result, 12 genes were found to be involved in the 15 pathways that directly relate to cancer. Especially, MAP2K1 and CCND1 were revealed in most cancer-related pathways. The pathway with the most genes mapped was MicroRNAs in cancer (hsa05206), followed by Proteoglycans in cancer (hsa05205), Pathways in cancer (hsa05200), Viral carcinogenesis (hsa05203), Prostate cancer (hsa05215), Endometrial cancer (hsa05213), Renal cell carcinoma (hsa05211), Non-small cell lung cancer (hsa05223), Colorectal cancer (hsa05210), Pancreatic cancer (hsa05212), and Bladder cancer (hsa05219). In other pathways associated with cancer, there are cell cycle (hsa04110), PI3K-Akt signaling pathway (hsa04151), Ras signaling pathway (hsa04014), and p53 signaling pathway (hsa04115). Cancer pathways and related gene lists are shown in Table 3.5.

Finally, limitations and utilization of the study are to be discussed. The target of specific diseases, such as biomarkers must be supported by experimental and clinical application, and predicting them among many associated factors is a hard task. For this purpose, utilizing the advantage provided by bioinformatics is important. The analyzed results through this research could be used for understanding the interactions among genes in carcinogenesis, the features of the formed network, the molecular functions, and biological pathways, and explain the common features among markers. These data would be able to provide useful information in creating a biomarker panel by multiple markers and development of new algorithms for

new drugs; they can also contribute to the selection of new markers by applying method to various cancers other than breast cancer. Moreover, the database for this research may be used effectively. In the database, there is important possible biomarker data for head and neck, colon, liver, lung, ovarian, pancreatic, and prostate cancer, and cancer-related gene data such as oncogenes, proto-oncogenes, and tumor suppressor genes. The data is searchable by gene type, chromosomes, and cancer type. To increase the convenience of searching, the database is equipped with various search functions. Additionally, BLAST and other bioinformatics tools were linked on the page, improving the convenience of the utilization. If such research is continued, the unsolved task of finding the solution to cancer would be assisted, and for this purpose, continued supplement and effort of the construed database is desirable.

CHAPTER V.

CONCLUSION AND SUMMARY

5.1 Conclusion

Advances in molecular biology have resulted in significant progress in cancer research and serve as a foundation for research in many fields. The present research sought to derive useful information through accumulated molecular biological data and analysis. Bioinformatics technology was applied, and various data that is used in the research was manipulated. Using this data, the proposed database was constructed. The outcome of the research was the selection of possible targets that reflect the features of breast cancer by analysis of bioinformatics and systems biology.

The reason for this approach is that biological process, molecular function of cancer genes, and their interaction of network were considered important in the expression of cancer. In addition, since research such as network analysis is being widely pursued, the construction of related data offers a good foundation for this research. By mapping the cancer gene ID on protein–protein interaction data, the network and interaction information specialized for cancer was obtained. Through this produced network, the extent of the interaction of central factors and how important they are can be determined. In addition, the detailed role, position, and pathways were identified by GO enrichment analysis. Finally, by referring to the analysis results and the existing clinical markers, targets as markers for breast cancer were sought.

5.2 Summary

In this research, the basic concepts of cancer and the necessity of biomarkers were reviewed, and bioinformatics analysis was applied to provide useful information to find targets as proper biomarkers to solve the problem of cancer. Breast cancer was selected as the type of cancer, and network analysis and gene ontology analysis were performed to explore the systems biological features of specific targets. The analysis is based on the hypothesis that markers are important in gene ontology and cancer network, and that there will be features common to these. In addition, by utilizing candidates considered possible breast cancer biomarkers by experiments, the research sought to select targets that are more potent.

