이 학 석 사 학 위 논 문

Drug response prediction model using a component based structural equation modeling method

구조방정식을 이용한

약물반응 예측 모형 설립 및 성능 비교

2017 년 8 월

서울대학교 대학원

협동과정 생물정보학과

김 성 태

Drug response prediction model using a component based structural equation modeling method

by

Sungtae Kim

A thesis
submitted in fulfillment of the requirement
for the degree of Master
in
Bioinformatics

Interdisciplinary Program in Bioinformatics
College of Natural Sciences
Seoul National University
Aug, 2017

Drug response prediction model using a component based structural equation modeling method

지도교수 박 태 성

이 논문을 이학석사 학위논문으로 제출함

2017 년 **8** 월

서울대학교 대학원

생물정보협동과정 생물정보학 전공

김 성 태

김성태의 이학석사 학위논문을 인준함

2017 년 **8** 월

위 원 장 원 성 호 (인)

부위원장 박 태 성 (인)

위 원 이 승 연 (인)

Abstract

Drug response prediction model using a component based structural equation modeling method

Sungtae Kim

Interdisciplinary Program in Bioinformatics

The Graduate School

Seoul National University

The liver is made up of many different types of cells. Mutations in those cells can be developed into several different forms of tumors as known as cancers. For this reason, it is hard to expect for a single type of liver cancer treatment to have a favorable prognosis for all cancer patients. If we can diagnose and classify the patients who are expected to have good responses to a single therapeutic drug, it will help to reduce the time on choosing appropriate therapeutic drug for each patient health efficiently. Therefore, nowadays, building decent prediction model became important. Up to date, several models such as linear/logistic regression (LR), support vector machine (SVM), random forest (RF) methods have been used for prediction. However, occasionally, these methods oversight the biological pathway

information with relations between metabolites, proteins, or DNAs. In this research, we selected possible biomarkers and constructed a drug called Sorafenib response prediction model for liver cancer patients using a component based structural equation model. Component based structural equation modeling method have been used in sciences, business, education and other fields. This method uses unobservable variables as known as latent variables and the structural equation model relationships between variables.

In our research, we applied this structural equation modeling method into biological structured information data. Currently, we have peptide level data with Multiple Reaction Monitoring (MRM) mass spectrometry. MRM is a highly sensitive and selective method for targeted quantitation of peptide abundances in complex biological samples. The advantage of our component based drug response prediction model is that it first merges peptide level data into protein level information which helps better biological interpretation later. Also, it uses alternating least squares algorithm and estimates both coefficients of peptides and proteins efficiently. It handles correlation between variables without constraint by a multiple testing problem. Using estimated peptide and protein coefficients, we have selected significant protein biomarkers with permutation test and constructed a Sorafenib response prediction model. Using drug response for liver cancer patients' MRM data, we composed a Sorafenib response prediction model for liver cancer and demonstrated that our prediction model successfully predicted a

drug response for liver cancer patients with high area under the curve (AUC) score.

Key words: Prediction Model, Component Based Structural Equation Modeling, Multiple Reaction Monitoring (MRM), Generalized Structured Component Analysis (GSCA)

Student number: 2015-20503

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Liver cancer, also known as hepatic cancer, is a cancer that originates in the liver. The primary liver cancer, known as hepatocellular carcinoma (HCC), is the most common form of liver cancer in adults and has different growth patterns [1,2]. It is the fifth most common form of cancer and the third leading cause of cancer death worldwide, accounting for more than 600,000 deaths each year [3,4]. However, as diverse treatment methods have been developed and utilized for treating hepatic cancer, it is still difficult for patients to have a favorable prognosis with any medicine. To improve cancer patients' status efficiently, we need to find an appropriate medicine for each patient. If we can diagnose and classify the patients with optimal therapeutic drugs depending on stage and growth patterns, it will reduce the time and the money spending on patients.

To diagnose and classify patients properly, building a decent prediction model be-came important. The classical ways of building cancer prediction model were based on linear/logistic regression, support vector machine or

random forest [7,8,9]. While these models are effective in prediction, they do not consider any structure or hidden information on the data which makes it difficult to derive more meaningful biological interpretations. Applying these in real data analysis, our Multiple Reaction Monitoring (MRM) data consist of 231 peptides from 124 proteins, each protein contains at least one single peptide to several peptides. When building prediction model, the classical methods only select the best peptides as variables which perform the best prediction performance. In reality, these classical procedures do not take any biological relationship between peptides and proteins into analysis.

In this research, we propose building of prediction model using component based structured equation modeling method which uses the peptide to protein biological structure. The advantage of component based structured equation modeling is that it generates latent variables. The latent variable is not observed variable but it is inferred from other observed variables. Using latent variables, we can collapse unstructured data into structured data. These latent variables help more feasible explanation on the results. In our case, multiple peptides can be merged into each belonging proteins which used as latent variables. By collapsing the peptides level data into proteins level data which reduce the dimension of data, we showed that the component based structured equation modeling method effectively covers a protein level analysis while taking all peptides into analysis simultaneously. On the contrary, often, the classical methods could not handle all peptides at once due to

multicollinearity. In real data analysis, we discovered possible protein biomarkers for a drug Sorafenib response. Using these protein biomarkers, we evaluated the performance of our drug response prediction model. Then, we compared performance of our prediction model with generalized linear model using logistic regression and logistic regression with ridge parameter by the area under the curve (AUC) score.

