The Development of Cancer Model System on Metabolism and Tumor Suppressor Effect Analysis using Systems Biology Approach

Young Sun Kim

The Graduate School

Yonsei University

Department of Biostatistics and Computing

The Development of Cancer Model System on Metabolism and Tumor Suppressor Effect Analysis using Systems Biology Approach

A Dissertation

Submitted to the Department of Biostatistics and Computing

and the Graduate School of Yonsei University

in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy

Young Sun Kim

December 2007

This certifies that the dissertation of

Young Sun Kim is approved.

Sang Won Han : Thesis Supervisor

Chang No Yoon : Thesis Committee Member #1

Ki Jun Song : Thesis Committee Member #2

Chang Bae Jin : Thesis Committee Member #3

Eun Jig Lee : Thesis Committee Member #4

The Graduate School

Yonsei University

December 2007

CONTENTS

Table and Figure Index

Abstract

The Development of Cancer Model System on Metabolism and Tumor Suppressor Effect Analysis using Systems Biology Approach

Kim Young Sun
Department of Biostatistics and Computing
The Graduate School
Yonsei University

Systems Biology is interpreted as systematic modeling the metabolism in the cells of an organism, the process of gene regulation and the signal transduction system. Systematic approach is solving problems by using the pathway made up of interactions as a base. Therefore, a hypothesis on the various phenomena seen in the cells has to be modeling and inferred by a systematic approach. This study researched the risk factor metabolism and genes in the cancer development process by using metabolic pathway, signaling pathway and disease pathway. A compound and an enzyme which are basic elements of a pathway were divided and studied by case by case for a systematic study. The first case studied the existence of correlation with thyroid disease by measuring the hormone profile regarding androgen and estrogen pathways. The second case estimated the hormones that become a risk factor when thyroid cancer

develops and studied about the pattern of risk factor hormones in the cancer development process. The third case studied the process of finding metabolisms that have a correlation with colon cancer. The fourth case studied the degree and order of influence BRCA1 gene, a breast cancer tumor suppressor, has on genes in the cell signaling pathway. As a result of the first case, thyroid gland disease decision-making system was developed and thyroid gland disease patients were able to be distinguishedby using the profile of androgen and estrogen metabolisms. In the results of the second case, 2-hydroxyestrone, 2-hydroxyestradiol, 2-methoxyestrone, 2-methoxyestradiol and 2-methoxyestradiol-3-methylether were estimated as risk factor hormones of thyroid cancer and an increase of frequency pattern during the development of cancer could be seen. The results from the third case showed that it was estimated that dichlorobenzene metabolism and styrene metabolism in xenobiotics metabolism have a correlation with colon cancer. It was seen in the results of the fourth case that there was a difference in the degree and order of influence in the BRCA1 gene, a tumor suppressor, depending on the genes in SWI/SNF (DNA regulatory complex), checkpoint arrest and transcription factor.

Key words : Thyroid Cancer, Colon Cancer, Breast Cancer, Androgen and Extrogen Pathway, Xenobiotics Metabolism, Tumor Suppressor, Decision-Making Systems, Lyapunov.

I. Introduction

1.1 Systems Biology

Systems Biology is a study of interpreting the interaction relationship by mathematically and statistically modeling the metabolism in the cells of an organism, the process of gene control and the signal-transduction system. [1] The information on the structure and function of constituents in cells, statistics and computer simulation technique are being demanded for this study and Bio-transformation can be mathematically modeling. Although the analysis unit on the existing disease system is limited to only one gene, systematic approach is solving this problem by using the pathway made up of interactions as a base. The ultimate goal of Systems Biology in relation to the study of diseases is focused on the prevention and treatment of diseases and also on the development of a new treatment. To achieve this goal, the model is extracted from the complex structure of genes or compounds classified as risk factors and decisions are made. [2] Also, different databases are being integrated and required data are being extracted. This study intends to analyze the compounds and enzymes in the metabolic and signaling pathways concerning the condition when there is cancer development. Metabolism stands for a metabolism pathway concerning the enzyme reaction of chemicals as a chemical process related to supplying required energy for a life process and synthesizing a new maintain substance. It also stands for a control pathway of the interaction and response of macro-molecule substance. [3] A metabolic pathway is composed of metabolites and enzymes above the network structure. The metabolites undergo the catabolism and anabolism processes and enzymes begin to take part. A metabolite is used to maintain the homeostasis of a specific function in the body. Up until recently, the purpose of the metabolic pathway analysis was to either

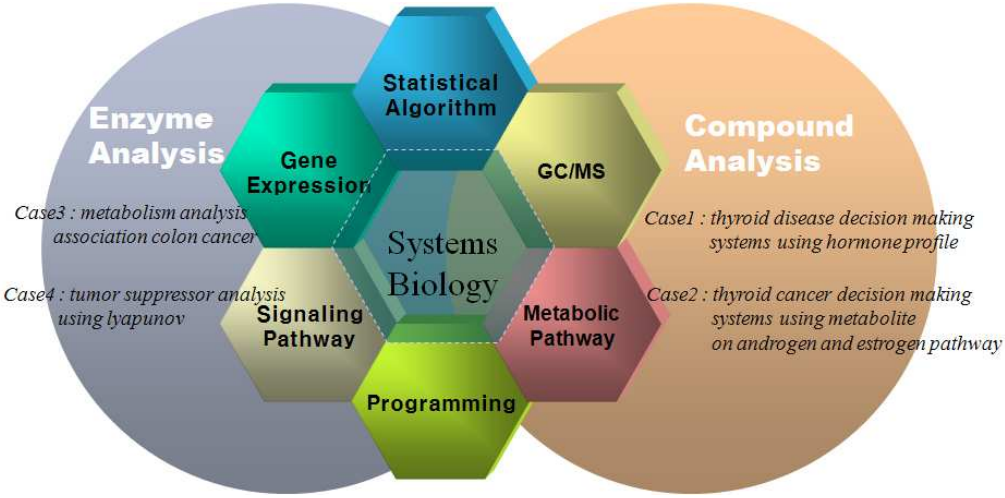
construct metabolic pathway or to analyze metabolic control but now the analysis is used for studies on diseases. The construction of metabolic pathway is a method of calculating possible pathways while eliminating unnecessary substances by drawing up a simultaneous linear equation and making compounds into variables. [4] The metabolic control analysis uses a partial differential equation similar to the Michaelis-Menten model by using the concentration of either a substance or equilibrium constant of an enzyme reaction [5]. A signaling pathway where a pathway maintaining its life by receiving signals from the outside is shown also consists of signal transduction substances and enzymes. A signaling pathway demonstrates various interactions that maintain the homeostasis of cells and individuals by correctly transmitting signals in each situation based on specificity, amplification, desensitization and integration. It plays the role of controlling a cell cycle and when there is a fault, a tumor develops. The study on the influence and location of the tumor suppressor gene in the pathway concerning tumor development is actively in progress.

1.2 Cancer Analysis Model using Systems Biology

The analysis on cancer where Systems Biology was used as a base, divides into two main points. The first point is the development of a disease diagnosing system for compounds. In detail, it is a decision-making system on the metabolite, a compound in the metabolic pathway, for thyroid gland disease and cancer patients. Case 1 aims to develop a disease diagnosing system by measuring the steroid hormone profile of thyroid gland disease patients. In order to do this, a purposive section of a purposive profile has to be found when the thyroid gland disease develops by measuring the hormone profile, an unknown metabolite, of thyroid gland disease patients. Also, a decision-making system using various distinct models has to be developed. Case 2 is

the analysis of the hormone pattern and development of a thyroid cancer decision-making system by measuring the metabolite that exists in the androgen and estrogen hormone pathway for thyroid cancer patients. The compound analysis uses the measured experimental data by using metabolic pathway structural data and GC/MS as in figure 1. The second point is the analysis of a correlation between an enzyme and cancer. Case 3 intends to find metabolism that have a correlation when colon cancer develops in a macroscopic perspective. Therefore, we aim to analyze whether there is a difference in the metabolism when cancer develops and to study through the variation of each metabolism if there is a correlation with cancer. Case 4 plans to analyze the range of tumor suppressor's influence and the stages of development in a microscopic perspective. Consequently, we aim to examine whether a tumor suppressor plays its role as an attractor by measuring the degree of influence it has on other genes that exist in the cell signaling pathway by using lyapunov exponent.

Figure 1. Category of analysis examples using various databases.



1.3. Introduction of Case 1. Study

: Thyroid Gland Disease Hormone Profile

To find out the genetic factors of outbreak of thyroid gland disease, we studied relationship between the disease and the thyroid gland disease decision-making system (TGDDM system) on the metabolic profile in the steroid hormone map. The first approach on profiling was conducted to steroids in urine, sugars, sugar alcohols and organic acid participated in Krebs cycle [6] and the study on the steroid profile was carried out live [7–11] after introduction of analysis method using glass capillary column and GC–MS. Furthermore, the constituent compound of the profiling was examined by the relative intensity and retention time information in LC–MS/MS [12], and data reduction is conducted through principal component analysis (PCA), which performs the pattern recognition in spectrum domain [13,14]. Contrary to the conventional researches, we analyze the metabolic profiling, which deals with the thyroid gland disease group and the normal group to develop the TGDDM

system. Conventionally, in the case of thyroid cancer, the effectiveness has been studied about the known metabolites, such as 2-hydroxyestrone, 2-hydroxyestradiol, 2-methoxyestrone, 2-methoxyestradiol, and 2-methoxyestradiol-3-methylether [15].

However, we examined a relation between the metabolic profiling and the thyroid gland disease, and developed the TGDDM system and its procedure. We analyzed the recognition method of the spectral patterns using PCA as the disease diagnosis procedures using metabolic profile [12–14], we developed the diagnosis system, which determined the thyroid gland disease group and the normal group based on the PCA scores acquired from above analysis by estimating the thyroid gland disease discriminant model using logistic regression.

1.4. Introduction of Case 2. Study

: Androgen and Estrogen Metabolite about Thyroid Cancer

The thyroid cancer is extremely common disease for women compare to men in South Korea. The cause of the outbreak would be of genetic and environmental factors. One of the environmental factors could be estimated as radiation [15]. A person with actinotherapy in head and neck owing to skin disease could be the main reason for young people and thyroid disease, tonsils fleshiness are likely to lead to the thyroid cancer. There were many thyroid cancer patients in Hiroshima, the city attacked by atomic bomb during the World War II or Chernobyl, Soviet Union. This is proved by an experiment [16]. In this research, we estimated that the cause of thyroid cancer is mainly due to abnormality in androgen and estrogen metabolic pathway and hormone [17,18]. We also learned the changes of metabolite, related to androgen and estrogen metabolic pathway during the spreading of cancer. The study attempted to develop the Thyroid Cancer Decision Making System (TCDMS) which demonstrates significant difference in the thyroid cancer. Moreover, as the thyroid cancer advances, we would like to monitor the hormone pattern that is estimated as the risk factor. Acquiring metabolite that is progressed by the thyroid cancer was difficult in this case. External factor is affected by the risk factor owing that when one is diagnosed with the cancer, then the other one gets cured. This makes it impossible to get data from normal status to cancer process. However, we made time series data through statistical prediction as it is essential in examining disease pattern for researchers of disease to have time series data of metabolite. We estimated a model that has a risk factor for outbreak and made time series data of posterior probability about patients based on the estimated model. We estimated the order of

estimated posterior probability as the degree of cancer and the pattern of metabolite in androgen and estrogen metabolic pathway.

1.5. Introduction of Case 3. Study
: Colon Cancer Gene Expression Data and Metabolism

One of the objects of genome bioinformatics is to understand the operation of metabolic pathway and describe it theoretically. Metabolism is the complete set of all chemical changes that occur in cells and organisms and it is the reaction of an enzyme-catalyst which forms a metabolic pathway. The chemical substances in a cell are dissolved and produced by enzymes, and the old cell is changed to a new one. [19] Here, the relationship between the enzymes and the substances are formed in a network structure and they are composed of 148 metabolic pathways. [20] In lecture studies, metabolic analysis was conducted through Mavrovouniotis technique which qualitatively draws out and enumerates metabolic pathway in an enzyme reaction test [21] and through a quantitative method known as metabolic control analysis. [22] Metabolic control analysis is a model designed to measure the effects of each enzymes on the entire metabolism and an approximation technique was developed based on differential equations such as the Michaelis-Menten model. [23] A method where weight, an element of metabolic pathway, is worked out using the linear apposition equation is sometimes used as well as a method using steady state data which knock-out a gene. However, if there are g members of genes within a metabolic pathway, a problem of requiring at least g^2 linear independent equation occurs. Van Someren's linear model was developed to solve this problem. [24] The object of this

study is to measure the effect on a metabolic pathway using measured gene expression data related to cancer. The essential parts that compose a metabolic pathway are compound and enzyme. [20] The enzyme is composed of genes. This study intends to statistically analyze what kind of change the genes in a metabolic pathway demonstrate when there is environmental variation such as cancer development. It also aims to estimate a metabolic pathway that has correlation with cancer and to analyze the changes of the metabolic pathways when a colon cancer is developed. Since the existing analysis of gene expression data weights its purpose on estimating a gene related to cancer, the level of the analysis unit was gene. Meanwhile, this study aims to test in a broad perspective whether metabolisms have significant reactions by measuring the effect of a metabolic pathway, the field of set of genes. This was defined as metabolic pathway effect and based on other experimental designs, whether there was any time effect when it was measured as time series was analyzed after testing the differences between the normal group and the colon cancer group. In order to carry out this study, gene expression data on colon cancer was used [25]. Also, KEGG [26] database was applied for location information and map information of metabolic pathway and in order to parse this using xml information, c program was used and SAS program was used for ANOVA analysis.

1.6. Introduction of Case 4. Study
: Breast Cancer Tumor Suppressor and SPAL Program

The cause of breast cancer depends on conditions such as one's menopausal status, underweight before menopause and overweight after menopause, [27] increase in number of estrogen hormones after menopause [28], age, drinking capacity and mutation of BRCA1 gene. One in every 500 of the general population in US. has

BRCA1 mutation and among family members with this mutation who are above 60 years of age, 54% is diagnosed with breast cancer and 30% is diagnosed with ovarian cancer. [29] BRCA1 gene is evidently thought to be correlated to breast cancer. There are many cases where abnormal BRCA1 gene is found in breast cancer patients. One of the functions of BRCA1 is repairing DNA damages. Therefore when malfunction occurs, it no longer operates at the same level as before due to outside stimuli and as a result tumor develops. The growth rate of a cell should accelerate if the function of BRCA1 is removed from the cell but instead it slows down due to an excess occurrence of RAD51. [30] The excess occurrence of RAD1 resulted in the acceleration of DNA mutation. BRCA1 gene is in a mutual relationship with its surrounding genes and the analysis on its effect is important. [31] BRCA1, a breast cancer tumor suppressor, plays an important role in relation to a cell cycle in the Cell signaling pathway. A tumor is developed through a variation of tumor suppressors, which is encoded with protein that suppresses the division of normal cells. The function of suppressing a cell division in a normal way stops operating and a tumor develops when there is a mutation of a tumor suppressor gene. Cancer is not developed through an independent variation of a tumor suppressor gene but through a variation in cells which occurs in many stages. The process of cancer development involves the variation of surrounding genes occurring through at least 7 stages. [32] Estrogen hormones and the growth factor deliver a signal to the surface of a breast cancer cell and stimulate the growth of a tumor. The tumor grows if a mutation occurs in the BRCA1 gene and the growth factor and signaling activity are not blocked. The signaling pathway passes through the surface of a breast cancer cell and if this pathway is blocked, the rate of breast cancer development can be slowed down. [33] This study aims to analyze the influence BRCA1, a tumor suppressor of breast

cancer in cell signaling pathway, has on its surrounding and related genes. The correlation between BRCA1 and genes that have the highest possibility of being affected by BRCA1 was analyzed. The data used for this analysis was gene expression data that had measured breast cancer cells in the human smooth muscle. An experimental design is divided into anatomical comparison design and time series design. The time series design exposed human smooth muscle cells in a serum for 24 hours and measured the gene expression data. This study plans to measure the correlation between BRCA1 and its surrounding genes when there are environmental changes and changes in the BRCA1 genefor breast cancer cell. In this circumstance, if the initial variation of the BRCA1 gene and surrounding genes are affected by BRCA1 and are in a non-linearity relationship, the correlation between the two genes can be analyzed through the lyapunov exponent. The lyapunov exponent shows a completely different result later as its difference of the initial condition exponential function widens. [34] Lyapunov exponent is an exponent that can measure the stability status of a system and is used as a basis of measuring the system's variation. Moreover, it can measure the non-linear relationship and useful information can be estimated if the relationship of genes with the possibility of having a value of another state from the beginning is measured by using initial changes. Lyapunov exponent of a physical concept and tumor suppressor of a medical concept have similar properties. [35] While an attractor draws out system changes, a tumor suppressor changes the cell cycle related genes when a mutation occurs. A local network, an aggregate for genes where a role as an attractor that rules its surrounding individuals and correlation exist, can be found by measuring the lyapunov exponent of a BRCA1 gene. A SPAL (Signaling Pathway Pattern Analysis using Lyapunov) program was created to analyze the correlation between BRCA1 genes and genes within the signaling pathway. This

program is programmed in C sharp language and it can calculate the lyapunov exponent of genes by using gene expression data and KGML data of KEGG database as an input value. [36]

II. Data and Methods

2.1 Case 1 Study Data and Methods

Since the metabolic profile measured through LC–MS/MS is a spectrum type data, its variables must be translated for data analysis Retention time interval changes variously to estimate the discriminant mode of thyroid gland disease using intensity values measured at the retention time of the metabolic profile. The variables are translated to estimate a proper statistic from the maximum, minimum, and mean values of the data measured simultaneously in each interval. The translated variables are used for screening a test to determine whether they are risk factors of outbreak of thyroid gland disease, and the optimal variables are selected for the TGDDM system. And also, the analysis dimension is reduced by principal component analysis extracting highly correlated variables as in existing papers [37]. Finally, we estimated the logistic model whose independent variables are PCA score and evaluate the model through the leave one out cross validation.