First, for efficient data management and analysis, a research-based database was construed. For this, oncogenes, proto-oncogenes, and tumor suppressor genes as cancer-related gene data, and biomarker data were brought from public databases such as NCBI, UniProt, and EDRN (NCI). Then, data was manipulated by Java s/w to achieve an integrated format. In addition, by saving these data on a MySQL server, efficiency in searching and data download services such as FASTA file were made possible; for this purpose, various search forms were equipped to the database. Computer languages used for the database construction were HTML Java, JavaScript, and JSP, and Tomcat was utilized for the web container. This database, specialized for cancer and biomarkers, was divided into three parts: cancer-related genes, biomarkers, and bioinformatics tool pages such as BLAST. For target search research, various data and bioinformatics tools were utilized based on the constructed database. To create a cancer network, protein interaction data and cancer gene data were needed; accordingly, 13,949 proteins, 122,708 interactions, and 3,399 cancer genes list were used. Next, for gene ontology analysis, 38 breast cancer–related genes and genes to form

the cancer network were used. Additionally, five markers used clinically for breast cancer were included in the research to see if they are important in network and gene ontology, and to verify future research results.

The produced network has 2,902 proteins and 17,383 interactions, and the largest cluster includes 2,551 proteins and 17,381 interactions. To see if the produced network has hub genes, it was checked whether the degree distribution, the basic parameter that reflects on the features of the network, follows power-law distribution. In drawing the log-log graph against the degree k and P(k) value, which gives the possibility that a random node would have the degree of k, it was discovered that it follows the power-law distribution. It is possible that such a scale-free network could explain the existence of the hub, and it can be seen that the produced network contains hub genes. Therefore, to decide on the proper hub genes ranking, degree distribution and betweenness, and stress centrality were calculated, as a result, 166 genes listed as top 150 hub genes were obtained. Subsequently, by performing gene ontology analysis, more important GO terms were filtered in breast cancer compared to general cancer. These GO terms included negative regulation of neuroblast proliferation (GO:0007406) and protein K6-linked ubiquitination (GO:0085020) in BP; beta-catenin binding (GO:0008013) and JUN kinase kinase activity (GO:0008545) in MF; and nuclear chromosome, telomeric region (GO:0000784), cell-cell adherens junction (GO:0005913) and BRCA1-A complex (GO:0070531) in CC. By applying them to cancer network genes, a 499 GO genes list with specific GO terms of breast cancer was obtained. To select targets for breast cancer as the final goal of the research by utilizing all the performed analysis results, the hub genes, specific GO genes, target candidates were expressed in a Venn diagram. The common fields formed among the three groups were set as the field for target selection; thereby, the targets that reflected the features of PPI network and gene ontology were filtered. The locations of the clinical markers were also

checked, since they were shown as the hypothesis; the verification of the research was shown as well. As the result, FLNB and HNRNPK (Hub genes); ANXA1, CALR, CCND1, CDH1, CRYAB, CTBP1, EIF3E, ERBB3, FAS, G3BP1, HSPB1, JUNB, KIT, MAP2K1, MUC1, PFN1, PGR, PRDX1, PTPRC, RHOA, and SOX2 (GO genes); and BRCA1, EP300, ERBB2, ESR1, FLNA, GRB2, HSP90AB1, NPM1, PCNA, PTPN11, and YWHAZ (Hub and GO genes) were selected as targets. This result could be useful information in experimental, clinical research to decide on new breast cancer markers.

BIBLIOGRAPHY

Aaronson SA. Growth factors and cancer. Science. 254(5035):1146–1153 (1991).

Alberts B, Bray D. Hopkin K, Johnson A, Lewis J, Raff M, Roberts K, Walte P. ESSENTIAL CELL BIOLOGY, 3rd Edition. Garland Science. (2009).

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J. Mol. Biol. 215(3):403–410 (1990).

American Cancer Society. Tumor Markers. (2012).

An O, Pendino V, D'Antonio M, Ratti E, Gentilini M, Ciccarelli FD. NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. Database. 2014:bau015 (2014).

Anderson MW, Reynolds SH, You M, Maronpot RM. Role of proto-oncogene activation in carcinogenesis. Environmental Health Perspectives. 98:13–24 (1992).

Andrae J, Gallini R, Betsholtz C. Role of platelet-derived growth factors in physiology and medicine. Genes & Development. 22(10):1276–1312 (2008).

Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT, Kerrien S, Khadake J, Kerssemakers J, Leroy C, Menden M, Michaut M, Montecchi-Palazzi L, Neuhauser SN, Orchard S, Perreau V, Roechert B, van Eijk K, Hermjakob H. The IntAct molecular interaction database in 2010. Nucleic Acids Research. 38(suppl 1):D525–531 (2010).