Chapter 2

Materials and Methods

2.1 Materials

Hepatocellular carcinoma (HCC) patient samples (n=115) were collected in Seoul National University Hospital from 2013 to 2015. When the patients were diagnosed as liver cancer, the patients were recommended to use a drug called Sorafenib as a treatment of liver cancer. Progress of liver cancer for each patient was diagnosed twice by a doctor. Doctor examined the first diagnosis of liver cancer when the patient entered the hospital. After six weeks from first diagnosis, doctors examined the progress of cancer status and classified the drug response status by Modified Response Evaluation Criteria in Solid Tumors (mRECIST) [11]. After the second examination, patients were divided into two groups: positive and negative drug responses. Positive drug response group consisted of patients with complete response (CR), partial response (PR), or stable disease (SD) with respect to mRECIST. CR and PR responses were diagnosed when the size of cancer reduced after

six weeks. Also, SD was diagnosed when the size of cancer did not decrease at the second visit. On the other hand, the negative drug response group consisted of patients with progress dis-ease (PD), where PD was diagnosed when the size of cancer increased after the first diagnosis.

Among all 115 patients (101 men and 14 women), 75 samples (64 men and 11 women) are grouped into the positive drug response group and 40 samples (37 men and 3 women) are grouped into the negative drug response group. For each patient, 231 peptides data were generated by multiple reaction monitoring (MRM) technique [10]. MRM is a highly sensitive and selective method for targeted quantitation of peptide abundances in complex biological samples. 231 peptides can be merged into 124 proteins. MRM technique measures the amount of peptides in the patient's blood. We used the log-transformed ratio of light peptide intensity to heavy peptide intensity. Light peptide intensity represents the amount of peptides from patient's blood measured by MRM technique, while heavy peptide intensity means the amount of artificially built peptides measured by MRM technique. Also, demographic information as age and sex are provided. The range of age varied from 34 to 84. For the gender, there were 101 male and 14 female samples.

2.2 Methods

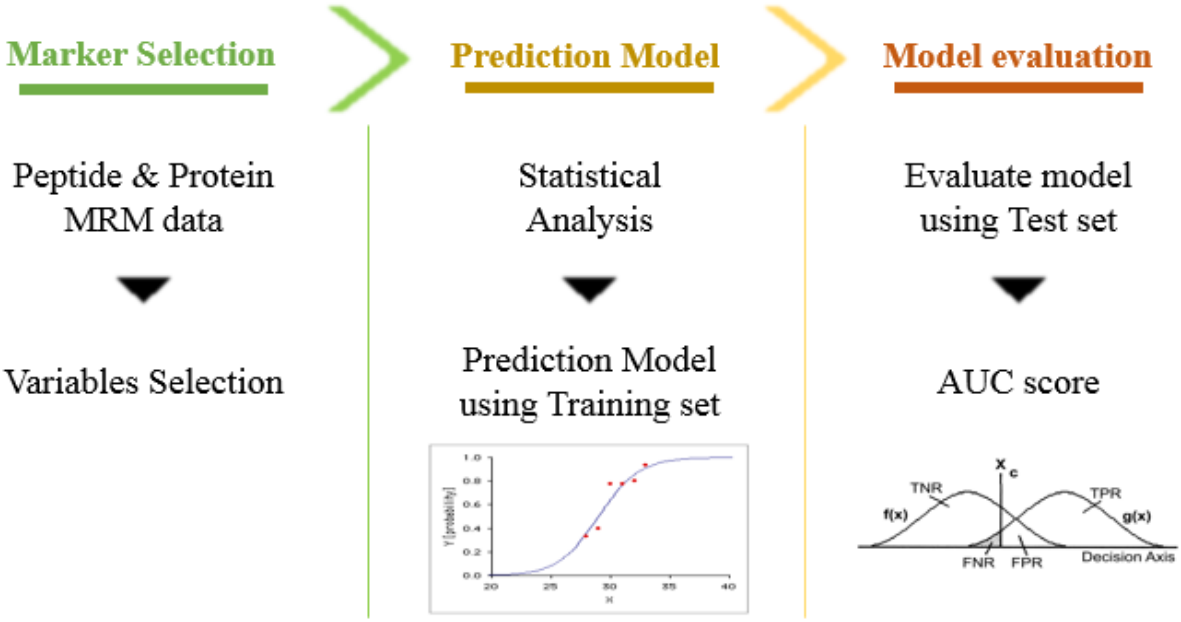

Figure 1. Schematic procedure of overall analysis

Overall schematic procedure for the research is described in Figure 1. At the beginning, we selected protein level biomarkers for drug Sorafenib response prediction model using MRM data. Second, we constructed prediction models via component based structural equation modeling method. At last, we evaluated the constructed drug response prediction model by AUC score.

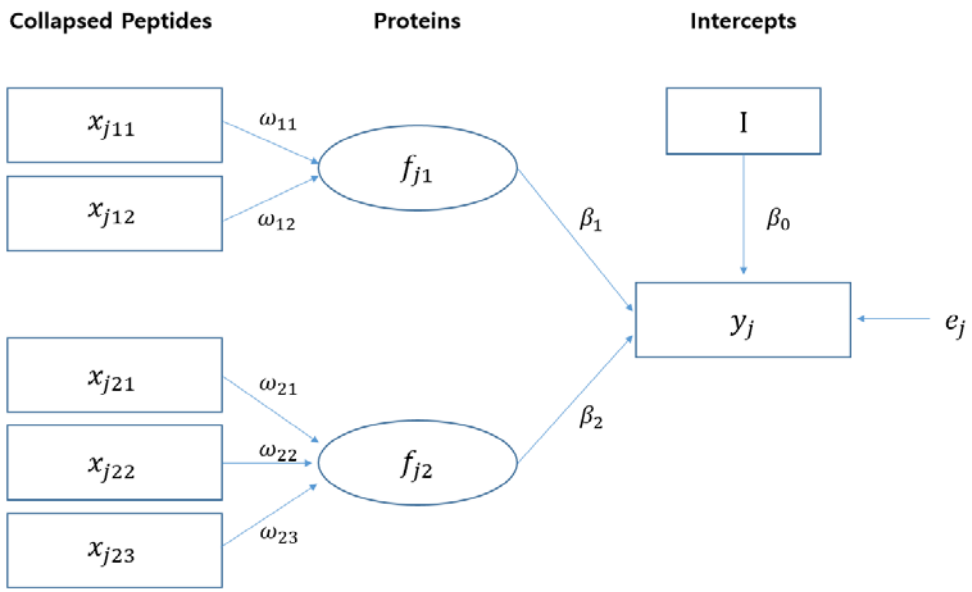

Figure 2. Example of proposed Generalized Structured Component Analysis (GSCA) model

A schematic example of our proposed drug response prediction model is described in Figure 2. This model merges peptide level MRM data into protein level information and estimates both peptides and proteins coefficients efficiently. In this example, two proteins are involved (K=2). Also, each protein consists of two or three peptides (T_k=2 and 3). Weight (w) and path coefficient (β) are estimated using the alternating least squares method [8].