2.1.1 Thyroid Gland Disease Data

Steroid and PUFA standards were purchased from Sigma (St. Louis, MO, USA). Isoprostane standards were obtained from Cayman Chemical Company (Ann Arbor, MI, USA). β-Glucuronidase/arylsulfatase from Helix pomatia was purchased from Boehringer Mannheim (Boehringer, Germany). As derivative reagents, N-methyl-N-trimethylsilyl-trifluoroacetamide (MSTFA), trimethylsilylchloride (TMCS), N-trimethylsilylimidazole (TMSI), pentafluorophenyldimethylsilyl (flophemesyl) chloride, and N,O-bis(trimethylsilyl)trifluoroacetamide containing 1% TMCS (BSTFA+ 1% TMCS) were obtained from Sigma. Oasis HLBTM cartridge was purchased from Waters (Milford, USA). All solvents were using a high-purified HPLC grade.

De-ionized water was distilled before use. We followed all of patients and normal samples at the Division of Endocrinology, Internal Medicine of Gyeongsang National University Hospital.We collected early morning urine samples from 10 normal controls, 14 hyperthyroidism patients, and 16 hypothyroidism patients. The mean age of female hyperand hypothyroidism patients were 44.33±17.37 (mean±S.D.) years and 41.24±15.60 years, respectively. Normal controls (mean age; 41.15±16.41 years) were selected from the general population whom had no evidence of thyroid disease and taken a physical examination for this study. Hyperthyroidism was diagnosed on the basis of clinical features, diffuse goiter and positive antithyroid antibodies. The diagnosis of primary hypothyroidism was based on elevated serum TSH concentrations (basal, >4.1 mIU/l). We also found low serum concentrations of free thyroxine (free T4), total T4 (TT4) and total triiodothyronine (TT3). Using saturation analysis, we measured free T4 (reference range, 12.0 – 28.0 pmol/l, Clinical Assays, Cambridge and NH). We measured the TT4 (reference range, 60.2 – 160.0 nmol/l) and TT3 (reference range, 1.2 – 3.1 nmol/l) by RIA as previously described (Clinical Assays).We measured the T3 uptake using a routine method and we calculated the fT4I by multiplying the T4 concentration by the T3 uptake result. The collected urine samples were stored at −20 ○C until they were analyzed.We measured the creatinine concentration in the urine using the Jaff´e method. Metabolic profile is a quantitative data acquired in test for the materials utilizing in metabolite urine [8]. The measured metabolite profile is consist of well-known hormones such as androsterone, dehydroepiandrosterone, DHEA, 2-hydroxyestradiol, 2-hydroxyestrone, 2-methoxyestrone and other unknown hormones. Let the metabolic profile measured at time j on ith observation value X_{ij}. Where the intensity of measured material denotes area and j denotes retention time. Since the area value at retention time, which explains the characteristics of thyroid gland disease group and normal group is independent variable that is the most important in constructing the TGDDM system and it should be measured at the same reference

time for all observations. The measured metabolic profile X_{ij} however, was not measured at the same reference time in each observation. Thus the same reference time is generated and given to each measured data X_{ij} for all observations. First, the measuring time point is changed to measureing time interval. Therefore, every observation value, j has area value at the measureing inverval $t = [t-\Delta,\ t+\Delta]$. Since there are several measuring time points in the measuring interval t, the data is translated by the following three statistic equations:

$$T_{it}^{\max} = \max[X_{il}, X_{il+1}, ..., X_{im}] \quad \cdots\cdots (1)$$

$$T_{it}^{\min} = \min[X_{il}, X_{il+1}, ..., X_{im}] \quad \cdots\cdots (2)$$

$$T^{mean} = mean[X_{il}, X_{il+1}, ..., X_{im}] \quad \cdots\cdots (3)$$

where, $t-\Delta \leq X_{il,}\ X_{il+1}\ ,\ \cdots\ , X_{im} \leq t+\Delta$.

The above calculated statistics denote the analytic variables T_{itg}^{*} in specific interval of observation j. Where g denotes the binary value representing the thyroid gland disease group and the normal group. Therefore, incrementing the retention time interval Δ as 0.25, 0.5, 1.0, 1.5, 2.0, and so on, the interval value that is most suitable explains thyroid gland disease and the statistic should be estimated at the interval. The screening test determines whether it is the risk factor to find out the optimal interval and its statistic.

2.1.2 Thyroid Gland Disease Risk Factor Analysis

The t-test is executed to see whether the translated independent variable T_{itg}^{*} is risk factor for outbreak of the thyroid gland disease. The t-test is a verifying method

to see whether there is a difference between the two groups. [38] If $\overline{T_0}$, S_0^2 denote mean and variance of normal group of the thyroid gland disease, and $\overline{T_1}$, S_1^2 denote the treatment group respectively, we can verify whether the specific interval in chromatography is risk factor of the outbreak of thyroid gland disease.

$$t = \frac{\overline{T_0} - \overline{T_1}}{\sqrt{\frac{S_0^2}{n_0} + \frac{S_1^2}{n_1}}} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (4)$$

the above t-test, we can estimate optimal interval value to differentiating the thyroid gland disease group and the normal group. And also, we can see among which statistic measure area values at the retention time interval is better to differentiate between the two groups.

2.1.3 Principal Component Analysis

Principal component analysis(PCA) is appropriate when you have obtained a measurement on a number of observed variables and wish to develop a smaller number of hypothetical variable (called principal components) that will account for most of the variance in the observed variables. The principal components may then be used as predictor or criterion variables in subsequent analyses. The hypothetical variable is also called as a latent variable, a theoretical variable and a construct variable. In this research, investigation of the Pearson Correlation on the observed variables, which is constructed based on the retention time interval shows that lots of variables seem to be highly correlated to each T_{ij}^* other. Thus we can calculate the principal component score through principal component analysis on the observed

variable , which is generated based on retention time interval. In some previous researches [13,14], they made efforts to clarify the type of materials by estimating the retention time interval of high correlation using PCA.

2.1.4 Discriminant Model Analysis about Thyroid Gland Disease

Discriminant model is a model to make an estimation and a prediction about the causal relation based on the independent variables, in case the response variable is a categorical type. This method, however, can not use the logistic regression in case the response variable is a binary type as the existence of disease. If y denote variable to represent thyroid gland disease or not (1:disease, 0:normal), and denote T^*_{itg} is statistic at the translated retention time interval, the logistic regression model can be expressed as equation (5). Estimation of parameter coefficient β_i can complete the model.

$$\log \frac{p(y=1|T^*_{itg})}{1-p(y=1|T^*_{itg})} = \alpha + \beta_i \times T^*_{itg} \quad \cdots\cdots(5)$$

Since the logistic regression is linear and has parametric characteristic, the performance of prediction on the new data maintains to some extent. On the contrary, the data mining method about the non-linear and non-parametric models has been studied profundly. The representative examples are neural network, decision tree, support vector machine and so forth. [39] This model has high prediction performance, but it has shortcoming to be over-fitted to training set of model estimation. Comparative evaluation should be performed to get an optimal model and it should be verified that the selected model is superior to other ones. In this paper, we estimate a model optimal onto thyroid gland disease prediction system using above various

models. As a result, validation sets are estimated 100 % for all models. Therefore, the thyroid gland disease dicriminant model is estimated using logistic regression model that is easily analyzable and applicable.

2.1.5 Thyroid Gland Disease Model Validation

The first criterion of model validation is how few independent variables are used to construct effective model. In this paper, we utilize screening analysis to select optimized variables by t-testing the explanatory variables and reduce the variables by PCA. The second criterion of model validation is how a stable result is produced when the model is applied to a new data. In other words, it means the possibility of generalization of the model. However excellent the prediction performance of the model is, it is useless if it can not be generalized. In this paper, since the number of data was not enough to be partitioned into the training set and the validation set, we verified using leave one out cross validation. [40] Leave one out cross validation is a method to utilize one of n data set as validation set and the other n-1 data sets as training set, so we can validate all the n data sets through n models. It can be used to analyze few data sets as in this case.

2.2. Case 2 Study Data and Methods

2.2.1 Thyroid Cancer Data Description

In Korea Institute of Science and Technology Bioanalysis and Biotransformation Research Center [41], we measured profiles of urine samples of 23 thyroid cancer patients and 20 normal people with Gas Chromatography-Mass Spectrometry-Selected Ion-Monitoring(GC-MS-SIM) system. Urine samples of people who were diagnosed with thyroid cancer were obtained from pre- and post-operative patients in luteal phase.

The GC column used for estrogen analysis, a fused silica capillary coated with crosslinked 5% phenylmethyl siloxane was used [42]. Among the patients of thyroid cancer, 18 were women and 5 were men, and their age distribution was from 26 to 60. They had thyroid cancer (14 people), thyroid tumor (6 people), Thyroid mass (1 person), Thyroid papillary cancer (1 person) and Goiter (1 person), probably a kind of thyroid cancer.

Table 1. Information of the 23 thyroid cancer patients.

No.	Sex	Age	Type
1	F	57	thyroid cancer
2	F	52	thyroid cancer
3	F	54	thyroid cancer
4	M	60	thyroid cancer
5	F	35	thyroid tumor
6	F	38	thyroid cancer
7	F	36	thyroid cancer
8	M	58	thyroid cancer
9	F	34	thyroid tumor
10	F	59	thyroid cancer
11	F	41	urine(pre)/thyroid tumor(post)
12	F	42	thyroid papillary cancer
13	F	28	thyroid cancer
14	F	36	thyroid tumor
15	F	53	thyroid cancer
16	M	32	thyroid mass
17	M	55	goiter
18	F	47	thyroid cancer(pre)/thyroid tumor(post)
19	F	36	thyroid cancer(pre)/thyroid tumor(post)
20	F	46	thyroid tumor(pre)/thyroid mass(post)
21	M	55	thyroid tumor
22	F	26	thyroid cancer(pre)thyroid mass(post)
23	F	36	thyroid cancer

We measured the densities of target hormones of the patient group and the normal group that exist in androgen and estrogen metabolic pathway through GC-MS System. We also measured the densities of 18 hormones through filtering densities of the measured hormones.

2.2.2 Thyroid Cancer Screening Test Analysis

We analyzed the measured hormones profiles through a t-test to find out whether they were the risk factors of the thyroid cancer. The t-test is a method to check difference between the two groups to examine whether metabolites in androgen and estrogen metabolic pathway are risk factors [43]. If we define the average and standard deviation of normal group as $\overline{X_0}$ and S_0^2, the average and standard deviation of cancer group as $\overline{X_1}$ and S_1^2, we can define t statistic by equation (6).

$$t = \frac{\overline{T_0} - \overline{T_1}}{\sqrt{\frac{S_0^2}{n_0} + \frac{S_1^2}{n_1}}} \quad \cdots\cdots(6)$$

Here, n_0 is sample size from the normal group and n_1 is from the cancer group. Through the t-test, we were able to perform a preliminary analysis on whether metabolites exist in androgen and estrogen metabolic pathway is the risk factor influencing the thyroid cancer. We can estimate a discriminant model of thyroid cancer through the risk factor metabolite and it can be independent variable in constructing the thyroid cancer decision-making system.

2.2.3 Thyroid Cancer Decision Making System

Through the metabolite in androgen and estrogen metabolic pathway, the study was able to develop a decision making system that could determine chances of the thyroid cancer. The models that were used for the decision making system are logistic regression, decision tree and neural network. We estimated a model that could distinguish the outbreak of thyroid cancer through a logistic regression model. The

logistic regression model is used in distinguishing two groups through independent variable when the target variable is binary type. When independent variables are $x_1, x_2, \ldots, x_p$, we can formulate equation (7) for target variable y.[44]

$$\log \frac{p(y=1 \mid x_1, x_2, \ldots, x_p)}{1-p(y=1 \mid x_1, x_2, \ldots, x_p)} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \quad \text{..... (7)}$$

The logistic regression model changes odd ratio value to log scale when y is 1 and target variable y is binary type [0, 1]. Target variable represents whether there is thyroid cancer [0: normal, 1: thyroid cancer]. We can also estimate the logistic regression model that measured metabolite hormones in androgen and estrogen metabolic pathway as independent variables. We recognized the influence of metabolites related with thyroid cancer through the estimated logistic regression model. As decision tree categorizes and expresses a probability of the thyroid cancer in a tree structure, it uses CHAID, CART, C4.5 Algorithm.[45,46] It separates distribution of hormones through chi-square statistic, gini index, entropy index in the thyroid cancer group and the normal group at maximum. The neural network starts out from the neurophysiology structure ad forms input layer, hidden layer and output layer. If we presume the variation that exist on the input layer as $x_1, x_2, \ldots, x_p$, the node H_j that consists the hidden layer and the target variable y that forms the output layer could be equal to equation (8, 9).

$$H_j = f_j(b_j + w_{1j} X_1 + w_{2j} X_2 + \ldots + w_{pj} X_p) \quad \text{.....(8)}$$

$$y = g(b_0 + w_{10} H_1 + w_{20} H_2 + \cdots + w_{j0} H_j) \quad \text{.....(9)}$$

Where f_j, g is activation function and w_{ij} is the weight for each node. We made the TCDMS through the three modes because the models have both the positive and negative points. Since the logistic regression basis on linearity, it could maintain identical accuracy even on new data but its accuracy is low in quality. While the decision tree and neural network seem to have high accuracy for their non-linearity and non-parametric characters, their accuracies are likely to decrease if there is new data. Therefore, all three models were used to make the TCDMS.

2.2.4 Posterior Probability

Upon making the TCDMS through the three models, the posterior probability will estimate probability of each observation subjected to the thyroid cancer from the system the possibility of thyroid cancer is higher if the posterior value of probability is high since the value is an index that demonstrates possibility of thyroid cancel for each observation. Hence, the posterior possibility value on each observation will display the rate of progress for the thyroid cancer thus the posterior value on the logistic regression model, decision tree and the neural network model is obtainable. The logistic regression model is linear and parametric thus we should find the estimate parameter coefficient from the equation (7). In this research, we were able to re-estimate the posterior probability as equation (10) on each subject through the model that is estimated with parameters. [47-49]

$$p(y=1\,|\,x_1, x_2, \ldots, x_p) = \frac{\exp(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p)}{1-\exp(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p)} \quad \text{..... (10)}$$

Calculated posterior probabilities of individuals are used to sort these individuals.

We can calculate the values of posterior probabilities in case of cancer $[y_i = 1]$ through estimated values of parameter coefficients. We can expect the position of spectrum that each individual goes to thyroid cancer if we sort the values of observed posterior probabilities. Although information that is estimated by posterior probability does not have correct time and intermission, we could estimate ordinary information about the degree of thyroid cancer. The study can estimate the posterior probability on observations when a model is estimated on the neural network. When M_{NN} is given as the neural network model on M_{tree} that is estimated through the decision tree model posterior probability value of each observation could be estimated as $P(y_i = 1 \mid M_{NN}),\ P(y_i = 1 \mid M_{tree})$.

2.2.5 Thyroid Cancer Model Validation

We evaluated the model for reliability and validity of the logistic regression, decision tree and neural network. The number of data to be used for Training Set is small which is to estimate a model and Validation Set for evaluating model, since the data was made with the 23 thyroid cancer patients and 20 normal people. Thus, we used the method of leave-one-out cross validation [50,51]. The leave-one-out cross validation method uses n-1 data out of n data for training set to estimate model and 1 data for validation set to check whether data has been legitimately classified. In case of using n data of validation set by once, n models are estimated and a general model that mix them. But we wanted to allow reliability and validity of the model through the logistic regression model, decision tree and neural network. And leave-one-out cross validation about n data is estimated using all data because we wanted to evaluate n validation data in this analysis. Therefore, posterior probability value estimated through the logistic regression, decision tree and neural network that has

reliability and validity can express the orders in continuous spectrum lines of the normal group and the disease group.

.