Assenov Y, Ramirez F, Schelhorn SE, Lengauer T, Albrecht M. Computing topological parameters of biological networks. Bioinformatics. 24(2):282–284 (2008).

Bachman KE, Park BH, Rajagopalan H, Herman JG, Baylin SB, Kinzler KW, Vogelstein B. Histone modifications and silencing prior to DNA methylation of a tumor suppressor gene. Cancer Cell. 3(1):89–95 (2003).

Bader GD, Betel D, Hogue CWV. BIND: the biomolecular interaction network database. Nucleic Acids Research. 31(1):248–250 (2003).

Baker SJ, Markowitz S, Fearon ER, Willson JK, Vogelstein B. Suppression of human colorectal carcinoma cell growth by wild-type p53. Science. 249(4971):912−915 (1990).

Barabasi AL, Albert R. Emergence of scaling in random networks. Science. 286(5439):509−512 (1999).

Barabasi AL, Oltvai ZN. Networks biology: understanding the cell's functional organization. Nature Reviews Genetics. 5(2):101−113 (2004).

Barbacid M. Ras genes. Ann. Rev. Biochem. 56(1):779−827 (1987).

Bast RC, Kufe DW, Pollock RE, Weichselbaum RR, Holland JF, Frei E. HOLLAND-FREI CANCER MEDICINE, 5TH Edition. B.C. Decker. (2000).

Basu T, Gutmann DH, Fletcher JA, Glover TW, Collins FS, Downward J. Aberrant regulation of Ras proteins in malignant tumour cells from type 1 neurofibromatosis patients. Nature. 356(6371):713−715 (1992).

Bates S, Phillips AC, Clark PA, Stott F, Peters G, Ludwig RL, Vousden KH. p14ARF links the tumour suppressors Rb and p53. Nature. 395(6698):124−125 (1998).

Bhatt AN, Mathur R, Farooque A, Verma A, Dwarakanath BS. Cancer biomarkers-current perspectives. Indian J. Med. Res. 132(2):129−149 (2010).

Biggar RJ, Jaffe ES, Goedert JJ, Chaturvedi A, Pfeiffer R, Engels EA. Hodgkin lymphoma and immunodeficiency in persons with HIV/AIDS. Blood. 108(12):3786−3791 (2006).

Biomarker Definition Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clinical Pharmacology & Therapeutics. 69(3):89−95 (2001).

Bird A. DNA methylation patterns and epigenetic memory. Genes & Development. 16(1):6−21 (2002).

Bishop JM. Viral oncogenes. Cell. 42(1):23−38 (1985).

Bishop JM. The molecular genetics of cancer. Science. 235(4786):305−311 (1987).

Blackburn EH, Szostak JW. The molecular structure of centromeres and telomeres.

Ann. Rev. Biochem. 53(1):163−194 (1984).

Bos JL. Ras oncogenes in human cancer: a review. Cancer Research. 49(17):4682−4689 (1989).

Brandes U. A faster algorithm for betweenness centrality. Journal of Mathematical Sociology. 25(2):163−177 (2001).

Brem S, Brem H, Folkman J, Finkelstein D, Patz A. Prolonged tumor dormancy by prevention of neovascularization in the vitreous. Cancer Research. 36(8):2807−2812 (1976).

Bryan TM, Englezou A, Gupta J, Bacchetti S, Reddel RR. Telomere elongation in immortal human cells without detectable telomerase activity. The EMBO Journal. 14(17):4240−4248 (1995).

Bushman Lab (Cancer Gene List, http://www.bushmanlab.org/links/genelists, accessed July 13, 2014).

Cancer Research UK (Breast cancer genes, http://www.cancerresearchuk.org/cancer-help/type/breast-cancer/about/risks/breast-cancer-genes, accessed July 25, 2014).

Cantley LC, Auger KR, Carpenter C, Duckworth B, Graziani A, Kapeller R, Soltoff S. Oncogenes and signal transduction. Cell. 64(2):281−302 (1991).