Suppose that there are *K* proteins and the *k*th protein contains T_k peptides, for *k*=1,…,*K*. To estimate parameters, the following penalized log likelihood function was maximized. Let $w_k = [w_{k1}, \cdots, w_{kT_k}]'$, $\beta = [\beta_0, \beta_1, \cdots, \beta_K]'$, and F = $[f_1, \cdots, f_N]'$, where $f_i = [1, f_{j1}, \cdots, f_{jK}]$.

$$\varphi_1 = \sum_{j=1}^{N} \log P(y_j; \gamma_i, \delta) - \frac{1}{2}\lambda_{pep} \sum_{k=1}^{K} \sum_{t=1}^{T_k} w_{kt}^2 - \frac{1}{2}\lambda_{prot} \sum_{k=0}^{K} \beta_k^2 \quad \cdots\cdots\cdots (1)$$

Here, λ_{prot} and λ_{Pep} represent the tuning parameters: one for peptides in a protein and the other is for proteins. Maximizing the above equation via iteratively reweighted least squares is same as minimizing the following penalized least-square function:

$$\varphi_2 = \sum_{j=1}^{N} v_j (z_j - \sum_{k=0}^{K} f_{jk}\beta_k)^2 + \lambda_{pep} \sum_{k=1}^{K} \sum_{t=1}^{T} w_{kt}{}^2 + \lambda_{prot} \sum_{k=0}^{K} \beta_k^2 \quad \cdots\cdots\cdots (2)$$

with respect to w_t and β [3].

After estimating w_t and β coefficients, we constructed a drug response prediction model as follows. Also, x is standardized before making the prediction model.

$$log\left(\frac{\pi_j}{1 - \pi_j}\right) = \beta_0 + \sum_{k} \left(\sum_{i} x_{jki} w_{ki}\right) \beta_k + AGE_j \beta_{age} + SEX_j \beta_{sex} \quad \cdots\cdots\cdots (3)$$

$$= \beta_0 + \sum_{k} f_{jk}\beta_k + AGE_j \beta_{age} + SEX_j \beta_{sex} \quad \cdots\cdots\cdots (4)$$

j: individual samples (*j* = 1,⋯,115)

k: proteins (*k* = 1, ⋯,124)

i: peptides (*i* = 1, ⋯,231)

After our final drug response prediction model has been constructed, we evaluate our drug response prediction model performance by Area Under the receiver operating characteristic Curve (AUC) score using training, validation, and test sets. At last, our drug response prediction model will be compared with generalized linear model by binomial family (GLM), generalized linear regression with ridge by binomial family (GLMwR), Support Vector Machine (SVM), and Random Forest (RF) methods.

Chapter 3

Results

3.1 Biomarker discovery

To evaluate our model, at the beginning, we randomly selected 39 out of 115 samples as a separate test set which will be used as an external test dataset to evaluate overall performance of the final drug response prediction model. The other remaining 76 samples are randomly divided into training set (n=38) and validation set (n=38) for 2-fold cross validation (CV) to select significant protein variables (Figure.3). Using the 76 samples with 100 times of replicate, we selected significant proteins using p values. The significance of protein coefficients were determined by 1000 times of permutation test in each replicate. Through comparing the path coefficient value of the original data with those from the permuted data, the p-values were computed. After computation of p values for each protein, we selected significant proteins as p-values less than 0.05.

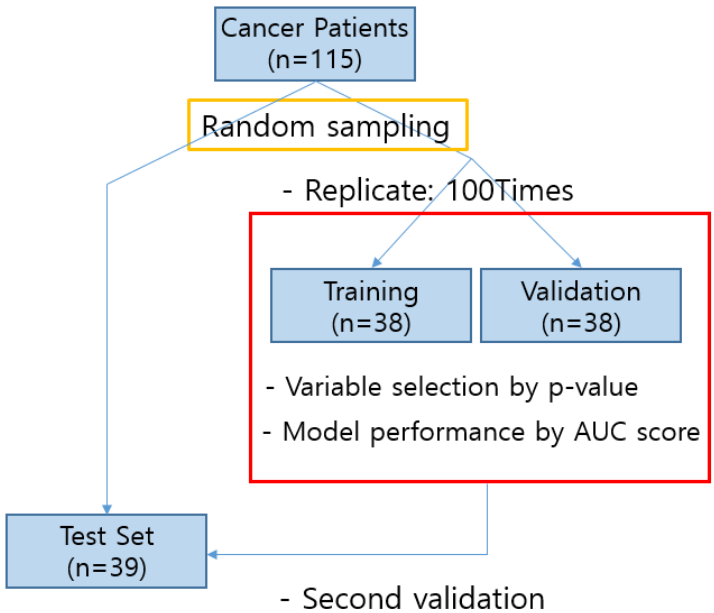

Figure 3. Sample separation for Training set, Validation set, and Test set

In the Fig. 4, the null distribution of path coefficients of each protein was calculated using repeated 1000 times by permutation test. Some of protein's path coefficient (beta) of came out as significant. Sample separation process for training and validation set was replicated 100 times. In each replicate, different number of proteins came out as significant.

