Prediction of prostate cancer 5-year recurrence via machine learning

Hsiang-Han Chen

chen4646@umn.edu

Introduction

Prostate cancer is the second most common cancer in men. There are ~ 1,000,000 new cases were diagnosed every year [1, 2]. Prostate cancer has caused many people die and also the recurrence is not uncommon. In clinical practice, physicians applied several well-established indices to evaluate the outcome of the prostate cancer patient, such as pretreatment prostate-specific antigen (PSA) and Gleason scores. However, it is still difficult to have precise manual predictions of prostate cancer recurrence. Prediction systems/tools of prostate cancer recurrence based on data-driven method is required for clinical decision making. Typically, the performances of those prediction systems are superior to manual decisions [3]. Several studies of developing prediction system based on machine learning method were also proposed [4-6]. They show good prediction performance of cancer outcome.

In this project, the objection is to apply machine learning tools to precisely predict the five-year recurrence of prostate cancer. Three machine learning models, linear SVM, nonlinear SVM (RBF kernel), and KNN, were applied to make predictions. Two types of genetic data, gene expression, and somatic mutation were used as feature input of machine learning methods. The prediction performances of different combinations of machine learning models and features were reported for comparison.

Data Preparation

In this project, I used two types genetic data, gene expression and somatic mutation, of prostate cancer patients to predict cancer recurrence. Both of genetic data and the corresponding phenotype data (including recurrence record) were downloaded from TCGA database [1]. Based on the phenotype and recurrence record, the normal samples and samples with incomplete recurrence/last follow up record were excluded. Samples without matched genetic data and phenotype data were also excluded. The remaining samples should contain both the genetic raw data and the recurrence record (or the last follow up record for non-recurrence samples). The genetic data was used to be the feature input of machine learning methods.The recurrence/last follow up record was used to determine the label of the sample for the recurrence prediction (positive class = recurrence; negative class = non-recurrence). The Kaplan–Meier analysis based on recurrence/last follow up record (as shown in figure 1) shows that there are many samples censor during the first five years (~80%). Considering the label of five-year recurrence, I need to decide how to handle these censored samples. There are two simplistic ways to address this issue:

1. Treating the censored data as the negative class because there is no evidence showing that the sample has recurrence event during the first five years. The risk of this strategy is to involve some recurrence samples into non-recurrence class.
2. Discarding the censored data from the following experiments performed in this project. This method can ensure the correct labels of both recurrence class and non-recurrence class. However, the model selected for making prediction could be just the sub-optimal model.

In this project, I applied the first strategy to label censored samples as non-recurrence. Therefore, the number of samples used to perform the following experiments would be 494 (93 recurrences and 401 non-recurrences).