Carmeliet P, Jain RK. Angiogenesis in cancer and other diseases. Nature. 407(6801):249−257 (2000).

Casciato DA, Territo MC. MANUAL OF CLINICAL ONCOLOGY, 6th Edition. Lippincott Williams & Wikins. (2009).

Ceol A, Aryamontri AC, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L, Cesareni G. MINT, the molecular interaction database: 2009 update. Nucleic Acids Research. 38(Data issue):D532−539 (2009).

Chen JS, Hung WS, Chan HH, Tsai SJ, Sun HS. In silico identification of oncogenic potential of fyn-related kinase in hepatocellular carcinoma. Bioinformatics. 29(4):420−427 (2013).

Chin L, Tam A, Pomerantz J, Wong M, Holash J, Bardeesy N, Shen Q, O'Hagan R, Pantiginis J, Zhou H, Horner JW, Cordo-Cardo C, Yancopoulos GD, DePinho RA.

Essential role for oncogenic Ras in tumour maintenance. Nature. 400(6743):468−472 (1999).

Chirenje ZM. HIV and cancer of the cervix. Best Practice & Research Clinical Obstetrics & Gynaecology. 19(2):269−276 (2005).

Cho Y, Gorina S, Jeffrey PD, Pavletich NP. Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. Science. 265(5170):346−355 (1994).

Crichton DJ, Mattmann CA, Thornquist M, Anton K, Hughes JS. Bioinformatics: biomarkers of early detection. Cancer Biomarkers. 9(1):511−530 (2011).

Croce CM. Oncogenes and cancer. N. ENGL. J. MED. 358(5):502−511 (2008).

Donehower LA, Harvey M, Slagle BL, McArthur MJ, Montgomery CA, Butel JS, Bradley A. Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours. Nature. 356(6366):215−221 (1992).

Doolittle RF, Hunkapiller MW, Hood LE, Devare SG, Robbins KC, Aaronson SA, Antoniades HN. Simian sarcoma virus onc gene, v-Sis, is derived from the gene (or genes) encoding a platelet-derived growth factor. Science. 221(4607):275−277 (1983).

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 10(1):48−54 (2009).

Ehrlich M. DNA methylation in cancer: too much, but also too little. Oncogene. 21(35): 5400−5413 (2002).

Elledge SJ. Cell cycle checkpoints: preventing an identity crisis. Science. 274(5293):1664−1672 (1996).

Epstein MA, Achong BG, Barr YM. Virus particles in cultured lymphoblasts from burkitt's lymphoma. The Lancet. 283(7335):702−703 (1964).

Esteller M. Cancer epigenomics: DNA methylomes and histone-modification maps. Nature Reviews Genetics. 8(4):286−298 (2007).

FDA (Drugs, http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm08337

8.htm, accessed July 25, 2014).

Feinberg AP, Tycko B. The history of cancer epigenetics. Nature Reviews Cancer. 4(2): 143−153 (2004).

Fidler IJ. The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. Nature Reviews Cancer. 3(6):45−458 (2003).

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M. Pfam: the protein families database. Nucleic Acids Research. 42(suppl 1):D222−230 (2014).

Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, Stratton MR, Futreal A. COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Research. 39(suppl 1):D945−950 (2011).

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. Nature Reviews Cancer. 4(3):177−183 (2004).

Genetics Home Reference (Breast cancer, http://ghr.nlm.nih.gov/condition/breast-cancer, accessed July 25, 2014).

Gray KA, Daugherty LC, Gordon SM, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2013. Nucleic Acids Research. 41(suppl 1):D545−552 (2013).

Gupta GP, Massague J. Cancer metastasis: building a framework. Cell. 127(4):679−695 (2006).

Hanahan D, Folkman J. Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis. Cell. 86(3):353−364 (1996).

Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 100(1):57−70 (2000).

Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 144(5):646−674 (2011).

Harley CB. Telomere loss: mitotic clock or genetic time bomb?. Mutation Research. 256(2):271−282 (1991).

Hartwell LH, Weinert TA. Checkpoints: controls that ensure the order of cell cycle events. Science. 246(4930):629–634 (1989).