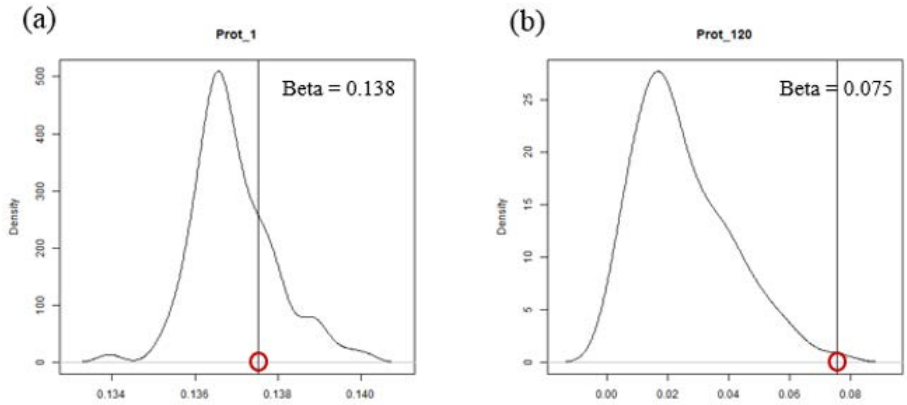

Figure 4. Estimation of path coefficient (beta) for each protein. (a) Not Significant Protein (b) Significant Protein

During the process on training/validation set with 76 samples, we repeated this process 100 times to check the consistency on selected significant proteins. Since our method uses two ridge parameters, we used same lambda values as 10 for computational efficiency. As a result, we selected top 6 significant proteins (APOC4, C163A, CD5L, IGJ, IC1, RET4) which were repeatedly selected as significant in the process of 100 replication. In other research, theses six proteins are found to be possible proteomic biomarker for hepatocellular carcinoma [13,14]. Also, to see consistency in the result, we used different lambda parameter and checked frequency of significant protein in same manner. Figure 5, 6, and 7 show selected protein frequency within circle. In this research, the optimal value of lambda is 10 in this research.

We repeat the process one more time only with those 6 protein MRM data for more accurate estimation of p value and path coefficients for the

prediction model. Then we calculate p values and path coefficients. In Table 1, p values for selected six proteins are shown: APOC4, C163A, CD5L, IGJ, IC1, RET4.

Table 1. P values for 6 proteins: APOC4, C163A, CD5L, IGJ, IC1, RET4

Protein	P value
APOC4	0.0061
C163A	0.0112
CD5L	0.0031
IC1	0.0102
IGJ	0.0142
RET4	0.0031

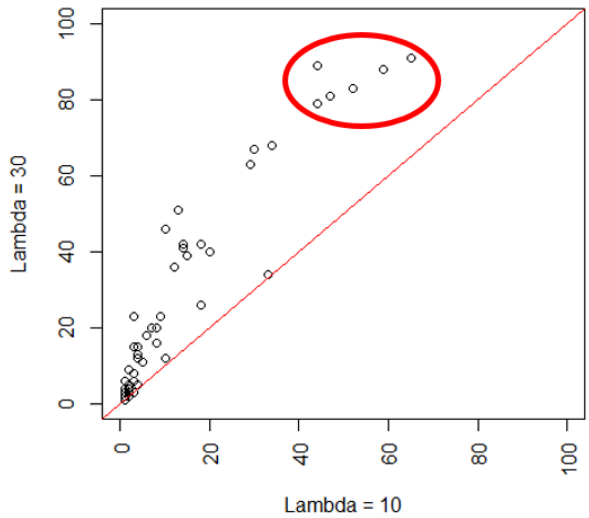

Figure 5. Plot of protein counts with lambda: 10 vs 30, Circled protein: CD5L, IGJ, APOC4, IC1, RET4, C163A (X axis: the frequency of each protein when lambda is 10, Y axis: the frequency of each protein when lambda is 30)

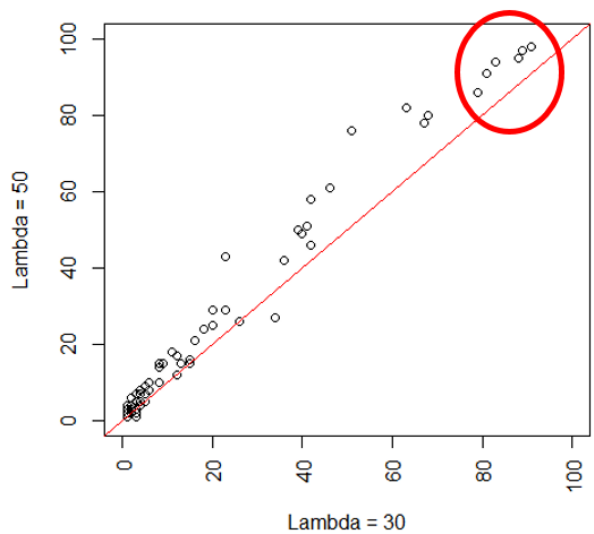

Figure 6. Plot of protein counts with lambda: 30 vs 50, Circled protein: CD5L, IGJ, APOC4, IC1, RET4, C163A (X axis: the frequency of each protein when lambda is 30, Y axis: the frequency of each protein when lambda is 50)

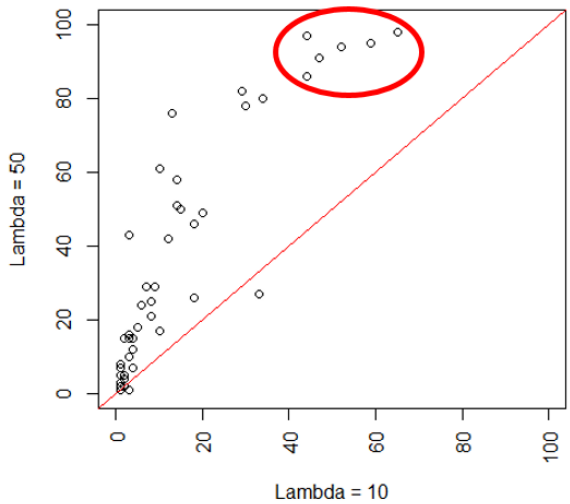

Figure 7. Plot of protein counts with lambda: 50 vs 10, Circled protein: CD5L, IGJ, APOC4, IC1, RET4, C163A (X axis: the frequency of each protein when lambda is 10, Y axis: the frequency of each protein when lambda is 50)