2.3. Case 3 Study Data and Methods

2.3.1 Metabolic Pathway Data

A metabolic pathway is composed in a network structure and a node which is a basic component of the network, contains chemicals and genes such as a compound and an enzyme. Metabolic pathways have been formatted in xml in 148 counts based on the KEGG database [52]. For this analysis, information on metabolic pathway structure was put into the genes that need location on network structure. Data being used for this analysis is gene expression data which was measured from the normal group and the colon cancer group. The genes are included in the enzymes within the metabolic pathway. Thus, a program which extracts only specific information from the KEGG data, made out of xml, was made. The KEGG DB includes information about compound and enzyme that interact with information of location, organism name and pathway number of a specific gene. As this analysis is to estimate the effect of a metabolic pathway in the broad perspective, only organism information o and pathway information p have been used.

2.3.2 Colon Cancer Gene Expression Data

Gene expression data is measured through two types of experimental design with Caco-2, Intestinal Epithelial Cells as targets. First measurement of gene expression value was carried out in a specific time point on 10 people from a normal group and 14 people from a cancer group. Second measurement of gene expression value was carried out in 11 time points by cultivating Caco-2 Cell for 26 days [25]. The data

used for this analysis was obtained through SMD (Stanford Microarray Database) [53]. A pre-analysis of gene expression data of normalization [54] and filtering used the tool provided by SMD. After analyzing through SAM, 15366 significantly expressed genes on colon cancer were estimated when FDR (False Discovery Rate) is 10.59%. This means that 57.54% of the total of 26706 genes is significantly expressed in relation to colon cancer. Moreover, it is estimated that 5626 genes are significant in relation to colon cancer when FDR is 10.33% of 7850 filtered genes. 71.67% of filtered genes are estimated to be significant genes. Similarly, too many number of genes show up when a colon cancer related gene is analyzed through gene expression data. Therefore, this study aims to search for correlation between colon cancer and metabolism through gene variation within a group called a metabolic pathway. The measured genes will be given new information through the metabolic pathway. If a slide is defined as j, a normal group or a cancer group as g and time point as t, gene i is defined as x_{ijgt}. The expression value x_{ijgt} will get metabolic pathway number p and organism name o as additional information through the metabolic pathway. As the subject of this analysis is homo sapience, information o concerning organism is omitted and instead pathway information p is added. Therefore, **expression value** of gene i is denoted as x_{ijgtp}.

2.3.3 Metabolism with Correlation to Colon Cancer through ANOVA Model

If the group is g, time is t, metabolic pathway is p and function is additive, gene expression value x_{ijgtp} can be defined as equation (11) below.

$$x_{ijgt} = \mu + \tau_{tp} + \gamma_{gp} + \rho_{tg} + (\tau\gamma)_p + (\tau\rho)_t + (\gamma\rho)_g + (\tau\gamma\rho)_{gpt} + \epsilon_{ijgpt} \quad \text{.....(11)}$$

Here, μ is the total average, τ_{tp} group effect, γ_{gp} time effect, ρ_{tg} pathway effect, $(\tau\gamma)_p$, $(\tau\rho)_t$, $(\gamma\rho)_g$, $(\tau\gamma\rho)_{gpt}$ interaction effect and ϵ_{ijgpt} a random error. A study hypothesis of equations (12-14) can be created through equation (11) as below.

$$H_0 : \tau_o = \tau_1 \quad \text{.....(12)}$$

$$H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_T \quad \text{.....(13)}$$

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_p \quad \text{.....(14)}$$

Equation (12) is a study hypothesis that shows whether or not gene expression data has a difference amongst groups, equation (13) a difference between time points and equation (14) a difference between metabolic pathways. The ANOVA model [55] can be used to solve a problem concerning the study hypothesis. The ANOVA model is a method that finds out between subject variances and within subject variances to analyze whether or not the effect of a subject subsists using the size of these ratios. To estimate whether a difference of effect amongst metabolic pathways exists as in the study hypothesis equation (14), the sum of square and mean square of between metabolic pathway and within metabolic pathway can be estimated like in table 2.

Table 2. ANOVA table for metabolic pathway effect test.

Source of Variation	Sum of Square	DF	Mean Square	F
Between metabolic pathway	$SS_P = n\sum_{p=1}^{p}(\overline{y_{i.}} - \overline{y_{..}})^2$	P-1	$MS_p = \frac{SS_p}{P-1}$	
Error(Within metabolic pathway)	$SS_E = SS_T - SS_P$	N-P	$MS_E = \frac{SS_E}{M-P}$	$F = \frac{MS_P}{MS_E}$
Total	$SS_T = \sum_{i=1}^{n}\sum_{p=1}^{p}(\overline{y_{ip}} - \overline{y_{..}})^2$	N-1		

Test statistic F-value can be created and estimated by finding MS_p of between metabolic pathway and MS_E of within metabolic pathway. As the F-value gets higher under the degree of freedom, the difference among effects of metabolic pathways gets higher. If an analysis is carried out using ANOVA, it is expected that the study hypothesis equation (12-14) will be rejected. When it is rejected, an additional analysis is required to find out how each effect is different between the various processes. This kind of analysis is called multiple comparison. When the paired test to find out the difference of $\rho_1, \rho_2, \cdots, \rho_p$ in equation (11) is put into practice, it demonstrates a method on how to control the significant level of each paired test. When the number of statistical test frequency is *m*, $\alpha' = \frac{\alpha}{m}$ is appointed as the significant level and the difference can be estimated. There are methods such as Student-Newman-Keuls(SNK) Method, Bonferroni method, Scheffe method and LSD method in the multiple comparison method. [56] This study based its analysis on the variation of genes within each metabolic pathway. The metabolic pathway is composed of many genes. When analysis unit uses genes, a significant gene, where up-regulation and down-regulation is possible, is found. On the other hand, even if there is a metabolic pathway that is composed only of significant genes for colon cancer, the effect of a metabolic pathway cannot be obtained if these genes display up-regulation and down-regulation simultaneously. Therefore, this study converted the **gene expression value** to absolute value and used it on the ANOVA model.

2.3.4 Sum of Square Metabolic Pathway

This study also aims to interpret the gene expression value from a different point of view using metabolic pathway. The correlation between a specific metabolic

pathway and colon cancer through the average expression volume of genes has been explained so far and from this point onwards, it will be explained through variation of genes. If a metabolic pathway has correlation with colon cancer, genes in the metabolic pathway intend to be expressed in a similar level. However, if there is no correlation with colon cancer, genes in the metabolic pathway intend to be expressed in random. There is a high probability of it being dispersed widely and randomly. Therefore, metabolic mean square (MSE) in table 2 will be compared. equation (15)

$$MS_E = \frac{\sum_{i=1}^{n}\sum_{p=1}^{P}(\overline{y}_{ip} - \overline{y_{i..}})^2}{N-P} \quad \text{.....(15)}$$

MSE in equation (5) is dissolved by p according to the metabolic pathway and $p^{(th)}$ MSE becomes as in equation (16).

$$MS_{EP} = \frac{\sum_{p=1}^{P}(\overline{y}_{ip} - \overline{y_{i..}})^2}{n_p} \quad \text{.....(16)}$$

Where n_p is the number of genes in $p^{(th)}$ pathway.

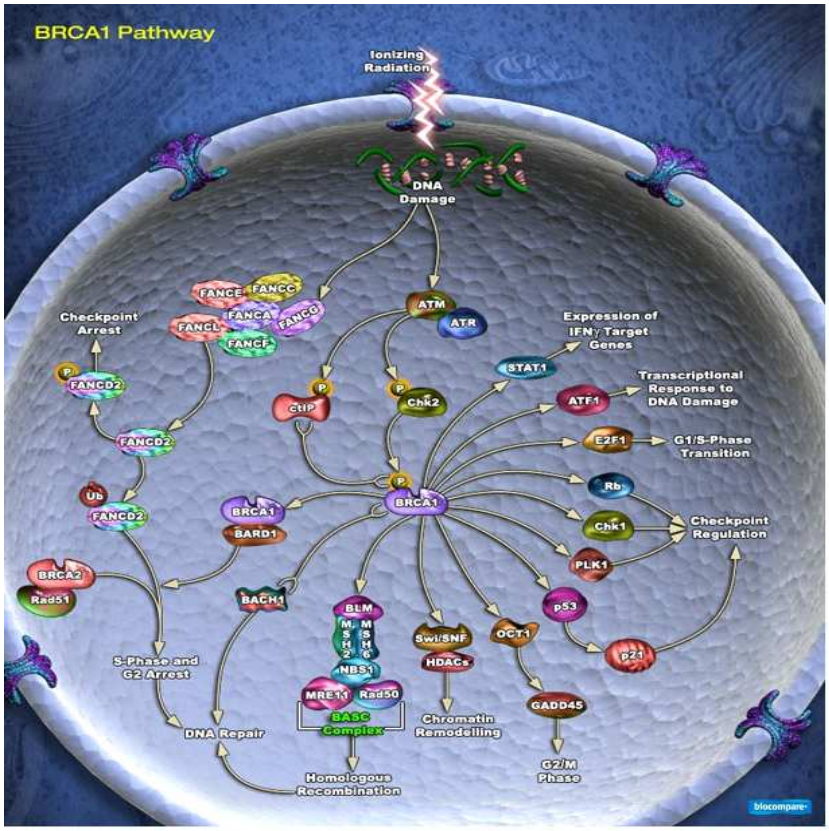
Through a normalized process, the entire data follows $N(0,1)$ which stands for standard normal distribution. Based on variance 1, values can go up or down depending on the metabolic pathway. If the MSE value is higher than 1, the genes in the metabolic pathway are distributed randomly and if the MSE value is near to 0, the genes in the metabolic pathway are distributed consistently. Consequently, it is possible to estimate that there is a correlation between the metabolic pathway and colon cancer.

2.4. Case 4. Study Data and Methods

2.4.1 Tumor Suppressor BRCA1's function and location

BRCA1 functions to suppress the tumor by regulating the process of repairing damaged DNA and to maintain genomic stability. However, its original function stops operating when a gene is mutated and as a result cancer development is stimulated. From the lecture study, BRCA1 plays an important role in cells such as DNA damage and repair, embryonic proliferation, transcription, centrosome duplication and apoptosis. [57] The study on BRCA1 gene was usually about transcriptional regulation and DNA repair. BRCA1 activates the enzyme after merging with a damaged DNA molecule and accelerates the processes that are needed for DNA repair. It plays the role of a recognition factor for DNA repair. [58] Although genes that demonstrated interaction with BRCA1 have been closely examined through many studies, not enough studies have been carried out to find out about their relationship. BRCA1 plays an important role in the pathway regarding DNA damage checkpoint. ATM and ATR activate chk1 and chk2 when DNA damage occurs. Chk1 and chk2 affects P53 and BRCA1 by being involved in the Cell Cycle control. BRCA1 activates its surrounding genes for DNA repair and prevents the cell from developing into a cancer. There is hyperphosphorylation of BRCA1 and an increase in its expression in the late G1 cell cycle and during S-phase and there is dephosphorylation after M-phase. The phosphorylation of BRCA1 is carried out by ATM. [59,60] BRCA1 is known to suspend the process of a cell cycle by merging with the hyperphosphorylated pRB.

Figure 2. Breast Cancer Tumor Suppressor BRCA1 Pathway.



Available at www.ambion.com

2.4.2 Serum stimulation Vascular Smooth Muscle cells

The required data for measuring the effect of the tumor suppressor BRCA1 on its surrounding genes were collected. We also looked for data which contained information on the response of genes according to time. In regards to time at this point, data which show specific response in a short period well is more appropriate than data which measured time for a long period like Cohort study. Each sample was measured at 0, 1,3,6,12,24 h and there are 13951 existing genes. The measured gene i is

defined as Z_{ij}. where, i stands for gene and j stands for time. Gene i is able to gain information on the signaling pathway. KEGG Database is providing compounds related to cell cycle and pathway information on genes in the form of XML. The information includes location, class and interaction regarding the signal pathway of genes. Gene expression data obtains information on Pathway and through this the distance of genes in Signaling Pathway can be seen.

2.4.3 Fractal Dimension and Lyapunov Exponent

Mandelbrot, a French mathematician, reported in his thesis [61] titled 'How long is the coastline surrounding England?' that the length of a coastline changes depending on the accuracy of the measurement unit level. A form of frequency pattern is generally shown in the case of either gene expression measured by time series or metabolism data. A characteristic of a frequency pattern is that there can be a distorted result of the data depending on the point of time of the measured time. Consequently, a problem of fractal dimension arises and this must be solved as the measurement value can change depending on the length of interval at the point in measurement and which point it is measured from. The fractal dimension is increased by its control based on the self-similarity and recursiveness. [62] On the other hand, linear moving average method was used to reduce fractal dimension. Therefore the current view was defined as j and prediction equation (17) was used to predict the time series value $Z_{i(j+l)}$ of future view $j+l$.

$$\hat{Z}_{i(j+l)}(n) = \widehat{Z_{ij}} + b_j l \quad \text{.....(17)}$$

1 stands for lead time which is used in the prediction and

$b_j = \frac{2}{N-1}(M_j - M_j^{'})$ is trend variation from the point of j. M_j represents the moving average while $M_j^{'}$ is the double moving average.[63] Gene expression value of the fractal dimension which was reduced by linear moving average method can be seen. Also, the fractal dimension can be estimated as in equation (18).

$$D = \lim(\epsilon \to 0)\frac{\ln N(0)}{\ln(1/\epsilon)} \quad \text{.....(18)}$$

D is defined as the dimension the time series data have and when ϵ is defined as the length of the structural form, the connection of measured value of an experiment from the present estimated time, the number required to cover the whole the structural form is defined as N.[64] If gene expression time series data according to fractal dimension is created by using the linear moving average method, the relationship between surrounding genes in the signaling pathway is found. In order to simulate the degree of the influence BRCA1, a tumor suppressor of breast cancer, has on its surrounding genes and the process of expression in the signaling pathway, an appropriate exponent is required. Many use pearson correlation coefficient to measure the interrelationship between BRCA1 and its surrounding genes. However, if these genes which show a frequency pattern are compared through pearson correlation coefficient, the possibility of drawing a distorted conclusion is high. Table 3 is the simulation function which has a 1:1 mapping. Apart from 1-polynomial function, the simulation function has a value of low pearson correlation coefficient despite the fact that there is no error.

Table 3. Pearson Correlation regarding Simulation Function.

Simulation function	Pearson correlation
1- polynomial function	1.0000
2- polynomial function	0.5630
3- polynomial function	0.9161
4- polynomial function	0.1287
Fourier transform function	0.1299
Auto correlation function	0.0184

As a result, an exponent that can express the characteristics of a frequency pattern is required. Lyapunov exponent can be used as an exponent that is able to express the correlation between two genes through their initial conditions. Lyapunov exponent is an exponent that is well-known through Lorenz's butterfly effect. A butterfly effect is the possibility of a storm breaking out from a long distance under the influence of the movement of a butterfly. Lyapunov exponent also shows a chaos pattern that is already known. [65]

$$\epsilon(0) = x_1(0) - x_2(0) \quad \text{.....(19)}$$

Based on equation (19), Lyapunov exponent λ can be calculated as in equation (20).

$$\epsilon(t) = t_1(t) - x2(t) \propto \epsilon(0)\exp(\lambda t) \quad \text{.....v.....(20)}$$

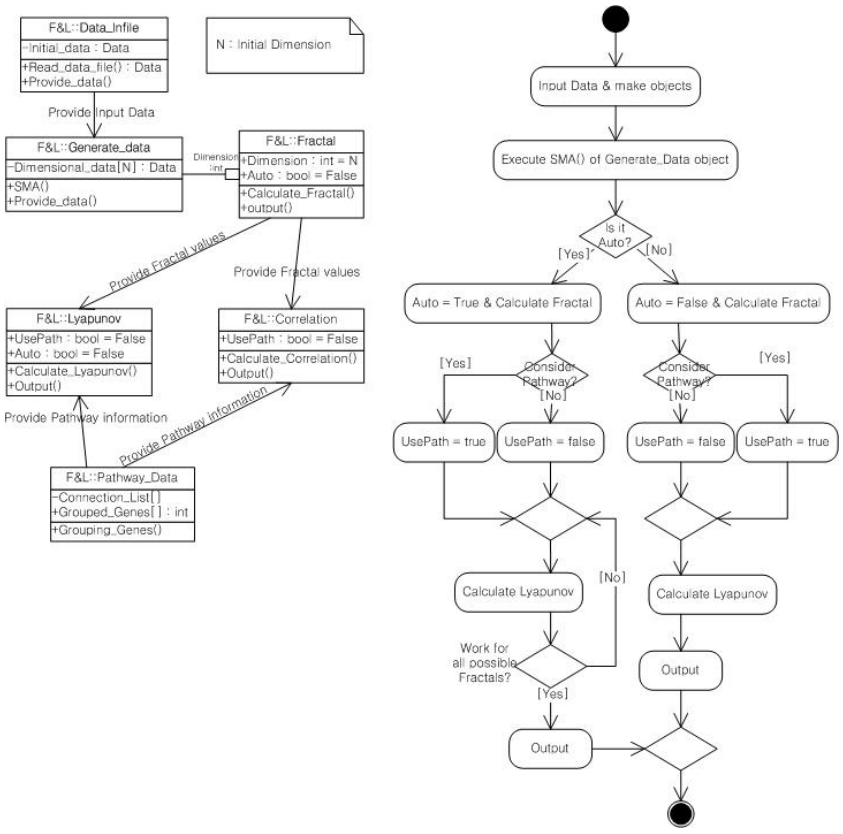
If the value of λ is positive, as the difference of the initial condition eventually increases as time passes, it can be said that the dynamic systems shows chaotic pattern. [66] The lyapunov exponent λ shows the amount of influence the change of initial genes has on the value which change depending on time. Furthermore, it does

not depend on linearity like pearson correlation coefficient. Since λ is expressed depending on time, the order of influence BRCA1 has on its surrounding genes can be analyzed.