Hartwell LH, Kastan MB. Cell cycle control and cancer. Science. 266(5192):1821–1828 (1994).

Hausen HZ. Viruses in human cancers. Science. 254(5035):1167–1173 (1991).

He TC, Sparks AB, Rago C, Hermeking H, Zawel L, da Costa LT, Morin PJ, Vogelstein B, Kinzler KW. Identification of c-Myc as a target of the Apc pathway. Science. 281(5382):1509–1512 (1998).

He Y, Zhang M, Ju Y, Yu Z, Lv D, Sun H, Yuan W, He F, Zhang J, Li H, Li J, Wang-Sattler R, Li Y, Zhang G, Xie L. dbDEPC 2.0: updated database of differentially expressed proteins in human cancers. Nucleic Acids Research. 40(suppl 1):D964–971 (2012).

Hirohashi S, Kanai Y. Cell adhesion system and human cancer morphogenesis. Cancer Sci. 94(7):575–581 (2003).

Hochegger H, Takeda S, Hunt T. Cyclin-dependent kinases and cell-cycle transitions: does one fit all?. Nature Reviews Molecular Cell Biology. 9(11):910–916 (2008).

Jensen PB, Hunter T. Oncogenic kinase signalling. Nature. 411(6835):355–365 (2001).

Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. Nature. 411(6833):41–42 (2001).

Jones PA, Laird PW. Cancer epigenetics comes of age. Nature Genetics. 21(2):163–167 (1999).

Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. Nature. Reviews Genetics. 3(6):415–428 (2002).

Jones PA, Baylin SB. The epigenomics of cancer. Cell. 128(4):683–692 (2007).

Jonsson PF, Bates PA. Global topological features of cancer proteins in the human interactome. Bioinformatics. 22(18):2291–2297 (2006).

Joy MP, Brock A, Ingber DE, Huang S. High-betweenness proteins in the yeast protein interaction network. BioMed Research International. 2005(2):96−103 (2005).

Jung KW, Won YJ, Kong HJ, Oh CM, Seo HG, Lee JS. Cancer statistics in Korea: incidence, mortality, survival and prevalence in 2010. Cancer Res. Treat. 45(1):1−14 (2013).

Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research. 28(1):27−30 (2000).

Kerr JFR, Wyllie AH, Curriem AR. Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. Br. J. Cancer. 26(4):239−257 (1972).

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Abdul Rahiman B, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. Human protein reference database-2009 update. Nucleic Acids Research. 37(suppl 1):D767−772 (2009).

Kim NW, Piatyszek MA, Prowse KR, Harley CB, West MD, Ho PL, Coviello GM, Wright WE, Weinrich SL, Shay JW. Specific association of human telomerase activity with immortal cells and cancer. Science. 266(5193):2011−2015 (1994).

Knudson AG. Mutation and cancer: statistical study of retinoblastoma. PNAS. 68(4): 820−823 (1971).

Kubbutat Michael HG, Jones SN, Vousden KH. Regulation of p53 stability by Mdm2. Nature. 387(6630):299−303 (1997).

Lane DP. p53, guardian of the genome. Nature. 358(6381):15−16 (1992).

Levine AJ. p53, the Cellular gatekeeper for growth and division. Cell. 88(3):323−331 (1997).

Llyin SE, Belkowski SM, Plata-Salaman CR. Biomarker discovery and validation: technologies and integrative approaches. Trends in Biotechnology. 22(8):411−416 (2004).

Loeb LA. Mutator phenotype may be required for multistage carcinogenesis. Cancer

Research. 51(12):3075−3079 (1991).

Lukas J, Parry D, Aagaard L, Mann DJ, Bartkova J, Strauss M, Peters G, Bartek J. Retinoblastoma-protein-dependent cell-cycle inhibition by the tumour suppressor p16. Nature. 375(6531):503−506 (1995).

Lukashev ME, Werb Z. ECM signaling: orchestrating cell behaviour and misbehaviour. Trends in Cell Biology. 8(11):437−441 (1998).

Madden SL, Cook DM, Morris JF, Gashler A, Sukhatme VP, Rauscher FJ. Transcriptional repression mediated by the WT1 wilms tumor gene product. Science. 253(5027):1550−1553 (1991).