3.2 Model Evaluation by AUC score

With selected proteins, we first constructed drug response prediction model using generalized structured component analysis (GSCA) as single protein. The performance of the prediction models is measured by AUC scores. In the prediction model, age and sex were used as the covariates. Table 2 shows the AUC score of our single protein prediction model compared with GLM, GLMwR, SVM, and RF. The performance of single protein prediction model shows similar AUC scores for three statistical methods, while the AUC scores varied from 0.60 to 0.90 depending on the proteins.

Table 2. Comparison of AUC score on our drug response prediction model compared with GLM, GLMwR, SVM, and RF

Protein	GSCA	GLM	GLMwR	SVM	RF
APOC4	0.617	0.611	0.611	0.540	0.584
C163A	0.697	0.703	0.697	0.702	0.677
CD5L	0.860	0.883	0.897	0.883	0.886
IC1	0.837	0.857	0.846	0.857	0.833
IGJ	0.717	0.709	0.700	0.709	0.834
RET4	0.803	0.829	0.826	0.829	0.767

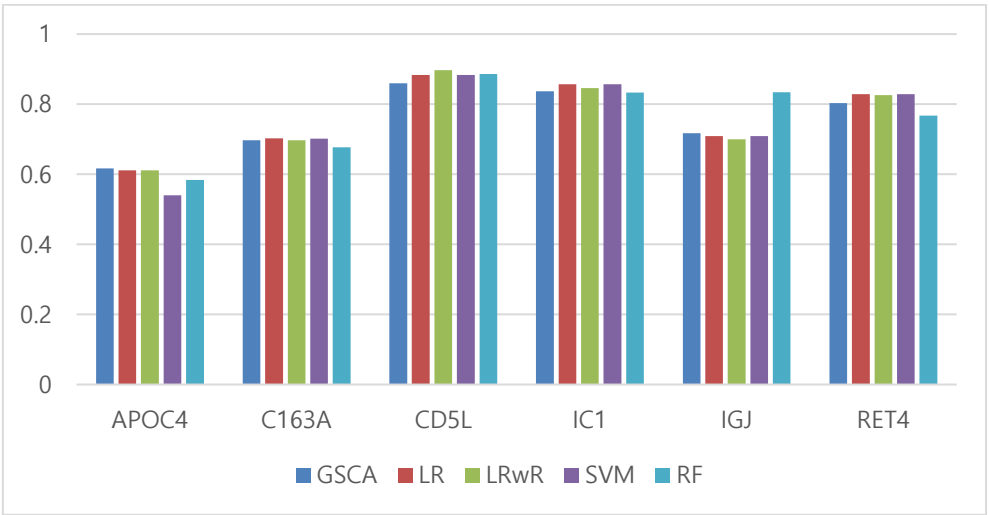

Figure 8. AUC values for the prediction models with single protein for each statistical method (GSCA, GLM, GLMwR, SVM, and RF)

The prediction model with two proteins had higher AUC scores compared to single protein models in all the three methods. Table 3 shows the AUC scores of the double proteins models. The AUC score varied from 0.73 to 0.95 which were higher than those of single protein prediction models.

The performance of GSCA models with two proteins was similar to those of GLM and GLMwR. It has higher performance or lower performance depending on the combination of proteins. The best performance of protein combination for all GSCA, GLM, GLMwR, SVM, and RF came out for the model with IC1 and IGJ proteins.

Table 3. AUC score of drug response prediction model with GSCA, GLM, GLMwR, SVM, and RF methods using combination of double proteins

Protein	GSCA	GLM	GLMwR	SVM	RF
APOC4 & C163A	0.851	0.837	0.837	0.791	0.790
APOC4 & CD5L	0.886	0.851	0.880	0.840	0.891
APOC4 & IC1	0.834	0.871	0.846	0.817	0.861
APOC4 & IGJ	0.886	0.814	0.897	0.914	0.971
APOC4 & Ret4	0.786	0.794	0.789	0.769	0.757
C163A & CD5L	0.866	0.883	0.897	0.849	0.891
C163A & IC1	0.894	0.917	0.914	0.866	0.901
C163A & IGJ	0.731	0.731	0.729	0.786	0.85
C163A & RET4	0.900	0.880	0.889	0.843	0.817
CD5L & IC1	0.923	0.923	0.934	0.929	0.959
CD5L & IGJ	0.854	0.931	0.886	0.92	0.931
CD5L & RET4	0.917	0.937	0.926	0.894	0.936
IC1 & IGJ	0.940	0.943	0.946	0.943	0.969
IC1 & RET4	0.871	0.891	0.897	0.874	0.839
IGJ & RET4	0.929	0.911	0.931	0.920	0.957