2.4.4 Signaling Pathway Pattern Analysis using Lyapunov (SPAL)

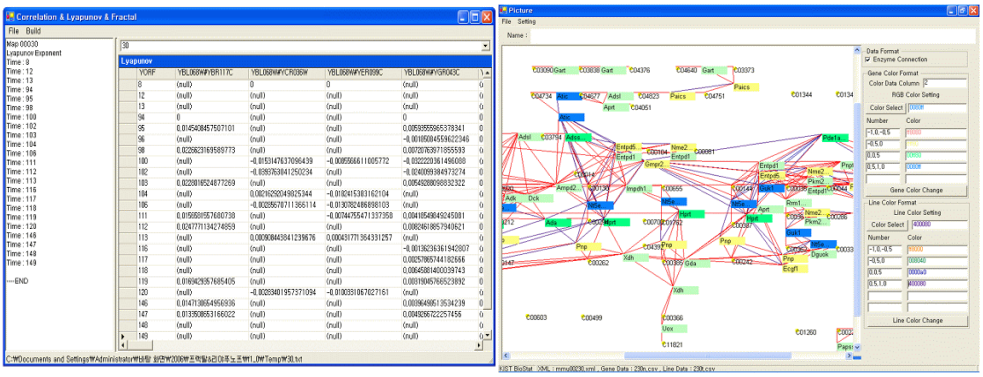
A program which measures lyapunov exponent λ from a manifold dimension and changes the fractal dimension was developed as in figure 3 through C Sharp program language for enzymes and compounds belonging to the signaling pathway.

Figure 3. Interaction Diagram about SPAL Program.



Fractal dimension, lyapunov exponent and pearson correlation coefficient are calculated by using gene expression data which includes signaling pathway information as an input. The prediction value on trend factor is estimated using the linear moving average method. This allows for prediction value on fractal dimension, lyapunov exponent and pearson correlation coefficient to be calculated. The information on signaling pathway which is used as an input uses KEGG pathway data and it is formatted as xml. [67] Gene expression data uses the format of SMD. Lyapunov exponent of the genes and their fractal dimension depending on time is as printed in figure 4. Also, the function of printing results in the pathway is included.

Figure 4. SPAL Program main screen and XML Visualiser.



III. Results

3.1 case 1 study : Results

3.1.1 Translation Variable and Risk Factor about Thyroid Gland Disease

The variables are generated with retention time's changed interval. Although several area values are included in the generated variables, only one statistic should be selected and used. Thus we need to decide the criterion to select the statistic and the interval size. The t-test is used to select the statistic and the interval size that explain the characteristics of the treatment group and the normal group. We could get the results of figure. 5, 6, 7 by changing the retention time interval from 0.5 to 3.0 variously and applying corresponding statistic to the t-test. In the figures, x-axis denotes the measured retention time, y-axis denotes translated interval, and z-axis denotes p-value. Therefore, the lower the values of z-axis, the better the variables explain characteristics of the treatment group and the normal group. In figure 5, the result using mean value at each interval as a statistic is shown, in figure 6, maximum value, and in figure 7, minimum value is used. As the findings of figures, there exist area values that can effectively discriminate between the treatment group and the normal group over 18 of retention time irrelevant to statistic. Also, overall distribution patterns show that the least p-value is at the interval 1.5.

Figure 5. Distribution of p-value with respect to mean value.

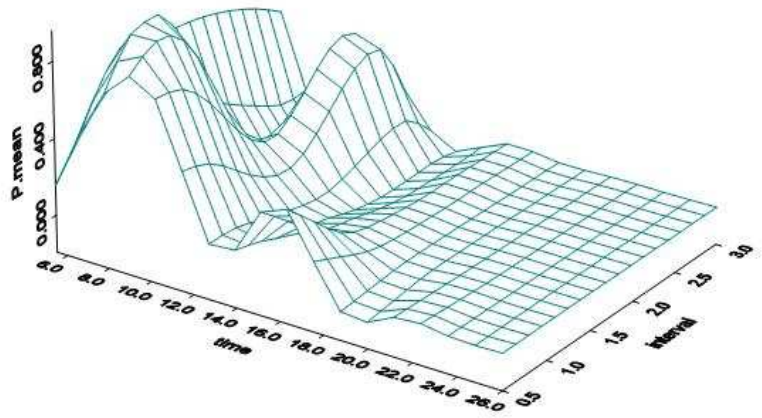


Figure 6. Distribution of p-value with respect to max. value.

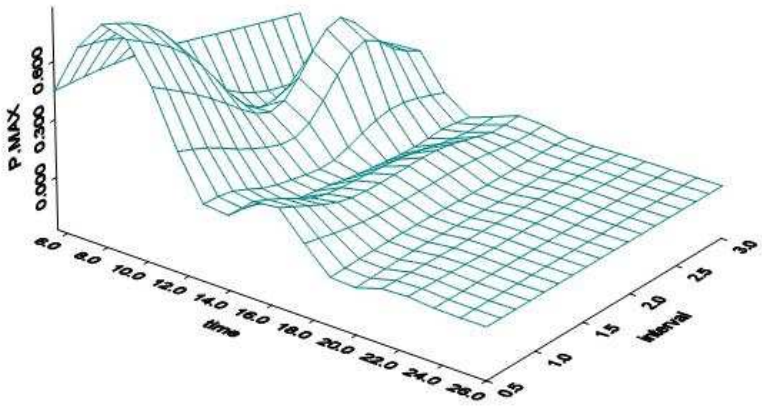
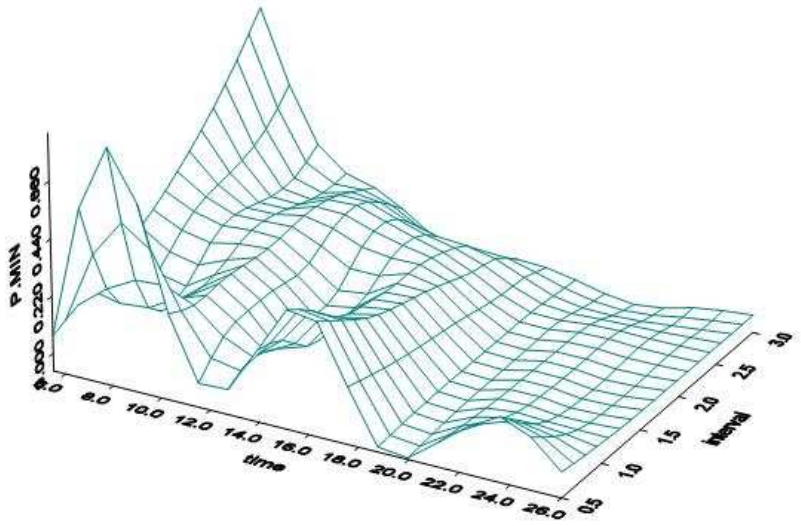


Figure 7. Distribution of p-value with respect to min. value.



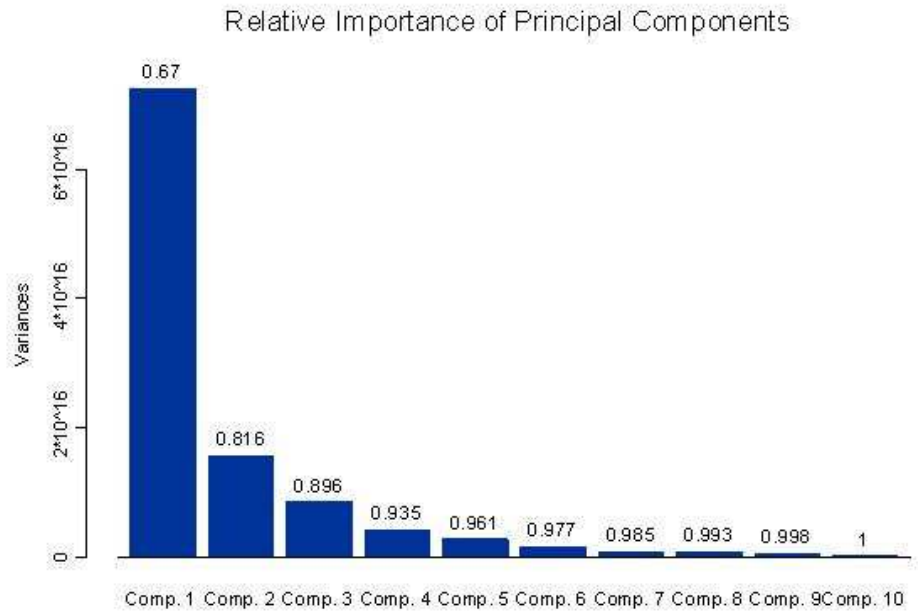
Therefore, the estimate of the optimal retention time interval becomes 1.5, and as a statistic, the maximum area value in that interval can be an important independent variable that explains–characteristic of the outbreak of the thyroid gland disease.

3.1.2 Principal Component Analysis Result

A good analysis uses small amount of variables to explain their maximum variance. Thus in order to reduce the number of highly correlated independent variables, principal component analysis is performed. The number of principal component is determined by choosing eigenvalues that is greater than 1, or determined with respect to the cumulative variance. Based on this consideration, the number of principal component is determined to be 3, and by this number of principal component, around

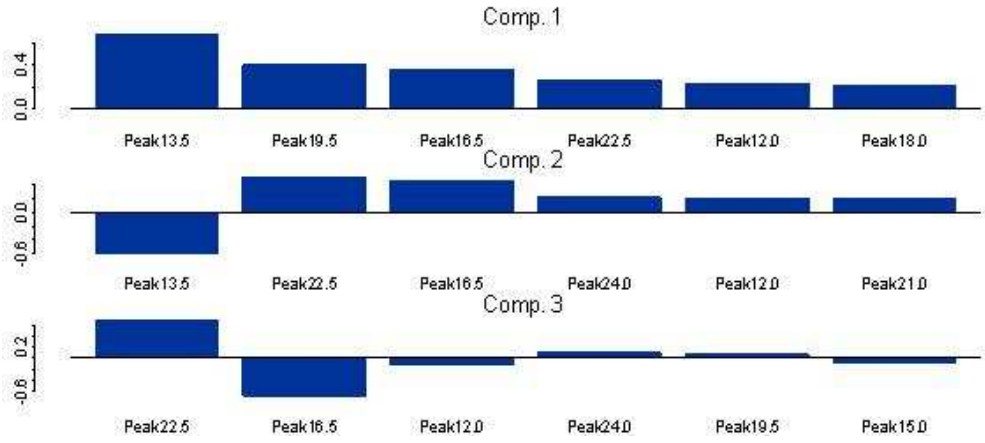
89.6 % of the data variance is explained as shown in figure 8.

Figure 8. Relative importance of Principal Components.



It is shown that the first principal component has a high correlation with the variables of Peak13.5, Peak19.5, and Peak16.5, which is in the retention interval between 12 and 13.5. [figure 9] The second principal component has negative correlation with Peak13.5 and has relatively higher correlation with Peak22.5 and Peak16.5. We can see that the third principal component has positive correlation with Peak22.5 and has negative correlation with 16.5. The 3 principal components are explaining 90 % of total variance of 13 explanatory variables.

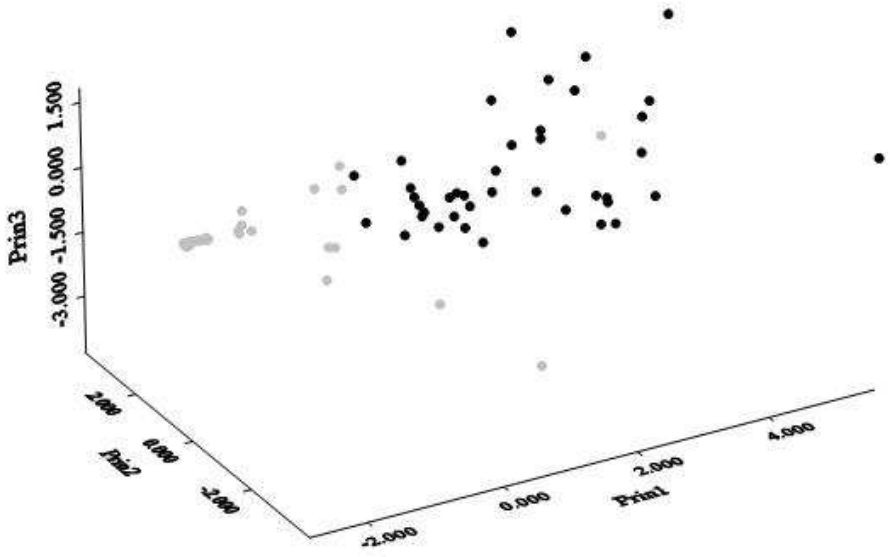
Figure 9. Principal component Scores.



Thus we showed how the 3 principal components are explaining the thyroid gland disease. To see the distribution of disease group and normal group about the reduced independent variables of 3 principal components, it is plotted in a 3-dimensional space as shown in the figure 10. First of all, the disease group and normal group are distinguishably distributed in a 3-dimensional space. In general, the normal group ($\overline{C1}_0$=1.19) is distributed more than the disease group ($\overline{C1}_1$=-1.37) depending on how large the first principal component is. For second principal component, on the contrary, the larger the value is, the more treatment group ($\overline{C2}_1$=0.30) is distributed than the normal group ($\overline{C2}_0$=-0.26). And for third principal component, the larger the value is, the more normal group ($\overline{C3}_0$=0.24) is distributed than the treatment group ($\overline{C3}_1$=-0.28). Thus, from the measured value area of the variables generated at retention time interval 1.5, we can conclude that there are many unknown metabolites in the metabolic profile**,** which take a similar effect on the thyroid gland disease and it can be estimated that these materials are significant to the outbreak of the thyroid

gland disease.

Figure 10. Distribution of the disease group and the normal group with the axes of 3 principal components.



3.1.3 Logistic Regression and Validation about Thyroid Gland Disease

We validated discriminant models of logistic regression, neural network, and decision tree, which discriminate between normal group and treatment group based on the principal component score of above analysis. Since all 3 models have the accuracy of 98.7 % in validation set of classification table, the logistic regression model among 3, which is easy to analyze, is used to construct thyroid gland disease discriminant system. From the analysis of logistic regression with input of 3 principal components

that are explaining 90 % of variance of explanatory variables, it showed that the first principal component had a parameter coefficient of -1.70 and took the largest effect on the model. Then the coefficient of the third principal component was -1.34 and the second principal component was 0.86.

Table 4. Result of Logistic regression.

Parameter	Normal	Treatment	Estimate	p-value
Intercept	.	.	-0.4866	0.3666
Prin1	1.19±2.17	-1.37±1.18	-1.7002	<.0001
Prin2	-0.26±1.68	0.30±1.91	0.8672	0.1613
Prin3	0.24±1.51	-0.28±1.36	-1.3104	0.0089

The model accuracy from leave-one-out cross validation revealed that the observation of one treatment group was miss-classified, and other observations was classified properly. Thus we developed the TGDDM system using logistic regression with input of measured 3 principal component

3.2. case 2 study : Result

3.2.1 Screening test about Thyroid Cancer

We conducted the t-test on 18 metabolites that exist in androgen and estrogen metabolic pathway with the 23 thyroid cancer patients and the 20 normal people as shown in table 5. We were able to find out that metabolites that are estimated as risk factors of thyroid cancer are 2 - hydroxyestrone, 2 - hydroxyestradiols, 2 - methoxyestrone, 2 - methoxyestradiols, 2 – methoxyestradiol 3 - methylether. They could be detected just before being exhausted to gland in metabolic map [figure 11]. These hormones show definite difference between the averages of thyroid cancer group and the normal group. Therefore, metabolites related with thyroid cancer are distributed

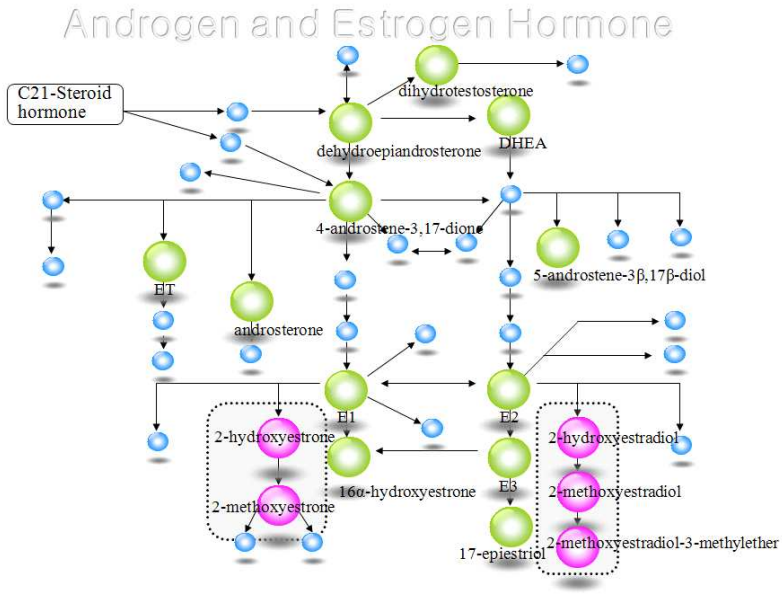
mainly after estrone (E3) and 17 β-estradiol (E2) in androgen and estrogen metabolic pathway. They could be estimated to have relation with outbreak of disease.