Malumbers M, Barbacid M. Ras oncogenes: the first 30 years. Nature Reviews Cancer. 3(6):459−465 (2003).

Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science. 298(5600):1912−1934 (2002).

Markowitz S. DNA repair defects inactivate tumor suppressor genes and induce hereditary and sporadic colon cancers. Journal of Clinical Oncology. 18(21):75−80 (2000).

Martin GS. Rous sarcoma virus: a function required for the maintenance of the transformed state. Nature. 227(5262):1021−1023 (1970).

Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell MJ, Kitano H, Thomas PD. The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Research. 33(suppl 1):D284−288 (2005).

Moore RG, Brown AK, Miller MC, Skates S, Allard WJ, Verch T, Steinhoff M, Messerlian G, DiSilvestro P, Granai CO, Bast Jr RC. The use of multiple novel tumor biomarkers for the detection of ovarian carcinoma in patients with a pelvic mass. Gynecologic Oncology. 108(2):402−408 (2008).

Morgan DO. Cyclin-dependent kinases: engines, clocks, and microprocessors. Annu. Rev. Cell Dev. Biol. 13(1):261−291 (1997).

Morin GB. The human telomerase terminal transferase enzyme is a ribonucleoprotein that synthesizes TTAGGG repeats. Cell. 59(3):521−529 (1989).

Morin PJ, Sparks AB, Korinek V, Barker N, Clevers H, Vogelstein B, Kinzler KW. Activation of β-catenin-Tcf signaling in colon cancer by mutations in β-catenin or Apc. Science. 275(5307):1787−1790 (1997).

Munemitsu S, Albert I, Souza B, Rubinfeld B, Polakis P. Regulation of intracellular β-catenin levels by the adenomatous polyposis coli (APC) tumor-suppressor protein. PNAS. 92(7):3046−3050 (1995).

Murphree AL, Benedict WF. Retinoblastoma: clues to human oncogenesis. Science. 223(4640):1028−1033 (1984).

NCI (Tumor Markers, http://www.cancer.gov/cancertopics/factsheet/detection/tumor-markers, accessed July 25, 2014).

Normanno N, Luca AD, Bianco C, Strizzi L, Mancino M, Maiello MR, Carotenuto A, Feo GD, Caponigro F, Salomon DS. Epidermal growth factor receptor (EGFR) signaling in cancer. Gene. 366(1):2−16 (2006).

Nurse P. A long twentieth century of the cell cycle and beyond. Cell. 100(1):71−78 (2000).

Oliveros JC. VENNY. an interactive tool for comparing lists with venn diagrams. (2007).

Onnela JP, Saramaki J, Hyvonen J, Szabo G, Lazer D, Kaski K, Kertesz J, Barabasi AL. Structure and tie strengths in mobile communication networks. PNAS. 104(18):7332−7336 (2007).

Origin (OriginLab, Northampton, MA).

Parangi S, O'Reilly M, Christofori G, Holmgeren L, Grosfeld J, Folkman J, Hanahan D. Antiangiogenic therapy of transgenic mice impairs de novo tumor growth. PNAS. 93(5):2002−2007 (1996).

Pardee AB. A restriction point for control of normal animal cell proliferation. PNAS. 71(4):1286−1290 (1974).

Roche Diagnostics Global Tumor Marker Workshop. (2013).

Salmon DS, Brandt R, Ciardiello F, Normanno N. Epidermal growth factor-related

peptides and their receptors in human malignancies. Critical Reviews in Oncology/Hematology. 19(3):183−232 (1995).

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The database of interacting proteins: 2004 update. Nucleic Acids Research. 32(suppl 1):D449−451 (2004).

Schaefer MH, Fontaine JF, Vinayagam A, Porras P, Wanker EE, Andrade-Navarro MA. HIPPIE: integrating protein interaction networks with experiment based quality scores. PLoS one. 7(2):e31826 (2012).

Scheffzek K, Anmadian MR, Kabsch W, Wiesmuller L, Lautwein A, Schmitz F, Wittinghofer A. The Ras-RasGAP complex: structural basis for GTPase activation and its loss in oncogenic Ras mutants. Science. 277(5324):333−339 (1997).