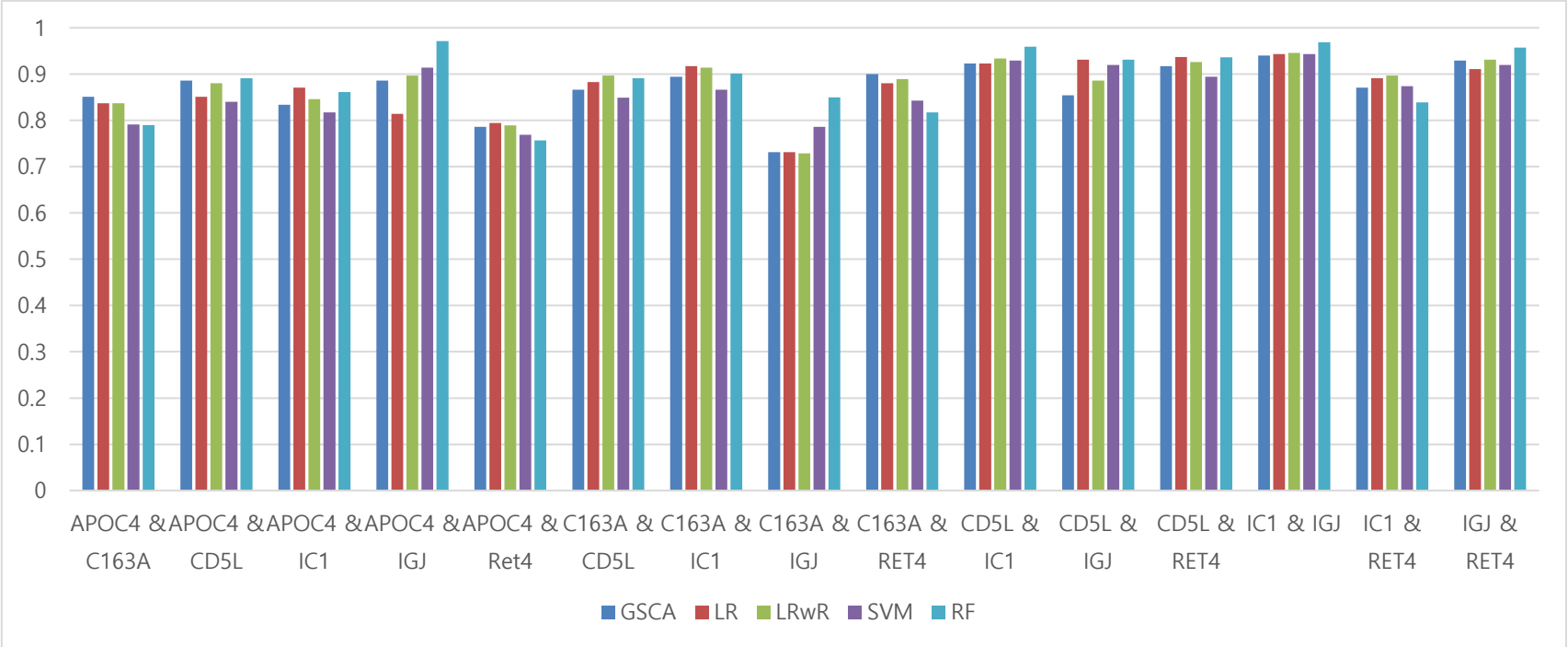

Figure 9. AUC values for the prediction models with two proteins for each statistical method (GSCA, GLM, GLMwR, SVM, and RF)

Using all 6 proteins, we also constructed GSCA drug response prediction model with estimated ω and β with covariates age and sex. In Figure 11, the AUC score by our GSCA model came out as 0.96 using the validation set. At first we tried to compare our prediction model with generalized linear model. However, it has a convergence problem due to high correlation among peptides, as shown in Figure 10. To solve this problem, we fit the logistic regression model with ridge penalty using “GLMNET” R Package. The convergence problem was resolved. The result is shown in Figure 11 and the AUC score of the generalized linear model with ridge parameter (GLMwR) model was 0.949 from the same validation set. As a result, our GSCA model had a slightly better AUC score compared to GLMwR, SVM, and RF.

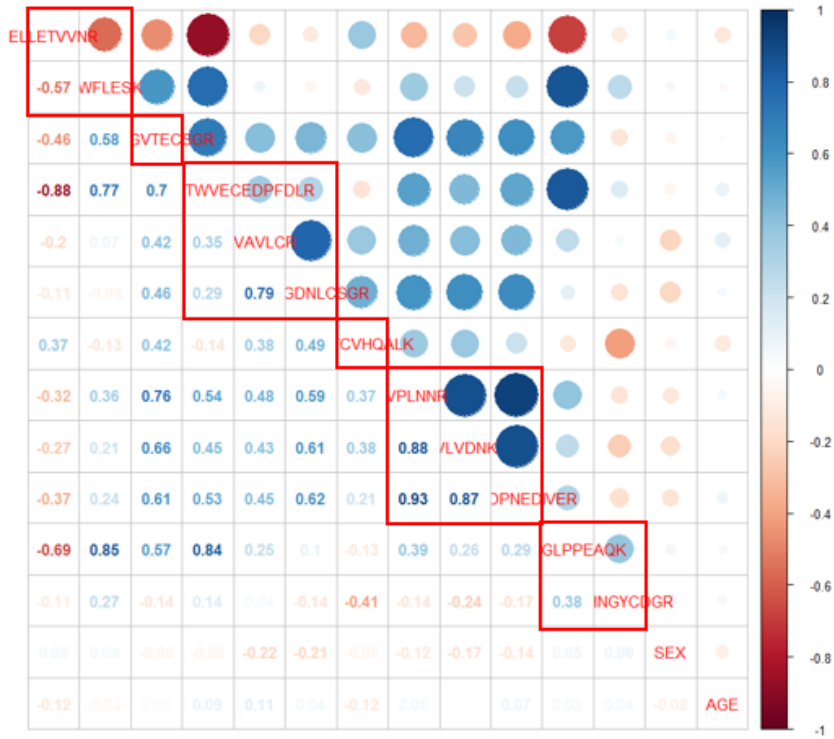

Figure 10. Correlations within peptides. Each red square box represents the peptides within same protein