Table 5. Results of t-test on the 23 thyroid cancer patients and the 20 normal people.

Hormone Name	Normal(meas±SD)	Cancer(meas±SD)	t value	p value
androsterone	7.411 ± 3.344	7.290 ± 2.967	0.13	0.9010
Etiocholanolone(ET)	6.103 ± 2.764	6.625 ± 3.013	-0.59	0.5570
dehydroepiandrosterone	1.110 ± 0.969	1.155 ± 0.990	-0.15	0.8818
4-androstene-3,17-dione	0.756 ± 0.578	0.795 ± 0.520	-0.23	0.8181
dihydrotestosterone	0.766 ± 0.546	0.720 ± 0.487	0.29	0.7713
16α-hydroxy DHEA	2.947 ± 1.648	3.167 ± 2.191	-0.38	0.7086
5-androstene-3β,17β-diol	0.404 ± 0.210	0.508 ± 0.381	-1.13	0.2655
estrone (E1)	14.422 ± 7.332	16.239 ± 8.248	-0.76	0.4487
17β-estradiol (E2)	12.862 ± 5.626	14.458 ± 6.369	-0.87	0.3880
estriol (E3)	18.924 ± 11.220	19.034 ± 10.647	-0.03	0.9741
16α-hydroxyestrone	17.816 ± 47.628	17.203 ± 8.009	0.26	0.7984
17-epiestriol	11.205 ± 3.772	9.512 ± 5.860	1.14	0.2606
16-epiestriol	13.313 ± 5.336	12.020 ± 4.509	0.85	0.3997
2-hydroxyestrone	498.677 ± 360.941	120.665 ± 63.594	4.61	0.0001
2-hydroxyestradiol	40.854 ± 29.326	15.633 ± 6.593	3.76	0.0005
2-methoxyestrone	83.524 ± 59.973	36.401 ± 17.064	3.39	0.0015
2-methoxyestradiol	93.910 ± 59.094	19.707 ± 10.973	5.52	0.0001
2-methoxyestradiol-3-methylether	59.207 ± 29.248	25.424 ± 14.224	4.7	0.0001

As the result of the t-test to 18 metabolites in androgen and estrogen metabolic pathway, hormones such as 2-hydroxyestrone, 2-hydroxyestradiol, 2-methoxyestrone, 2-methoxyestradiol,2-methoxyestradiol and 3-methylether could be estimated as risk factors.[p-value≤0.05]

Figure 11. Risk factors distribution of thyroid cancer in androgen and estrogen metabolic pathway.



3.2.2 Thyroid Cancer Decision Making System

We estimated the logistic regression model that has independent variables such as 23 thyroid cancer patients, 20 normal people and 18 metabolites in androgen and estrogen metabolic pathway. Before examining, we cleaned up the data of 18 metabolites using the process of normalization and missing value imputation. Since the number of data is not enough for the validation, we calculated 43 logistic regression models. Through estimated parameters and mean value of the test statistics, we achieved thyroid cancer discriminant model as shown in table 6. Metabolites that cause much effects to logistic regression model were 2 - hydroxyestrone [β coefficient = - 26.5840],and 2 - methoxyestrone [β coefficient = - 24.1155], 2 - methoxyestradiol [β coefficient = - 18.4875], 2 - methoxyestradiol-3 - methylether [β coefficient = - 15.6975]. Parameters that influence much in thyroid cancer discriminant model could

be found mainly in hormones being emitted to gland in androgen and estrogen metabolic pathway.

Table 6. 43 average logistic regression parameters.

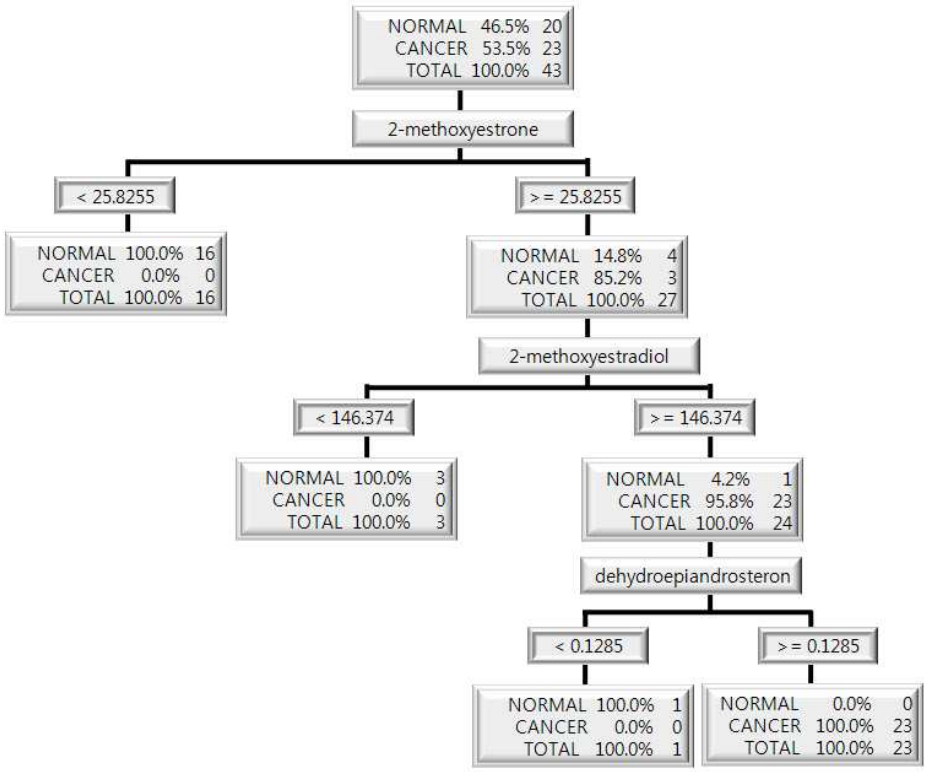
Hormone Name	Parameter Coeff.	Standard Error	Wald χ^{2v}	p value
Intercept	-4.9626	25.6818	0.04	0.8468
androsterone	5.7586	45.6279	0.02	0.8996
Etiocholanolone(ET)	7.4678	53.1024	0.02	0.8882
dehydroepiandrosterone	7.4316	44.4975	0.03	0.8674
4-androstene-3,17-dione	-1.4520	32.9935	0.00	0.9649
dihydrotestosterone	-1.6786	32.1477	0.00	0.9584
16α-hydroxy DHEA	-8.3316	51.0439	0.03	0.8703
5-androstene-3β,17β-diol	5.7400	29.3010	0.04	0.8447
estrone (E1)	1.2972	41.7609	0.00	0.9752
17β-estradiol (E2)	-1.0553	58.0360	0.00	0.9855
estriol (E3)	-1.6432	30.9132	0.00	0.9576
16α-hydroxyestrone	0.6458	39.1798	0.00	0.9868
17-epiestriol	-6.9175	46.4806	0.02	0.8817
16-epiestriol	8.5029	48.5531	0.03	0.8610
2-hydroxyestrone	-26.5840	107.1	0.06	0.8039
2-hydroxyestradiol	9.4641	129.9	0.01	0.9419
2-methoxyestrone	24.1155	76.7396	0.10	0.7533
2-methoxyestradiol	-18.4875	102.5	0.03	0.8569
2-methoxyestradiol-3-methylether	-15.6975	91.6540	0.03	0.8640

3.2.3 TCDMS using Decision Tree

We estimated decision tree that determines thyroid cancer targeting 18 metabolite hormones. First,2-methoxyestrone, 2-methoxyestradiol and were selected among the 18 input metabolite as the factors that compose decision tree. It means hormones such as 2-methoxyestrone and 2-methoxyestradiol are selected as the key factors in the logistic

regression. 4-androstene- 3 and 17-dione were not used as important factors for the logistic but were selected for decision tree. Although the normal group and the thyroid cancer group of 4-androstene-3,17-dione didn't show significant difference as they are independent, the hormones such as 2-methoxyestrone and 2-methoxyestradiol were performing important role under controlled condition.

Figure 12. Estimated Decision Tree to determine Thyroid cancer.



As shown in the above figure 12, it is design to determine threshold value of the three hormones in the normal group and the thyroid cancer group. This result demonstrates rule-based such as [If ~ then ~] which could be useful for thyroid cancer analysis system. First, is numerical value of 2-methoxyestradiol reaches more than 25.82 and if numerical value of methoxyestrone is under 146.374, 0.1285 with 4-androstene-3,17-dione in the thyroid cancer that satisfy all three conditions, it is thyroid cancer of 100% accuracy. Whereas for the normal group, they are classified as a normal group if the numerical value is more than 146.374 for 2-methoxyestrone even if the normal group's numerical value is under 25.82 or more of 2-methoxyestradiol.

3.2.4 TCDMS using Neural Network

We made a neural network model in two hidden layers by inputting the hormones of the 18 metabolite. It was set on iteration 500,000 so that they don't fall into local minimization and the result of fitting, the weight for each hidden layer is as following. As a result of determining the Thyroid cancer group and the normal group through a neural network model, we were able to make a model with 100%accuracy.

Table 7. The neural network analysis result to establish TCDMS.

	From	To	Weight
1	androsterone	H11	-0.429375681
2	Etiocholanolone(ET)	H11	-0.616635629
3	dehydroepiandrosterone	H11	-0.914548297
4	4-androstene-3,17-dione	H11	-2.000986713
5	dihydrotestosterone	H11	-1.714964804
6	16α-hydroxy DHEA	H11	1.4601873375
7	5-androstene-3β,17β-diol	H11	-1.995114495
8	estrone (E1)	H11	-1.006136578
9	17β-estradiol (E2)	H11	-0.416675087
10	estriol (E3)	H11	-4.461939711
11	16α-hydroxyestrone	H11	1.6871608253
12	17-epiestriol	H11	-1.245385734
13	16-epiestriol	H11	0.9568905473
14	2-hydroxyestrone	H11	3.3183719019
15	2-hydroxyestradiol	H11	1.2129467013
16	2-methoxyestrone	H11	-0.733370592
17	2-methoxyestradiol	H11	4.718249406
18	2-methoxyestradiol-3-methylether	H11	10.112553426
19	BIAS	H11	6.2256201288
20	H11	GROUP	-8.940296225
21	BIAS	GROUP	-0.34053237

3.2.5 Leave-one-out cross validation for TCDMS

The data used for this analysis is consisted of the 23 people of the thyroid cancer group and 20 normal people from the normal group. The number of model was insufficient to divide in a training set and a validation set for an evaluation. Therefore, 1 data among the total 43 observation was eliminated for the validation set and the remaining 42 data were used models for the training set The validation set was made

by eliminating one each from the 43 people and 43 models have been fitting using the remaining data. As below, the leave-one-out cross validation table is made on the 43 models.

Table 8. Leave-one-out cross validation table for TCDMS.

Model	Accuracy	Sensitivity	Specificity	False Positive	False Negative
logistic Regression	43/43	20/20	23/23	0/20	0/23
Decision Tree	43/43	20/20	23/23	0/20	0/23
Neural Network	43/43	20/20	23/23	0/20	0/23

Each of the 43 models was precisely classified for possible thyroid cancel for the validation set. All three models demonstrate 100% accuracy. Therefore, the TCDMS that determines possible thyroid cancer using hormone estimation shows 100% accuracy.

3.2.6 Posterior Probability Pattern Analysis

We calculated posterior probability of 43 people through estimated TCDMS. We calculated the ranks of subjects according to calculated values of posterior probability. If the rank is high, probability to be thyroid cancer becomes high. Through the estimation, the posterior probability and rank are obtained. The each estimated rank is the order that shows the degree of disease risk. Since Decision tree calculates standard results based on each critical value of metabolite, the posterior probability value is divided in 0 and 1.0. Therefore, it is not practical information as a posterior probability value for hormone pattern. The probability value and rank for the logistic regression model and the neural network model are estimated and result can be seen in the table 9.

Table 9. Estimated posterior probability and rank table through TCDMS.

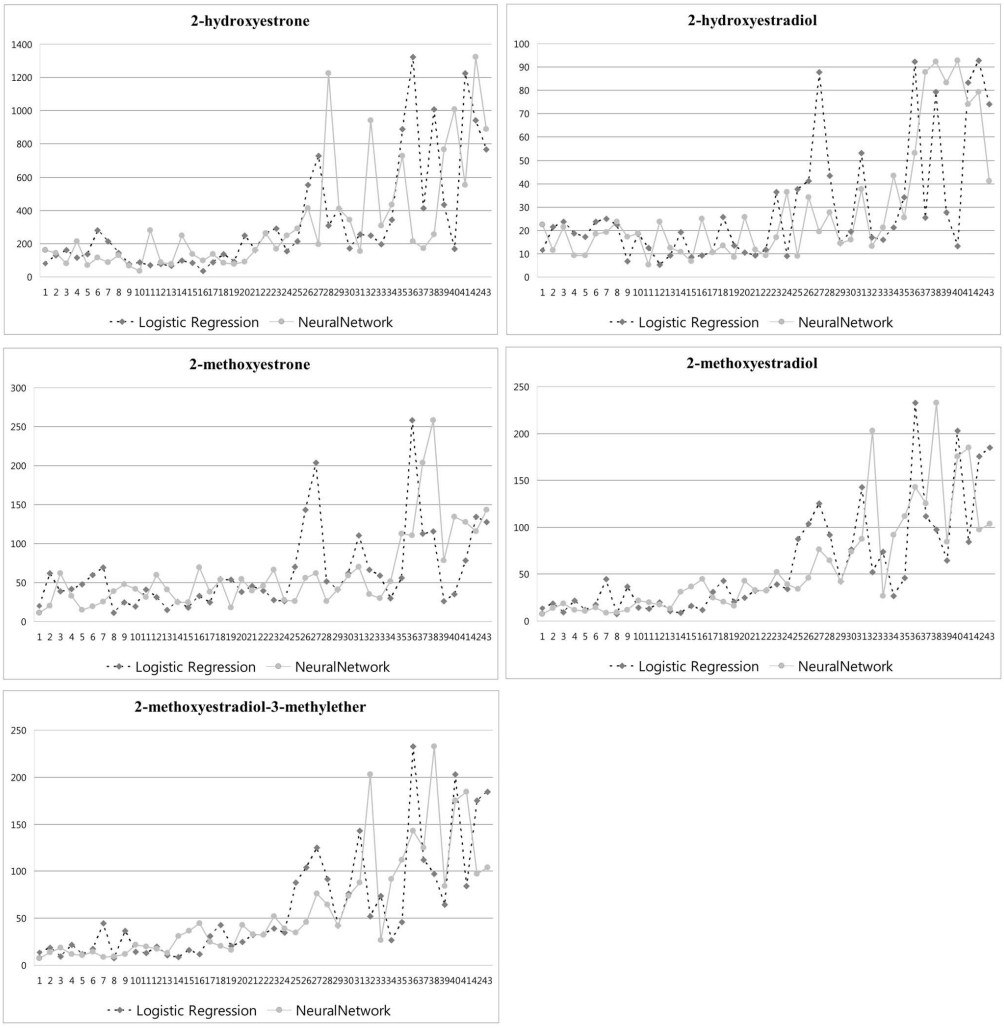
ID	GROUP	Posterior Prob. Tree	Posterior Prob. Logistic	Posterior Prob. Neural Network	Rank(Logistic)	Rank(NN)
1	cancer	1.0000000	0.9980790	0.9993111	21	22
2	cancer	1.0000000	1.0000000	0.9999068	37	36
3	cancer	1.0000000	0.9999903	0.9999068	31	38
4	cancer	1.0000000	0.9993590	0.9997032	24	25
5	cancer	1.0000000	0.9998270	0.9999068	27	37
6	cancer	1.0000000	1.0000000	0.9999068	42	32
7	cancer	1.0000000	1.0000000	0.9999068	36	39
8	cancer	1.0000000	0.9999999	0.9999068	33	30
9	cancer	1.0000000	0.9999929	0.9996486	32	23
10	cancer	1.0000000	1.0000000	0.9997222	35	26
11	cancer	1.0000000	1.0000000	0.9999068	40	41
12	cancer	1.0000000	0.9982680	0.9984273	22	21
13	cancer	1.0000000	0.9998820	0.9999065	29	29
14	cancer	1.0000000	1.0000000	0.9999068	34	33
15	cancer	1.0000000	0.9992140	0.9996487	23	24
16	cancer	1.0000000	1.0000000	0.9998891	39	28
17	cancer	1.0000000	0.9998570	0.9999068	28	34
18	cancer	1.0000000	1.0000000	0.9999068	43	43
19	cancer	1.0000000	1.0000000	0.9999068	41	42
20	cancer	1.0000000	1.0000000	0.9999068	38	40
21	cancer	1.0000000	0.9997330	0.9999068	26	35
22	cancer	1.0000000	0.9996580	0.9999068	25	31
23	cancer	1.0000000	0.9999582	0.9998890	30	27
24	normal	0.0000000	0.0002580	0.0001842	13	5
25	normal	0.0000000	0.0019650	0.0005993	20	17
26	normal	0.0000000	0.0005260	0.0005233	17	14
27	normal	0.0000000	0.0000000	0.0001841	2	2
28	normal	0.0000000	0.0000000	0.0001841	8	1
29	normal	0.0000000	0.0003620	0.0006114	15	19
30	normal	0.0000000	0.0002020	0.0002931	12	11
31	normal	0.0000000	0.0000000	0.0005675	7	16
32	normal	0.0000000	0.0006390	0.0006664	18	20
33	normal	0.0000000	0.0000290	0.0005286	9	15
34	normal	0.0000000	0.0000000	0.0002307	4	10
35	normal	0.0000000	0.0000000	0.0001841	1	3
36	normal	0.0000000	0.0000000	0.0003375	6	12
37	normal	0.0000000	0.0004330	0.0001841	16	4
38	normal	0.0000000	0.0000000	0.0001875	5	9
39	normal	0.0000000	0.0000760	0.0003758	11	13
40	normal	0.0000000	0.0012040	0.0006045	19	18
41	normal	0.0000000	0.0000000	0.0001850	3	8
42	normal	0.0000000	0.0002890	0.0001844	14	7
43	normal	0.0000000	0.0000380	0.0001843	10	6