Schlessinger J. Cell signaling by receptor tyrosine kinases. Cell. 103(2):211−225 (2000).

Serrano M, Hannon GJ, Beach D. A new regulatory motif in cell-cycle control causing specific inhibition of cyclin D/CDK4. Nature. 366(6456):704−707 (1993).

Sevik M. Oncogenic viruses and mechanisms of oncogenesis. Turk. J. Vet. Anim. Sci. 36(4):323−329 (2012).

Sherr CJ. Cancer cell cycles. Science. 274(5293):1672−1677 (1996).

Sherr CJ. Principles of tumor suppression. Cell. 116(2):235−246 (2004).

Sherr CJ, McCormick F. The Rb and p53 pathways in cancer. Cancer Cell. 2(2):103−112 (2002).

Shimbel A. Structural parameters of communication networks. Bulletin of Mathematical Biophysics. 15(4):501−507 (1953).

Siegel R, Naishadham D, Jemal A. Cancer statistics, 2013. CA: A Cancer Journal for Clinicians 63(1):11−30 (2013).

Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, van Auken K, Wang X, Shi X, Reguly T, Rust JM, Winter A, Dolinski K, Tyers M. The BioGRID interaction database: 2011 update. Nucleic Acids Research. 39(suppl 1):698−704 (2011).

Steeg PS. Metastasis suppressors alter the signal transduction of cancer cells. Nature Reviews Cancer. 3(1):55−63 (2003).

The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nature Genetics. 25(1):25−29 (2000).

The UniProt Consortium. Activities at the universal protein resource (UniProt). Nucleic Acids Research. 42(suppl 1):D191−198 (2014).

Thompson CB. Apoptosis in the pathogenesis and treatment of disease. Science. 267(5203): 1456−1462 (1995).

Vashisht S, Bagler G. An approach for the identification of targets specific to bone metastasis using cancer genes interactome and gene ontology analysis. PLoS one. 7(11): e49401 (2012).

Vaux DL, Korsmeyer SJ. Cell death in development. Cell. 96(2):245−254 (1999).

Vigneri P, Wang JY. Induction of apoptosis in chronic myelogenous leukemia cells through nuclear entrapment of BCR-ABL tyrosine kinase. Nature Medicine. 7(2):228−234 (2001).

Vogelstein B, Lane D, Levine AJ. Surfing the p53 network. Nature. 408(6810):307−310 (2000).

Vogelstein B, Kinzler KW. The multistep nature of cancer. Trends in Genetics. 9(4):138−141 (1993).

Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nature Medicine. 10(8):789−799 (2004).

Wang J, Duncan D, Shi Z, Zhang B. Web-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Research. 41(Web Server issue):W77−83 (2013).

Watt FM. The extracellular matrix and cell shape. Trends in Biochemical Sciences. 11(11):482−485 (1986).

Weinberg RA. Tumor suppressor genes. Science. 254(5035):1138−1146 (1991).

Weinberg RA. The retinoblastoma protein and cell cycle control. Cell. 81(3):323−330 (1995).

Weinberg RA. THE BIOLOGY OF CANCER. Garland Science. (2006).

Yeatman TJ. A renaissance for Src. Nature Reviews Cancer. 4(6):470−480 (2004).

Yokota J. Tumor progression and metastasis. Carcinogenesis. 21(3):497−503 (2000)

Yonishi-Roucah E, Resnitzky D, Lotem J, Sachs L, Kimchi A, Oren M. Wild-type p53 induces apoptosis of myeloid leukaemic cells that is inhibited by interleukin-6. Nature. 353(6333):345−347 (1991).

Yoon J, Blumer A, Lee K. An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality. Bioinformatics. 22(24):3106−3108 (2006).

Yoshida BA, Sokoloff MM, Welch DR, Rinker-Schaeffer CW. Metastasis-suppressor genes: a review and perspective on an emerging field. Journal of the National Cancer Institute. 92(21):1717−1730 (2000).

Young LS, Rickinson AB. Epstein-barr virus: 40 years on. Nature Reviews Cancer. 4(10): 757−768 (2004).

Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Computational Biology. 3(4):e59 (2007).

Zhu L. Tumour suppressor retinoblastoma protein Rb: a transcriptional regulator. European Journal of Cancer. 41(16):2415−2427 (2005).

ABSTRACT (Korean)

Bioinformatics Research on the Investigation of Target Genes for Breast Cancer

바이오인포매틱스 기법을 활용한 유방암
타겟 유전자 탐색 연구

제 미 경
서울대학교 보건대학원 보건학과
바이오인포매틱스 전공

암이라는 질병은 전 세계적으로 여전히 문제가 되고 있으며, 시급한 해결이 요구되는 질병이다. 암을 대상으로 한 수 많은 연구가 이루어져 왔고, 많은 진단법과 치료법들이 제시되어 왔지만 현재까지 완전한 해결책은 없는 실정이다. 바이오마커 역시 암이라는 질병이 가지는 문제점들에 대한 하나의 해결책으로 유용하게 사용되고 있고, 새로운 마커를 발견하기 위한 연구가 다방면으로 이루어지고 있다. 이러한 연구에서, 바이오인포매틱스는 마커로서의 타겟을 발견하기 위해 암과 관련된 네트워크 및 유전자 온톨로지 분석에 사용될 수 있다. 암은 유형별로 특성이 다르고, 임상적으로도 접근방법이 다르기 때문에 본 연구에서는 유방암이라는 하나의 암을 대상으로 한정하여 진행하였다. 먼저, 효율적인 데이터 관리와 분석 기반 마련을 위해 암과 바이오마커에 특화된 데이터베이스를 구축하였다. 데이터베이스 구축에 필요한 데이터는 NCBI, NCI, UniProt과 같은 여러 대규모 공공 저장소로부터 가져왔으며, Html, Java, JavaScript, MySQL, JSP 등의 프로그래밍 언어들을 활용하여 목적에 맞게 가공하였다. 다음으로 수집된 데이터들을 활용하여 암 네트워크를 생성하였다. 이는 단백질 상호작용 데이터에 암 유전자 ID를 mapping

시킴으로써 생성되었으며, 생성된 네트워크가 scale-free nature를 가지는지 확인하기 위해 degree distribution의 멱함수(power-law) 분포 여부를 살펴보았다. 그 결과 생성된 암 네트워크가 멱함수 분포를 보임으로써 hub의 존재를 확인할 수 있었으며, 그래프 특성을 반영하여 hub genes의 rank를 결정하였다. 이에 따라 degree distribution, betweenness centrality, stress centrality와 같은 네트워크 파라미터를 계산하여 Top 150의 hub genes을 얻었다. 다음 분석으로 유방암 관련 주요 유전자들과 암 네트워크를 형성하는 2,902개의 유전자들을 활용한 GO enrichment analysis를 수행하여 유방암의 특성을 반영한 specific GO terms을 필터링하였다. 이를 암 네트워크의 2,902개의 유전자에 적용시켜 결과적으로 암 네트워크에 관여하면서 특히, 유방암 specific GO terms를 가지는 유전자 리스트를 얻었다. 분석의 마지막 단계에서 PPI network analysis와 GO enrichment analysis로 획득된 유전자들과 타겟 후보자들이 형성하는 Venn diagram 분석을 통해 타겟이 선정되었으며, 여기에는 임상에서 사용되고 있는 마커 다섯 개에 대한 결과가 참조되었다. 최종적으로 유방암의 임상 마커들과 유사한 특성을 가지면서 PPI network와 GO에서 유의하게 나타나는 34개의 유전자가 타겟으로서 선별되었으며, 이들이 다음 유방암 마커 후보로서 우선적으로 이용될 수 있으리라 기대하는 바이다. 또한 이러한 방법이 마커를 결정하는데 있어 잠정적인 후보자를 걸러줄 수 있다면, 실험적으로나 임상적으로나 유용할 것이라 생각된다.

주요어: 데이터베이스, 마커, 바이오인포매틱스, 유방암, 유전자 온톨로지, 타겟, PPI 네트워크

학 번: 2013-21873