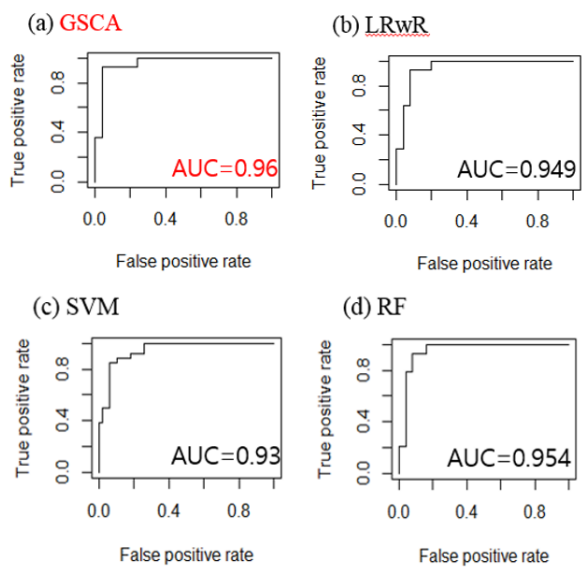

Figure 11. AUC Score of 6 proteins model (a) GSCA (b) GLMwR (c) SVM (d) RF

3.3 Simulation Study

For the Simulation model, we assume the true model contains RET4 (Significant protein with 2 peptides) and APOA1 (Nonsignificant protein, with 7 peptides), with parameters estimated by GSCA. The second simulation model is as following

$$log\left(\frac{\pi_j}{1-\pi_j}\right) = \beta_0 + \left(\sum_{i=1}^{2} x_{ji}\omega_i\right)\beta_{RET4} + \left(\sum_{i=3}^{9} x_{ji}\omega_i\right)\beta_{APOA1} + AGE_j\beta_{age} + SEX_j\beta_{sex} \quad \cdots\cdots\cdots (6)$$

From the estimated βs and ωs, derived from the MRM data, we estimated $\pi_1, \pi_2, \cdots, \pi_{115}$. Using theses estimated $\pi_1, \pi_2, \cdots, \pi_{115}$, the simulation responses were generated from the Bernoulli distribution. We constructed GSCA, GLM, GLMwR, SVM, and RF drug response prediction models. For each statistical methods, we measured the AUC score. We repeated the whole process 1000 times, and obtained 1000 AUC scores for each of the GSCA, GLM, GLMwR, SVM, and RF models. The boxplot of 1000 AUC scores are shown in Figure 12. In table4, we then calculated the mean of the 1000 AUC scores based on those models.

Table 4. Mean AUC scores of GSCA-based Simulation model

Methods	Mean AUC
GSCA	0.7270
GLM	0.6515
GLMwR	0.6812
SVM	0.6838
RF	0.6346

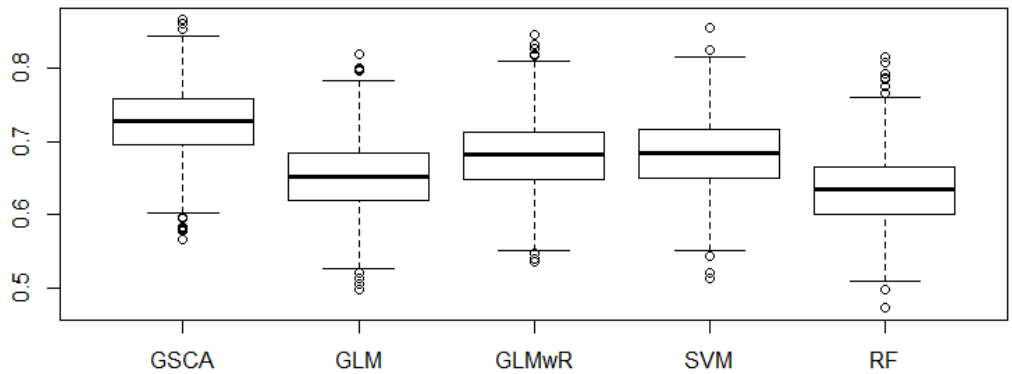

Figure 12. Box plots of ranges of 1000 AUC scores of GSCA-based simulation data

Chapter 4

Discussion

In this research, we have proposed a Sorafenib drug response prediction model for liver cancer patients using component based structural equation modeling method. We used generalized structured component analysis (GSCA) method to construct the drug Sorafenib response prediction model for Korean Hepatocellular carcinoma (HCC) patients with MRM proteomic data with some demographic variables. The advantage of GSCA is that it makes latent variables which are not directly observed variable but it is collapsed from other observed variables. With these latent variables, we can organize unstructured data into structured form. The latent variables help better explanation on the results. For example, our GSCA model collapsed several peptide MRM data into several proteins as latent variables. Unlike other classical methods as linear/logistic regression, support vector machine, and random forest, our approach using GSCA considers peptide to protein structure of peptide to protein biological structure. On the other hand, other classical prediction model methods does not consider the structure of

biological information as a component in the model. Using the peptide level data, we found significant proteins for possible biomarkers for building a Sorafenib response prediction model.

In the model building process, 1000 permutation tests were performed to find the significant peptides within each 100 replicates. Also, different ridge parameter value has been used to see the consistency of important protein. As a result, six significant proteins were selected: APOC4, C163A, CD5L, IGJ, IC1, RET4. All the 6 proteins were reported as possible bio markers related with cancers [13,14]. Of these 6 proteins, CD5L is well known for liver cancer biomarker [10,11]. With these six proteins, we successfully constructed a drug response prediction model. By comparing with generalized linear model with ridge penalization, the performance of our GSCA model with AUC score 0.96 showed a slightly better prediction result compared to generalized linear model with ridge parameter which had AUC score 0.949. Since GSCA and other statistical methods had high AUC score, it might be due to characteristic of MRM data. In further research, we can apply this overall prediction model building schematic process with GSCA method to other cancer data for constructing prediction model.