Since algorithm to estimate a solution for the logistic regression and neural network is different, the posterior probability does not have identical value. However,

as a result of pearson correlation coefficient estimation to confirm correlation of posterior probability through the logistic regression and neural network, the two models have extreme correlation showing $r = 1.0000 (p - value < 0.00001)$ and had nearly identical result. In another simplified analysis result of Pearson correlation coefficient on the rank of the two models, we were able to obtain high correlation at $r = 0.9003 (p - value < 0.00001)$. This means, that although the rank of the thyroid cancel on the observation of the 43 people and the rank of the neural network through the logistic regress are not exactly the same, it showed a result of near conformity. If arrange the patients of thyroid cancer and the normal people in line through the estimated rank, it shows the disease transfer to the normal people. The figure 13 demonstrates hormone changes of risk factors in thyroid cancer through a screening test. It refers to high level of thyroid cancel as the rank goes up. As rank goes up, the 5 hormones usually maintain high numeric value. In addition, larger vibration is found as the rank gets higher. Frequency pattern is a special trait of time series. While 2 – hydroxyestrone demonstrates stable hormone value of under 300 from the normal people under the rank 20, it showed high value of more than 300 from the cancer patient from the rank 22. While the normal people demonstrate stable distribution at under 30 with 2 – hydroxyestradiols, it showed above 30 as it advanced to thyroid cancer. Whilst 2 – methoxyestradiols and 2 - methoxyestradiol - 3 – methylethers showed almost no sign of tendency among the normal people under the rank 20, it clearly showed sign of tendency as it advanced to cancer. While 2 – methoxyestradiols demonstrates stable value of under 50 at the normal group of under the rank 20, the value increased as the rank increased. 2 - methoxyestradiol - 3 – methylethers include relatively large vibration under the rank 20 and remains at under 50but show above 50 as the rank increased. While at the low rank, in other words, the group that is stable with thyroid cancer does not show the pattern, the thyroid cancer which is the 21 rank showed tendency of increase. Hormone distribution is low at the normal group and for the thyroid cancer group, it generally show high value

and seldom low value as well.

Figure 13. Pattern monitoring of risk factors according to posterior probability.



3.3 case 3 study : Results

The gene expression data measured from 10 people from the normal group and 15 people from the colon cancer group was analyzed using the ANOVA model.

The study hypothesis is on whether there are differences of effect among metabolic pathways when colon cancer developed. 110 metabolic pathways out of 148were used for this analysis. Since the probability of 110 metabolic pathway effects being 0 is low, SNK analysis was used to test the correlation between a cluster with high effect and colon cancer by building a cluster between metabolic pathways.

3.3.1 metabolic pathway effect analysis using ANOVA model

Through $abs(x_{ijg})$ that has been converted into absolute value and gene expression value x_{ijg}, this study tested the difference of effects in metabolic pathways and whether there is a difference of effects between the colon cancer and the normal groups. ANOVA test was done to estimate if there were any differences between group effect and among metabolic pathways as illustrated in table 10. The results showed a significant difference in metabolic pathways and interaction effect (Path*Group). (under 0.05 significant level).On the contrary, a significant difference of effect between the normal group and the cancer group was not shown. This suggests that even if many numbers of genes are expressed, the genes of the cancer group cannot be reflected due to up and down regulations. These problems can be solved if the absolute value of gene expression is used. In the case of $abs(x_{ijg})$ which uses absolute variation if data, the difference of effect (p-value < 0.0001) between normal and cancer groups can be seen as in table 10.

Table 10. ANOVA table for metabolic pathway effect and group effect test.

	$x_{ijg} = \mu + \tau_p + \rho_g + (\tau\rho)_{pg} + \epsilon_{ijgp}$					$abs(x_{ijg}) = \mu + \tau_p + \rho_g + (\tau\rho)_{pg} + \epsilon_{ijgp}$				
	DF	SS	MS	F	Pr>F	DF	SS	MS	F	Pr>F
Path	110	4370.33	39.73	30.84	<.0001	110	2066.62	18.78	27.24	<.0001
Group	1	0.06	0.06	0.05	0.8319	1	113.58	113.58	164.67	<.0001
Path*Group	109	687.81	6.31	4.90	<.0001	109	339.26	3.11	4.51	<.0001

Gene expression shows the difference in the group effect and the difference in the metabolic pathway effect through table 10. It is evident that interaction between metabolic pathway and the group exists and the main purpose of this study is to check whether or not a difference between metabolic pathway and effect exists. Hence, ANOVA test was carried out by dividing the data into normal group and colon cancer group. We were able to check that there was a difference of effect between metabolic pathways as a result of ANOVA test in the normal group and the colon cancer group through table 11.

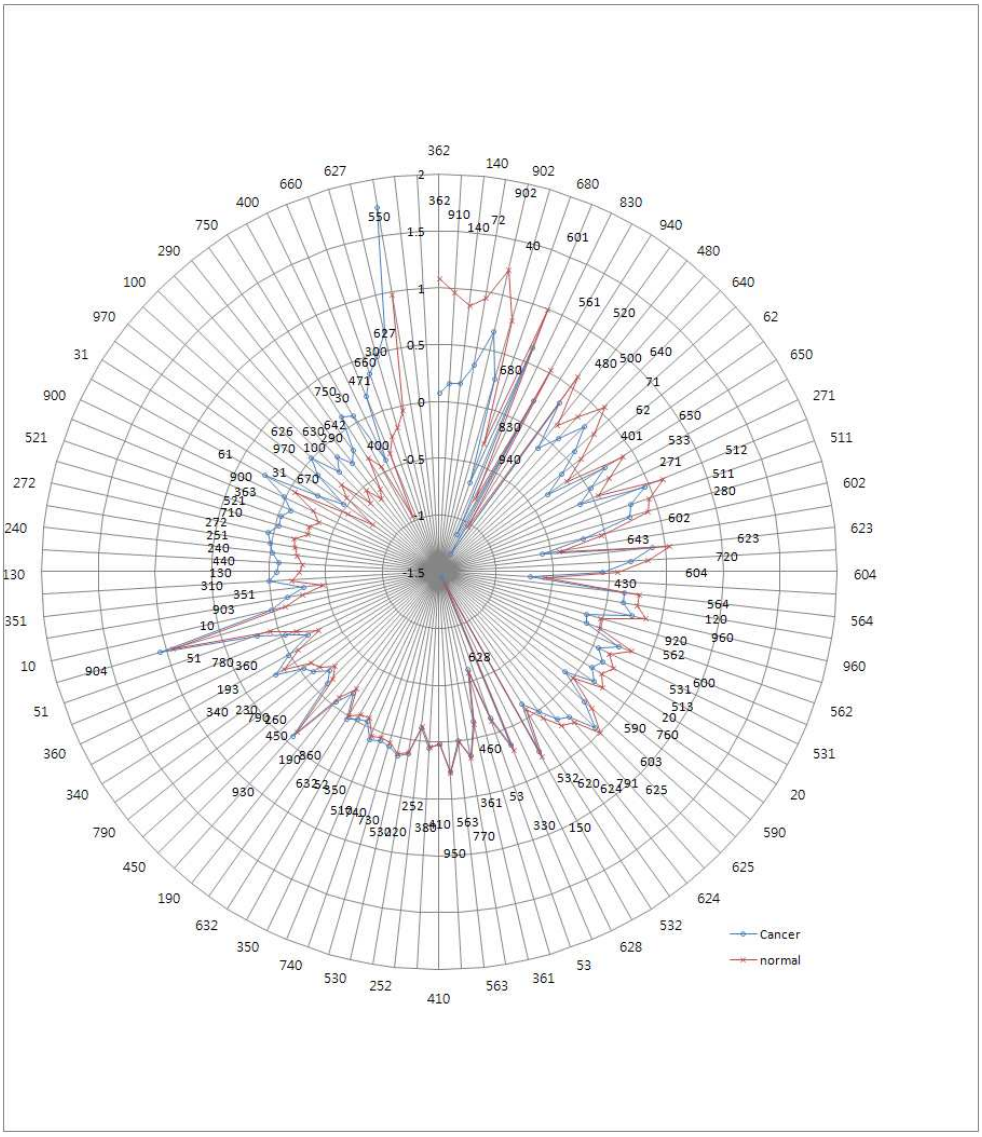
Table 11. ANOVA table for metabolic pathway effect test between normal group and colon cancer group.

	$x_{ijg} = \mu + \rho_g + \epsilon_{ijgp}$					$abs(x_{ijg}) = \mu + \rho_g + \epsilon_{ijgp}$				
	DF	SS	MS	F	Pr>F	DF	SS	MS	F	Pr>F
Path (Normal)	110	2997.56	27.25	16.70	<.0001	110	1226.41	11.14	12.70	<.0001
Path(Cancer)	109	1848.67	16.96	16.11	<.0001	109	1172.55	10.75	19.19	<.0001

The Student-Newman-Keuls (SNK) Test was carried out in order to verify how many clusters metabolic pathways are formed into and what they are at the highest level cluster if there was a difference between metabolic pathways. SNK test for normal group and colon cancer group was done separately. The SNK test shows that the normal group is divided into 11 levels of cluster and the colon cancer group is

divided into 21 levels as in figure 14. The level of effect is divided into 11 clusters and we intend to estimate the correlation between metabolic pathway and colon cancer that belongs to the cluster with the highest expression level.

Figure 14. Average expression value of genes regarding metabolic pathway of normal group and colon cancer group.



In comparison to the normal group, colon cancer group is divided into more various levels. The correlation between metabolic pathway and colon cancer will be analyzed after the SNK test. Figure 14 is a picture expressing the pathway effect on metabolic pathway between the normal group and the colon cancer group where genes in Glycosphingolipid biosynthesis metabolism and Retinol metabolism in the normal group can be seen being highly expressed. In the colon cancer, on the other hand, genes in Phenylpropanoid biosynthesis metabolism or Methane metabolism can be seen being highly expressed.

3.3.2 Significant metabolic pathway about colon cancer using ANOVA

Metabolic pathway that had estimated high expression levels among the cluster in SNK test of the normal and the colon cancer group was selected. There is a possibility that the selected metabolic pathway which used normal group data will equally show high expression level in the colon cancer group which is comparable to the normal group. Therefore, the normal group and the comparable colon cancer group that was used in the SNK test will carry out a t-test and metabolic pathway that show a difference between the two groups will be selected. T-test on the metabolic pathway that was selected through SNK test regarding colon cancer group was carried out. Even though the metabolic pathway has a high expression level, there is a possibility of both groups having a high expression level at the same time so we carried out a paired *t* test regarding the expression level of the two groups.

Table 12. Metabolic pathway in high level and the paired t test result.

Group	Metabolic Pathway	DF	t	Pr>t
Normal	Glycosphingolipid biosynthesis	346	-3.67	0.0003
	Retinol metabolism	127	-0.21	0.8372
	Synthesis and degradation of ketone bodies	110	2.64	0.0137
	Nitrogen metabolism	885	-7.12	<.0001
	Ethylbenzene degradation	201	-1.77	0.0787
	Fluorene degradation	104	0.03	0.9763

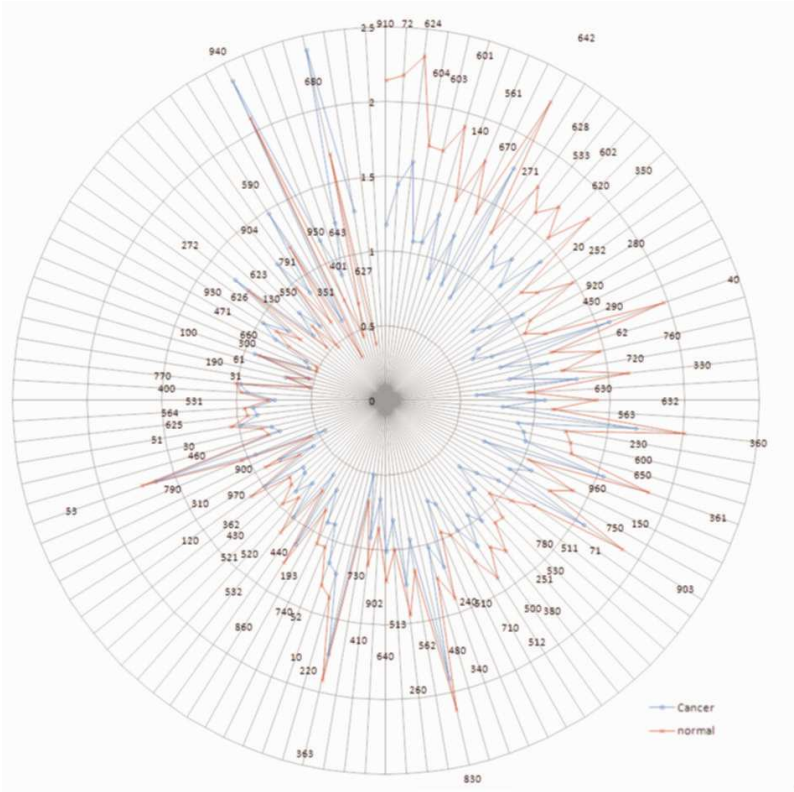
Group	Metabolic Pathway	DF	t	Pr>t
Cancer	Phenylpropanoid biosynthesis	185	1.14	0.2548
	Methane metabolism	177	2.11	0.0365
	Fluorene degradation	104	0.03	0.9763
	Retinol metabolism	127	-0.21	0.8372
	Ascorbate and aldarate metabolism	332	0.48	0.6320
	Phenylalanine metabolism Phenylalanine metabolism	523	-2.12	0.0346

As seen in table 12, Retinol metabolism and Fluorene degradation metabolism is the metabolic pathway with a high expression level in both normal and colon cancer groups. This metabolic pathway shows high expression level where there is no correlation to colon cancer. The metabolic pathways which show high level of expression in the colon cancer group and significant difference from the normal group are known as methane metabolism and phenylalanine metabolism. As for methane metabolism, its correlation with colon cancer was verified in previous studies. [68] Furthermore, based on a lecture study, it can be believed that Phenylalanine Metabolism is correlated to colon cancer. [69] The selected methane metabolism and phenylalanine metabolism are manifested at a high level when cancer is developed and show a significant difference from the normal group. Although the two metabolic pathways were clinically verified in previous studies, the difference of effect between the metabolic pathways as well as its correlation with diseases can be estimated when gene expression data is analyzed using the ANOVA model and SNK test.

3.3.3 Significant metabolic pathway about colon cancer using MSE

We calculated the MSE value according to each metabolic pathway regarding the normal group and the colon cancer group. The result of calculating MSE of the metabolic pathway showed a distribution of MSE between 0.33 and 2.31 in the normal group and a distribution between 0.45 and 2.41 in the colon cancer group. In other words, the variation of genes in metabolic pathway in the colon cancer group is higher compared to the normal group.

Figure 15. Distribution of MSE value according to the metabolic pathway of the normal group and the colon cancer group.



The MSE values of the metabolic pathway are shown in figure 15. As in the figure, the MSE value of 2.40 is the highest under the colon cancer condition in Methane Metabolism (KEGG Pathway ID =680) and 2.36 is the second highest in the Phenylpropanoid Biosynthesis (KEGG Pathway ID =940). On the other hand, Methane metabolism shows a relative high MSE value of 1.69 in the normal group but it shows a difference of 0.71 to the colon cancer group. Phenylpropanoid Biosynthesis shows 2.09 in the normal group while the difference is only 0.27 in the colon cancer. Therefore, if the normal group is considered as the control, Methane Metabolism displays significant variation in the colon cancer but Phenylpropanoid Biosynthesis maintains a similar value to the normal group. In the case of the colon cancer group, if the variation of genes in comparison to the normal group is considered as the control, it is the same as table 13.