Bibliography

1. Asnacios A, Fartoux L, Romano O, et al. Gemcitabine plus oxaliplatin (GEMOX) combined with cetuximab in patients with progressive advanced stage hepatocellular carcinoma: Results of a multicenter phase 2 study. Cancer. 2008;112:2733–2739.
2. Fong,Y, Dupey,DE,, Feng M, Abou-Alfa G. Cancer of the liver. In: DeVita VT, Lawrence TS, Rosenberg SA, eds. DeVita, Hellman, and Rosenberg's Cancer: Principles and Practice of Oncology. 10th ed. Philadelphia, Pa: Lippincott Williams & Wilkins; 2015:696-714.
3. American Cancer Society. Cancer Facts & Figures 2016. Atlanta, Ga: American Cancer Society; 2016. American Joint Committee on Cancer.
4. Liver. In: AJCC Cancer Staging Manual. 7th ed. New York, NY: Springer; 2010:191–195.
5. Lee, S., Choi, S., Kim, Y. J., Kim, B. J., Hwang, H., Park, T., & T2d-Genes Consortium. (2016). Pathway-based approach using hierarchical components of collapsed rare variants. Bioinformatics, 32(17), i586-i594.
6. Jang, J. Y., Park, T., Lee, S., Kim, Y., Lee, S. Y., Kim, S. W., ... & Hirono, S. (2016). Proposed Nomogram Predicting the Individual Risk of Malignancy in the Patients With Branch Duct Type Intraductal Papillary Mucinous Neoplasms of the Pancreas. Annals of Surgery. Ma, Y., Ding, Z., Qian, Y., Shi, X., Castranova, V., Harner, E. J., & Guo, L. (2006). Predicting cancer drug response by proteomic profiling. Clinical cancer research, 12(15), 4583-4589.
7. Visser, H., le Cessie, S., Vos, K., Breedveld, F. C., & Hazes, J. M. (2002). How to diagnose rheumatoid arthritis early: a prediction model for persistent (erosive) arthritis. Arthritis & Rheumatism, 46(2), 357-365.
8. Spitz, M. R., Etzel, C. J., Dong, Q., Amos, C. I., Wei, Q., Wu, X., & Hong, W. K. (2008). An expanded risk prediction model for lung cancer. Cancer prevention research, 1(4), 250-254.
9. Huang, C. L., Liao, H. C., & Chen, M. C. (2008). Prediction model building and feature selection with support vector machines in breast cancer diagnosis. Expert Systems with Applications, 34(1), 578-587.
10. Chambers, A. G., Percy, A. J., Simon, R., & Borchers, C. H. (2014). MRM for the verification of cancer biomarker proteins: recent

applications to human plasma and serum. Expert review of proteomics, 11(2), 137-148.
11. Rabouhans, J. (2011, January). A radiologist's guide to the modified Response Evaluation Criteria in Solid Tumours (mRECIST) assessment of therapy for hepatocellular carcinoma. European Congress of Radiology 2011.
12. Hwang, H. (2009). Regularized generalized structured component analysis. Psychometrika, 74(3), 517-530.
13. Gray, J., Chattopadhyay, D., Beale, G. S., Patman, G. L., Miele, L., King, B. P., ... & Reeves, H. L. (2009). A proteomic strategy to identify novel serum biomarkers for liver cirrhosis and hepatocellular cancer in individuals with fatty liver disease. BMC cancer, 9(1), 271.
14. Braconi, C., Meng, F., Swenson, E., Khrapenko, L., Huang, N., & Patel, T. (2009). Candidate therapeutic agents for hepatocellular cancer can be identified from phenotype-associated gene expression signatures. Cancer, 115(16), 3738-3748.

초 록

간암은 간의 대부분을 구성하는 간 세포에서 시작되는 악성 종양을 얘기한다. 간에서 악성종양이 발병 시 간 주위를 구성하는 다른 세포와 함께 변이가 일어나 다양한 종류의 악성종양이 생성될 수 있다. 이러한 이유 때문에 간암환자에게 한 가지 약물치료법을 사용하였다 하더라도 모든 환자에게서 긍정적인 효과를 기대할 수는 없다. 의사들이 환자의 건강을 빠르고 효과적으로 치료하기 위해서 우리는 특정 약물치료법에 반응이 좋은 환자를 진단하고 구분하는 방법이 필요하다. 그로 인해 정확하고 사용하기 쉬운 약물 반응 예측모형을 만드는 것은 매우 중요하다. 현재까지 연구자들은 약물 반응 예측모형을 만드는 데에 선형/로지스틱 회귀분석 (Linear/Logistic Regression), 지지벡터기계 (Support Vector Machine), 랜덤 포레스트 (Random Forest) 등의 방법을 이용하여 예측 모형을 만들었다. 그러나 때때로 이러한 방법들은 생물학적 정보나 변수간의 상관관계를 간과하여 예측 모형을 만드는 경우가 생기게 된다.

본 논문에서는 우리는 일반화 구조 성분 분석 (Generalized Structured Component Analysis)를 통해 간암 환자들을 위한 소라페닙 (Sorafenib) 약물 반응 예측 모형을 만들고 이에 대한 성능평가를 하였다. 일반화 구조 성분 분석의 장점은 펩타이드 (Peptide) 레벨의 정보를 이용하여 단백질 레벨의 분석을 통해 생물학적 해석에 용의하고 예측 모형을 만들 시 펩타이드와 단백질들의 계수를 효과적으로 추정한다. 다중 반응 관찰 (Multiple Reaction Monitoring) 질량분석을 통해 측정한 소라페닙 약물반응 자료를 이용하여 우리는 간암환자들의 소라페닙 물반응성에 대한 모델을 만들었고 이 모델을 이용하여 간암

환자들의 소라페닙 약물에 대한 반응성 예측에 대한 성능평가를 성공적으로 하였다.

주요어: 예측모형, 다중 반응 관찰 (MRM) 질량분석, 구조방정식, 일반화 구조 성분 분석 (GSCA)

학 번: 2015-20503