Table 13. Metabolic pathway of high correlation to colon cancer through MSE.

Metabolic Pathway	MSE_{normal}	MSE_{cancer}	$Diff = MSE_{normal} - MSE_{cancer}$
1,4-Dichlorobenzene degradation	0.38	1.28	0.9
Methane metabolism	1.69	2.4	0.71
Styrene degradation	0.67	1.23	0.56
Novobiocin biosynthesis	0.45	0.89	0.43
Alkaloid biosynthesis	0.72	1.15	0.42
1,1,1-Trichloro-2,2-bis(4-chlorophenyl)ethane (DDT) degradation	0.33	0.6	0.27
Phenylpropanoid biosynthesis	2.09	2.36	0.27

Table 13 illustrates 7 metabolic pathways of a high $Diff = MSE_{normal} - MSE_{cancer}$ value. There has been an estimation that the correlation with the colon cancer would be significant in Methane Metabolism and Phenylpropanoid Biosynthesis as in table 12. It is known from lecture studies that Alkaloid metabolism, a basic chemical substance

containing nitrogen, emits nitric oxide from the colon tumor when colon cancer is developed. [70] Styrene, one of ambient aromatic hydrocarbons with benzene nucleus, and Dichlorobenzene, substituent of benzene, seem to have a high probability of development of colon cancer and other various cancers when they are exposed to Styrene or benzene. [71] 1,1,1-Trichloro-2,2-bis(4-chlorophenyl)ethane (DDT) degradation, 1,4-Dichlorobenzene degradation and Styrene degradation are xenobiotics metabolism. When xenobiotics metabolism is exposed to harmful substances from the outside and when a part of these accumulate in cells, it plays a role in easily emitting by changing into a soluble condition. [72] At this point, a process of neutralizing takes place through cytochrome P450 (CYP), the phase I enzymes, and glutathione-S-transferase (GST), the phase II enzymes. Genes such as CYP is known to be related to the correlation between smoking and human cancer risk. [73] Through the gene expression experiment, it is also known to have correlation with colon cancer. [74] When MSE value is used, the correlation with colon cancer could be estimated by finding the metabolic pathway of colon cancer group which has a higher variation compared to the normal group. Moreover, this could be verified once more through the lecture study on previous clinical experiment.

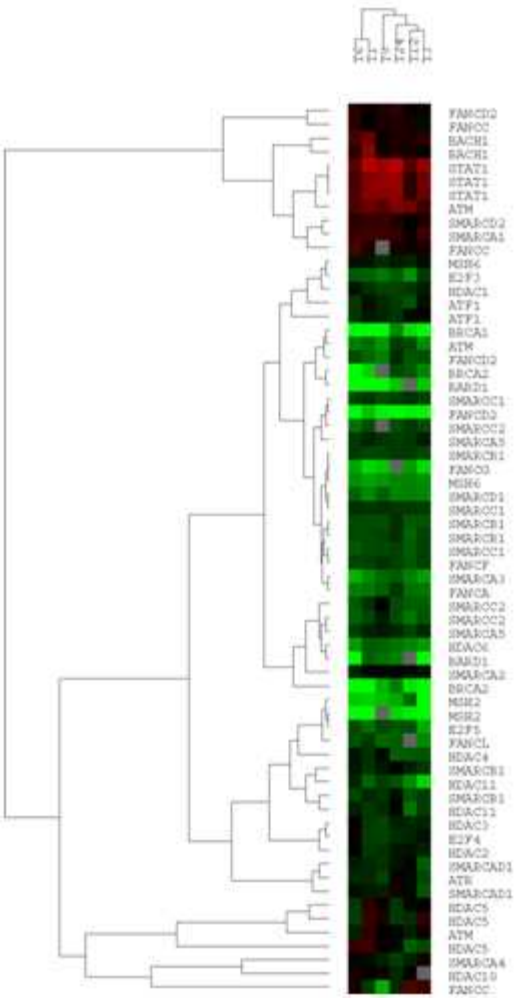
3.4. case 4 study : Results

3.4.1 Tumor Suppressor BRCA1and Signaling Pathway gene

The normalization and filtering processes were carried out for 13951 gene expression data measured by exposing vascular Smooth Muscle cells in serum. As a result of observing 39 genes in the same Pathway as BRCA1, a tumor suppressor of breast cancer, gene expression value of the BRCA1 gene was high at each point so it is excluded when it goes through the filtering process. However, as the purpose of

this study is to analyze the correlation between BRCA1 and its surrounding genes according to the gene expression value pattern, we did not delete the genes that have been through filtering. As a result of carrying out a clustering analysis for only 39 genes in the cell signaling Pathway, BRCA1 had the closest distance to ATM, FANCD2, BRCA2 and BARD1.

Figure 16. The Cluster Analysis results regarding BRCA1 gene and existing genes in BRCA1 Pathway.

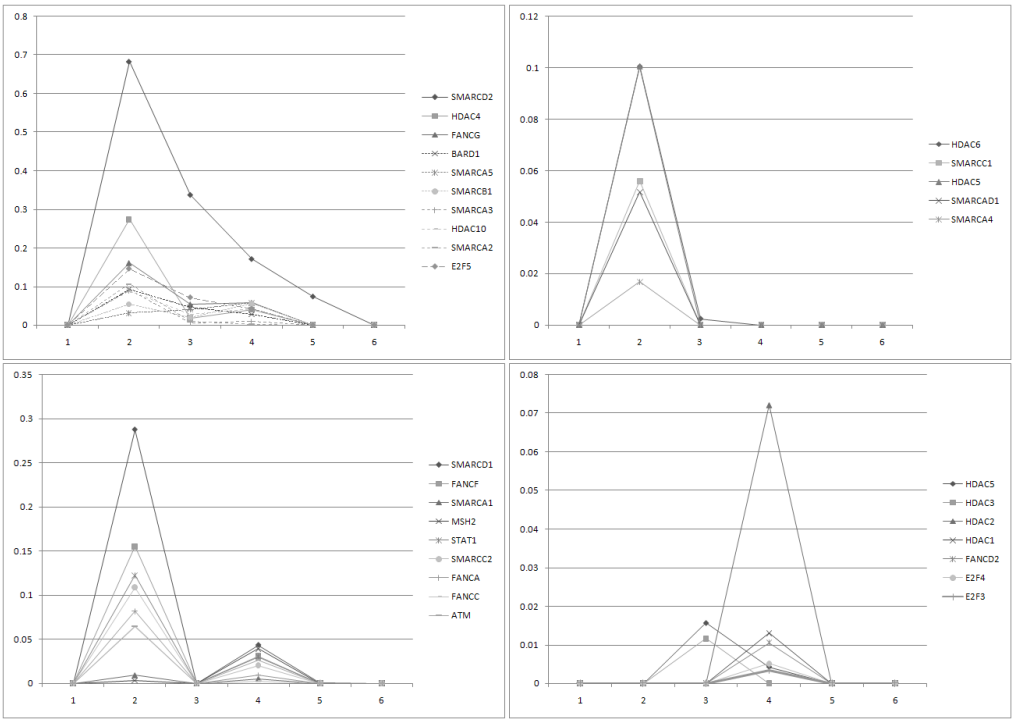


On the contrary, Spearman coefficient, which shows non-parametric correlation between the two genes, indicate a high interrelationship with genes such as MSH2, SMARCD, SMARCB, SMARCC and ATM. ATM is a gene related to phosphorylation of BRCA1 and SMARCD, SMARCB and SMARCC are SWI/SNF, which is a DNA regulatory complex. Only BRCA1 and ATM are shown as genes with a high interrelationship when clustering analysis and spearman coefficient, which shows non-parametric correlation, are compared.

3.4.2 Tumor Suppressor BRCA1 and Signaling Pathway gene

The BRCA1 gene and 39 genes in the signaling pathway were analyzed through SPAL program. It was possible to observe the pattern of genes that were effected in regards to the attractor gene through SPAL program. The variation of Lyapunov exponent was observed according to time using the initial condition of change with the attractor.

Figure 17. The variation of Lyapunov exponent of genes in BRCA1 and BRCA2 pathway.



Lyapunov exponent is defined to have a significant correlation in relation to an attractor on the section that obtains positive value. The Lyapunov exponent between genes in BRCA1 pathway and BRCA1 gene is divided into four main groups as in figure 17. In the first group there are E2F5, HDAC10, HDAC4, FANCG, BARD1, SMARCD2, SMARCA5, SMARCB1, SMARCA3 and SMARCA which are genes that

have a positive lyapunov exponent from the beginning time until the end time of the observation. These genes have an intimate correlation regarding the changes in BRCA1 while they change themselves as well. SMARCD2, SMARCA5, SMARCB1, SMARCA3 and SMARCA2 which belong to SWI/SNF, a DNA regulatory complex, are in charge of remodeling chromatins for transcription machineries to approach the genes. BRCA1 has a high possibility of interacting with BRG1 among SWI/SNF complex. Among Histone deacetylases that are in charge of chromatin remodeling, HDAC10 and HDAC4 are influenced by BRCA1 across the whole section. In the second group there are HDAC6, HDAC5, SMARCC1, SMARCAD1 and SMARCA4 which are genes that are effected byBRCA1 only in the beginning of the observation. They are all involved in chromatin remodeling but in contrast with the first group, they are involved only in the beginning. In the third group there are ATM, FANCF, FANCA, FANCC, SMARCA1, SMARCD1, MSH2, STAT1 and SMARCC2 which are genes that are influenced by BRCA1 gene in the beginning of the observation and once more later. ATM is a gene that influences the phosphorylation of BRCA1. Conversely, ATM obtained a positive lyapunov exponent value in relation to BRCA1 gene in two phases. FANCF, FANCA and FANCC are genes that do cell cycle checkpoint arrest. In the fourth group there are HDAC5, HDAC3, HDAC2, HDAC1, FANCD2, E2F4 and E2F3 which are genes that are influenced by BRCA1 in the middle time of the observation. Usually they are transcription factors that increase the expression level as BRCA1 and genes in charge of chromatin remodeling are phosphorylated. While genes belonging to SWI/SNF, a DNA regulatory complex, are consistently effected by BRCA1 from the beginning, genes that are related to cell cycle checkpoint arrestare influenced twice, in the beginning and middle, and the transcription factors that phosphorylate BRCA1 are effected towards the end.

IV. Discussion

4.1 Case 1. Study Discussion

Conventionally, the studies on outbreak cause of thyroid gland disease were concentrated to the environmental effects. Amongst the people in the same environmental position however, there was a significant difference in the outbreak of the disease, thus the previous studies concluded that this phenomenon could be caused by genetic factor. More researches are required to find out which hormone influences the outbreak of the disease however; there isn't any clarified or completed study result on metabolites in steroid map. Therefore, we have studied t the TGDDM system using the metabolic profile, which includes measured value of unknown metabolites. The parallel research on the known metabolites affecting thyroid gland disease reveals that the hormones such as 2-hydroxyestrone, 2-methoxyestrone, 2-hydroxyestradiol, 2-methoxyestradiol, and 2-methoxyestradiol-3-methylether are the important risk factors of outbreak of the disease, but it merely means they are just a part of the risk factors, not all the factors. This is the reason why we are developing the TGDDM system considering all the materials through metabolic profile. With respect to retention time interval size, the number of unknown materials in the interval is different. In this research, we classified the data by translating variables and t-test and constructed the TGDDM system using logistic regression, which take above classified variables as an explanatory variable in order to properly find out materials that explain the characteristics of the thyroid gland disease. In the further study, we are going to emphasis our research on clarifying unknown materials that are included in the system.

4.2 Case 2. Study Discussion

This study is to develop TCDMS through hormones in androgen and estrogen metabolic pathway and estimate how the patterns of hormones change as it progress to cancer. Therefore we examined the availability of risk factors and estimated the three models of thyroid cancer with hormones that exist in androgen and estrogen metabolic pathway in order to find the causes of thyroid cancer outbreak from metabolic pathway. We found that metabolites of 2 - hydroxyestrone, 2 - hydroxyestradiols, 2 - methoxyestrone, 2 – methoxyestradiols and 2 - methoxyestradiol - 3 - methylethers were risk factors of thyroid cancer. Furthermore, we were able to estimate a superior model that had 100% accuracy through TCDMS. We monitored the density change of hormone as thyroid cancer grew through posterior probability of subject estimated from TCDMS and the patterns of metabolites through the model and found that the densities of hormones maintained high value with unstable distribution as thyroid cancer advances. It is anticipated that the hormone pattern could be used as a useful element in diagnosing thyroid cancer as TCDMS and thyroid cancer advance. Thyroid cancer can be diagnosed through measuring hormone and a degree of the cancer could be estimated by the pattern monitoring.

4.3 Case 3. Study Discussion

Analyzing the correlation between metabolism and diseases is a very important task. This research used the analysis unit of metabolic pathway, a field of sets derived from genes, through gene expression data, rather than a gene. A test was carried out using ANOVA model to see if there is a difference of effects between metabolic pathways in colon cancer and was able to estimate that methane metabolism and

phenylalanine metabolism have a significant correlation with the colon cancer. Through SNK test, the fact that a metabolic pathway divides up into many levels and reacts when colon cancer is developed could be seen. Also, we were able to see that Dichlorobenzene and Styrene Metabolism are highly correlated to colon cancer through MSE. Similarly, a metabolic pathway with a high correlation to colon cancer could be found through variation of gene expression. Styrene and Dichlorobenzene which are correlated to colon cancer are substances that are related to benzene known as carcinogen. In the case of cancer development, there is a lot of variation in genes in Styrene metabolism and Dichlorobenzene metabolism in the body. These metabolic pathways are about xenobiotics that can be considered to have correlations to colon cancer. Although there are no clinical experiment results on the correlation between colon cancer and Novobiocin biosynthesis, a metabolism regarding an antibiotic that Actinomyces produces, it is estimated that a correlation exists. There has not been any mention of correlations to Alkaloid biosynthesis metabolism either which is related to the production of basic organic compound containing nitrogen but apparently a large amount of nitric oxide is produced. We tried to find metabolism that showed high variation in colon cancer through this research. As a result, we were able to find metabolisms related to high variation including those already known for their correlations through clinical experiment results. As this analysis is from a broad perspective, we must find correlation between mutual genes, significant genes in colon cancer and expression pathway within the pathway through detailed analysis about genes that subsist within the metabolic pathway. However, there is a limit in interpreting this as the pre-existing analysis algorithm because the genes have to be tested under the structure called network. We intend to study about genes and colon cancer which forms a network structure in future studies.

4.4 Case 4. Study Discussion

We developed the SPAL program to analyze the correlation of time series data that showed frequency pattern and analyzed the genes in BRCA1, a breast tumor suppressor, and BRCA1 Pathway. This program calculates the lyapunov exponent of the genes by inputting the gene expression data and pathway data. Lyapunov exponent is an exponent that measures chaos stability, which the initial condition between the two systems changes as time passes. This was applied and used as an exponent measuring the correlation of tumor suppressor gene related genes. It was possible to observe the order of response of genes in the Pathway by measuring the degree and point of response against the tumor suppressor. As a result, while genes belonging to SWI/SNF, a DNA regulatory complex, obtaina positive lyapunov exponent with BRCA1 from the beginning, genes that cell cycle checkpoint arrest obtain a positive exponent in the beginning and middle. On the other hand, genes that are transcription factors which phosphorylate BRCA1 usually obtain a positive lyapunov exponent towards the end. In particular, SMARCD2 genes have a quite high lyapunov exponent with BRCA1. The influence that BRCA1 has on its surrounding genes and its order for knocking-out a gene can be analyzed using the SPAL program. This program is created especially to find out the order and genes that tumor suppressor affects regarding gene expression data on cancer and it is possible to apply it to various cancers.

V. Conclusion

The purpose of this study is to measure the influence metabolic pathway, signaling pathway and disease pathway which are made up of a network structure has when a disease develops and to estimate the compounds and genes that become a risk factor in this process. There were not many cases in the previous analysis where information on pathway concerning experimental data on diseases was used. When the information on pathway is used, plentiful conclusions are drawn as they are drawn after considering the surrounding genes or metabolisms. Also, the results can be applied like a new hypothesis. This study developed procedures and research methods required for analyzing measured data of phenomena occurring in the cells of an organism based on a pathway. Pathways are divided into enzymes that are composed of compounds and genes which fall under the chemical heading. When a compound is measured through an experiment, the analysis procedure is shown through a case study for the analysis on the correlation with diseases depending whether the profile is mixed with many substances or whether the metabolite well-known. As for the enzyme, an analysis from a macroscopic perspective and an analysis procedure from a microscopic perspective are shown by case by case using gene expression data. This study aimed to analyze from a systematic science perspective using the previous statistical method. We modified the data into information using the statistical method and developed a system. There were various conclusions drawn through the analysis of the four cases on the risk factor of cancers, metabolism and the influence tumor suppressor has on range and order. This study analyzed the thyroid gland disease, thyroid cancer, colon cancer and breast cancer. By applying the analysis procedure of the four case studies on other various cancers, their causes and risk factors can be studied.

Reference and Notes

[1] G. Kitano, Foundations of Systems Biology. The MIT Press, Cambridge, MA, 2001.

[2] EU Projects Workshop Report on Systems Biology. 2004. Available at http://www.wtec.org/sysbio/

[3] D.L. Nelson and M.M. Cox, Lehninger Principles of Biochemistry. Worth Publisher, Korea, 2002.

[4] M.L. Mavrovouniotis, Identification of Qualitatively Feasible Metabolic Pathways in Artificial Intelligence and Molecular Biology. MIT Press, Cambridge, MA, 1993.

[5] Raaijmakers and G.W. Jeroen, Statistical analysis of the Michaelis-Menten equation. Biomet 43 (1987) 793-803.

[6] E.C. Horning and M.C. Horning, in A. Zlatkis(Editor), Advances in Chromatography. University of Houston Press, Houston, 1970

[7] C.H.L. Shachleton, J.A. Gustafsson and F.L. Mitchell, Steroids in newborns and infants; The changing pattern of urinary steroid excretion during infancy. Acta Endocrinol (Copenh). 744 (1973) 157-167.

[8] J.A. Luyten and G.A. Rutten, Analysis of steroids by high-resolution gas-liquid chromatography. II. Application of urinary samples. J Chromatogr 91 (1974) 393-406.

[9] H. Ludwig, J. Reiner and G. Spiteller, Investigation of the steroids in blood with the combination glass capillary column gas chromatography-mass spectrometry. Chem Berichte 110 (1977) 217-230.

[10] C.D. Pfaffenberger and E.C. Horning, High-resolution biomedical gaschromatograpy; Determination of human urinary steroid metabolites using glass open tubular capillary columns. J Chromatogr 112 (1975) 581-594.

[11] C.D. Pfaffenberger and E.C. Horning, Sex differences in human urinary steroid

metabolic profiles determined by gas chromatography. Anal Biochem 80 (1977) 329-343.

[12] H. Yamanaka, M. Nakajima, K. Nishimura, R. Yoshida, T. Fukami, M. Katoh and T. Yokoi, Metabolic profile of nicotine in subjects whose CYP2A6 gene is deleted. Eur Jour of Pharma Sci 22 (2004) 419–425.

[13] H. Wu, X. Zhang, X. Li, Z. Li, Y. Wu and F. Pei, Comparison of metabolic profiles from serum from hepatotoxintreated rats by nuclear-magnetic-resonance-spectroscopy-based metabonomic analysis. Analyt Biochem 340 (2005) 99–105.

[14] A.M. Lavermana, M. Brasterb, W.F.M. Ro¨lingb and H.W. van Verseveldb, Bacterial community structure and metabolic profiles in a forest soil exhibiting spatially variable net nitrate production. Soil Bio. & Biochem xx (2005) 1–8.

[15] N.D. Tronko, O.O. Bobylyova, T.I. Bogdanova, et al. Thyroid gland and radiation (Ukrainian-American Thyroid Project). International Congress Series 1258 (2003) 91-104.

[16] J. Robbins and A.B. Schneider. Radioiodine-induced thyroid cancer: Studies in the aftermath of the accident at Chernobyl. Trends Endocrinol Metab 9 (1998) 87-94.

[17] S.T. Bennett and C. Schmutzler, Thyroid gland goes genomic. International Symposium: Genetics of Thyroid Disease. Trends Genet 13 (1997) 468.

[18] K.M. Kim, B.H. Jung, D.S. Lho, W.Y. Chung, K.J. Paeng and B.C. Chung, Alteration of urinary profiles of endogenous steroids and polyunsaturated fatty acids in thyroid cancer. Cancer Lett 202 (2003) 173-9.

[19] D. Voet and J.G. Voet, Biochemistry. Wiley & Sons, New York, 1995.

[20] H. Ogata, S. Goto, W. Fujibuchi, and M. Kanehisa, Computation with the KEGG pathway database. BioSystems 47 (1998) 119-128.

[21] V.N. Reddy, M.N. Liebman and M.L. Mavrovouniotis, Qualitative analysis of biochemical reaction systems. Comput Biol Med 26 (1996) 9-24.

[22] H. Kacser and J.A. Burns, The control of flux. Symp Soc Exp Biol. 27 (1973) 65−104.

[23] U.S. Pasaribu, A.G. Hawkes and S.J. Wainwright, Statistical assumptions underlying the fitting of the Michaelis-Menten equation. J App Stat 26 (1999) 327-341.

[24] E.P. van Someren, L.F.A. Wessels and M.J.T. Reinders, Linear modeling of genetic network from experimental data. Proc ISMB 8 (2000) 355-366

[25] A.M. Sääf, J.M. Halbleib, X. Chen, S.T. Yuen, L.S. Yi, W.J. Nelson and P.O. Brown, Parallels between Global Transcriptional Programs of Polarizing Caco-2 Intestinal Epithelial Cells in vitro and Gene Expression Programs in Normal and Colon Cancer. Mol Biol Cell (15 August 2007)

[26] M. Kanehisa, S. Goto, S. Kawashima and A. Nakaya, The KEGG databases at GenomeNet. Nucleic Acids Res 30 (2002) 42-46.

[27] B. Tehard, P.H. Lahmann, E. Riboli and F. Clavel-Chapelon, Anthropometry, breast cancer and menopausal status: Use of repeated measurements over 10 years of follow-up -results of the French E3N Women's cohort study. Int J Cancer 111 (2004) 264-269.

[28] J. Manjer, R. Johansson, G. Berglund, L. Janzon, R. Kaaks, A. Agren and P. Lenner, Postmenopausal breast cancer risk in relation to sex steroid hormones, prolactin and SHBG (Sweden). Cancer Causes Control. 14 (2003) 599-607.

[29] D.F. Easton, D. Ford and D.T. Bishop, Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. American Journal of Human Genetics. 56 (1995) 265-271.

[30] R.W. Martin, B.J. Orelli, M. Yamazoe, A.J. Minn, S. Takeda, and D.K.

Bishop, RAD51 Up-regulation Bypasses BRCA1 Function and Is a Common Feature of BRCA1-Deficient Breast Tumors. Cancer Res 67 (2007) 9658-9665.

[31] J.T. Chi, E.H. Rodriguez, Z. Wang, D.S.A. Nuyten, S. Mukherjee, et al., Gene Expression Programs of Human Smooth Muscle Cells: Tissue-Specific Differentiation and Prognostic Significance in Breast Cancers. PLoS Genet 3 (2007) 1770-1784.

[32] W.H. Elliott and D.C. Elliott, Biochemistry and molecular biology. Oxford University Press, Oxford, 1997.

[33] M. Razandi, A. Pedram, E. Rosen, and E.R. Levin, BRCA1 inhibits membrane estrogen and growth factor receptor signaling to cell proliferation in breast cancer. Mol Cell Biol 24 (2004) 5900-5913.

[34] Y.B. Pesin, Characteristic Lyapunov Exponents and Smooth Ergodic Theory. Russian Math Surveys. 32 (1977) 55-114.

[35] J. Banks, V. Dragan and A. Jones, Chaos : a mathematical introduction. Cambridge University Press, Cambridge, New York, 2003.

[36] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno and M. Hattori, The KEGG resources for deciphering the genome. Nucleic Acids Res 32 (2004) 277-280.

[37] C.H.L. Shackleton, Profiling steroid hormones and urinary steroids. J Chromatography 379 (1986) 91-156.

[38] B. Rosner, Fundamentals of Biostatistics Duxbury Press, California, 2005.

[39] Y.S. Kim and S.Y. Sohn, Screening test data analysis for liver disease prediction model using growth curve. Biomed Pharm 57 (2003) 482-488.

[40] G.C. Cawley and N.L.C. Talbot, Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. Pattern Recog 36 (2003) 2585-2592.

[41] Korea Institute of Science and Technology Bioanalysis and Biotransformation Research Center. Available at http://www.kist.re.kr

[42] S.H. Lee, K.M. Kim, B.H. Jung, W.Y. Chung, C.S. Park and B.C. Chung, Eestrogens in female thyroid cancer: alteration of urinary profiles in pre- and post-operative cases. Cancer Lett 189 (2003) 27-32.

[43] C.T. Le, Introductory biostatistics. Wiley-Liss, Inc, New York, 2003

[44] W.W. Daniel, Biostatistics: a foundation for analysis in the health sciences. Wiley, New York, 1983.

[45] L. Breiman, J.H. Friedman, R.A. Olsen and C.J. Stone, Classification and Regression Trees. Wadsworth International Group, Belmont, 1984.

[46] J.R. Quinla, C4.5 Programs for Machine Learning. San Mateo, Morgan Kaufmann, 1993.

[47] M.Y. Hu, M. Shanker and M.S. Hung, Estimation of posterior probabilities of consumer situational choices with neural network classifiers. Int J Res Mark 1999 (16) 307-317.

[48] A.W. Ambergen and W. Schaafsma, Interval estimates for posterior probabilities in a multivariate normal classification model. J Multivariate Anal 16 (1985) 432-439.

[49] G.C. Chow, A comparison of the information and the posterior probability. criteria for model selection. J Econometrics 16 (1981) 21-33.

[50] G.C. Cawley and N.L.C Talbot, Fast exact leave-one-out cross-validation of sparse. Neural Networks 17 (2004) 1467-1475.

[51] T. Zhang, Leave-One-Out Bounds for Kernel Methods. Neural Comput 15 (2003) 1397-1437.

[52] M. Kanehisa, S. Goto, M. Hattori, K.F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa, From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34 (2006) 354-357.

[53] J. Demeter, C. Beauheim, J. Gollub, T. Hernandez-Boussard, H. Jin, D. Maier, J.C. Matese, M. Nitzberg, F. Wymore, Z.K. Zachariah, P.O. Brown, G. Sherlock and C.A. Ball, The Stanford Microarray Database: implementation of new analysis tools and open source release of software. Nucleic Acids Res 35 (2007) 766-770.

[54] V.G. Tusher, R. Tibshirani and G. Chu, Significance analysis of microarrays applied to the ionizing radiation response. PNAS 98 (2001) 5116-5121.

[55] D.C. Montgomery, Design and Analysis of Experiments. John Wiley & Sons, Cambridge, MA, 2006.

[56] Y. Hochberg, A.C. Tamhane, Multiple comparison procedures, John Wiley & Sons Inc., New York, 1987.

[57] S.G. Brodie and C. Deng, BRCA1-associated tumorigenesis: what have we learned from knockout mice? Trends Genet 17 (2001) 18-22.

[58] T.T. Paull, D. Cortez, B. Bowers, S.J. Elledge, and M. Gellert, From the Cover: Direct DNA binding by Brca1. PNAS 98 (2001) 6086-6091.

[59] J.M. Gudas, T. Li, H. Nguyen, D. Jensen, F.J. Ⅲ Rauscher and K.H. Cowan, Cell cycle regulation of BRCA1 messenger RNA in human breast epithelial cells. Cell Growth Differ 7 (1996) 717-723.

[60] Y. Liu, D.M. Virshup, R.L. White and L.C. Hsu, Regulation of BRCA1 phosphorylation by interaction with protein phosphatase 1alpha. Cancer Res 62 (2002) 6357-6361.

[61] B.B. Mandelbrot, How Long is the Coast of Britain: Statistical Self-similarity and Fractal Dimension, Science 155 (1967) 636-38.

[62] P. Fischer and W.R. Smith, Chaos, fractals, and dynamics. Marcel- Dekker, New York, 1985.

[63] P.J. Brockwell and R.A. Davis, Introduction to time series and forecasting.

Springer, New York, 1996.

[64] C. Bandt, J. Flachsmeyer and H. Haase, Topology, measures, and fractals. Akademie Verlag, Berlin, 1992.

[65] L. Arnold and V. Wihstutz, Lyapunov exponents : proceedings of a workshop held in Bremen, Springer-Verlag, Berlin, 1986.

[66] S. Ellner, D.W. Nychka, and A.R. Gallant, LENNS, a program to esitimate the dominant Lyapunov exponent of noisy nonlinear systems from time series data, Institute of Statistics Mimeo Series 2235 (BMA Series 39), North Carolina State University, 1992.

[67] M. Kanehisa, S. Goto, M. Hattori, K.F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa, From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34 (2006) 354-357.

[68] J.A. Flick, S.R. Hamilton, F.J. Rosales, and J.A. Perman, Methane excretion and experimental colonic carcinogenesis. Digestive Dis Sci 35 (1990) 221-224.

[69] E. Cosentini, I. Haberl, P. Pertschy, B. Teleky, R. Mallinger, G. Baumgartner, E. Wenzl and G. Hamilton, The differentiation inducers phenylacetate and phenylbutyrate modulate camptothecin sensitivity in colon carcinoma cells in vitro by intracellular acidification. Int J Oncol 19 (2001) 1069-74.

[70] K.L. Cai, G.B. Wang, L.J. Xiong , K.L. Cai, G.B. Wang and L.J. Xiong, Effects of carbon dioxide and nitrogen on adhesive growth and expressions of E-cadherin and VEGF of human colon cancer cell CCL-228, World J Gastroenterol 9 (2003) 1594-1597.

[71] M. Gerin, J. Siemiatycki, M. Desy and D. Krewski, Associations between several sites of cancer and occupational exposure to benzene, toluene, xylene

and styrene: results of a case control study in Montreal. Amer J Indust Med 34 (1998) 144-156.

[72] M. Lang and O. Pelkonen, Metabolism of xenobiotics and chemical carcinogenesis. IARC Sci Publ 148 (1999) 13-22.

[73] D.W. Nebert,. Role of genetics and drug metabolism in human cancer risk. Mut Res 247 (1991) 267-81.

[74] S. Veeriah, T. Kautenburger, N. Habermann, J. Sauer, H. Dietrich, F. Will, B. Louise and P. Zobel1, Apple Flavonoids Inhibit Growth of HT29 Human Colon Cancer Cells and Modulate Expression of Genes Involved in the Biotransformation of Xenobiotics. Mol Carcs 45 (2006) 164–174.

국 문 요 약

시스템 생물학적 접근법을 이용한 암 물질대사 및 종양탄압자 분석 모형 개발

Systems Biology란 생물 세포내에서 일어나는 물질대사, 유전자 조절과정, 신호전달 체계 과정을 시스템적 접근을 통해 모형화 하고 해석하는 것을 목적으로 한다. 시스템적 접근이란 상호작용으로 구성된 경로 구조 자료를 기반으로 하여 여러 연구 가설들을 해결하는 것을 의미한다. 따라서 세포내에서 나타나는 다양한 현상들에 대한 가설을 시스템적 접근으로 모형화 하고 추론할 수 있다. 본 연구에서는 물질대사 경로, 세포신호 경로, 질병 경로 등을 이용하여 암 발병 과정에서의 위험인자인 물질대사와 유전자들을 연구하였다. 체계적 연구를 위하여 경로를 구성하는 기본 요소인 합성물과 효소로 나누어 사례별로 연구를 하였다. 첫 번째 사례 연구에서는 안드로겐 에스트로겐 물질대사 경로에 대한 호르몬 프로필을 측정하여 갑상선 질환과 연관성이 존재하는지 연구하였다. 두 번째 사례 연구에서는 갑상선암 발병시 위험인자가 되는 호르몬들을 추정하고 암발병 과정에서의 위험인자 호르몬들의 패턴에 대해 연구하였다. 세 번째 사례 연구에서는 대장암 발병시 연관성이 존재하는 물질대사를 찾아내는 과정에 대해 연구하였다. 네 번째 사례 연구에서는 유방암 종양탄압자인 BRCA1 유전자가 세포 신호 경로에 있는 유전자들에 미치는 영향력의 정도와 순서에 대해 연구 하였다. 첫 번째 사례 연구 결과 갑상선 질환 의사 결정 시스템을 개발하였으며 안드로겐 에스트로겐 물질대사의 프로필을 이용하여 갑상선 환자들을 판별해 낼 수 있었다. 두 번째 사례 연구 결과 2-hydroxyestrone, 2-hydroxyestradiol, 2-methoxyestrone, 2-methoxyestradiol, 2-methoxyestradiol-3-methylether 등이 갑상선암의 위험인자 호르몬으로 추정되어졌

으며 암 발병시 증가 추세의 빈도 패턴을 나타내고 있음을 볼 수 있었다. 세 번째 사례 연구 결과 생체 이물 물질대사에 속한 Dichlorobenzene 물질대사와 Styrene 물질대사가 대장암과 연관성이 있다고 추정되어졌다. 네 번째 사례 연구 결과 종양 탄압자인 BRCA1 유전자는 DNA regulatory complex인 SWI/SNF에 속한 유전자와 checkpoint arrest에 속한 유전자, transcription factor에 속한 유전자에 따라서 영향력의 정도와 순서들 간에 차이가 존재함을 볼 수 있었다.

핵심되는 말 : 갑상선암, 대장암, 유방암, 안드로겐 에스트로겐 물질대사 경로, 생체이물 물질대사, 종양탄압자, 의사결정시스템, 리야프노프 지수