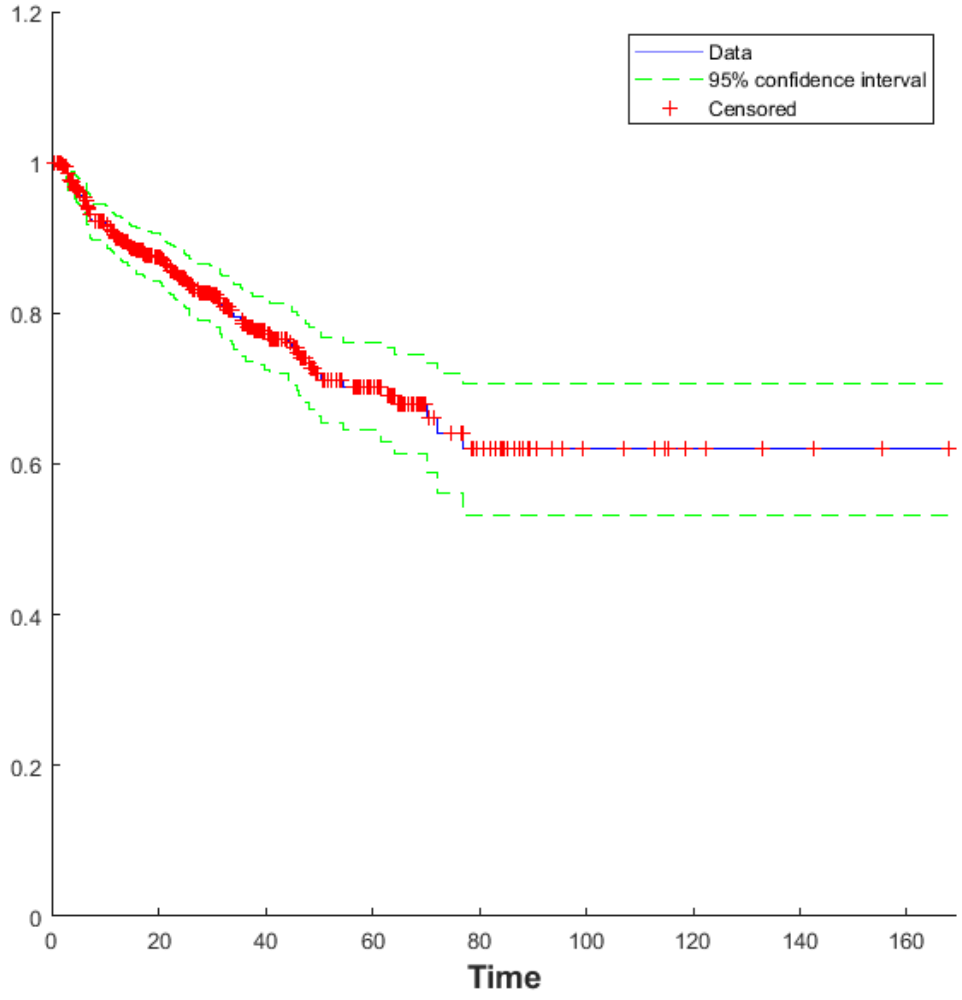

Figure 1. Kaplan–Meier plot for the recurrence records of prostate cancer samples from TCGA

Experiment Design

The procedure of feature extraction of the 494 prostate cancer samples is shown in figure 2. For both genetic data types, there are 93 recurrence samples and 401 non-recurrence samples which is a very unbalanced dataset for machine learning. Without applying an appropriate strategy, the model selection of machine learning was only based on the majority class. To handle this unbalance, the sample set (for prediction task) includes all recurrence samples and just a part of non-recurrence samples. The sample set contains 93 non-recurrence samples which were randomly selected from the all 401 non-recurrence samples. The order of the 93 recurrence samples is also permuted before the prediction task. The feature selection is based on the t-test between the 93 recurrence samples and the 401 non-recurrence samples. Several different numbers of features with the lowest p-value were selected as the feature input of machine learning.

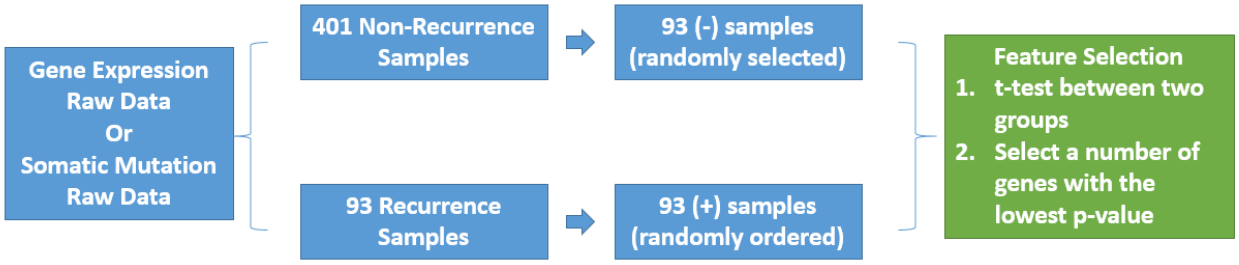

Figure 2. Feature extraction procedure

The experimental procedure is shown in figure 3. After a number of features have been selected, they would be feed into three different machine learning methods, linear SVM, non-linear SVM (RBF kernel), and KNN. Then through three-fold cross-validation to estimate the optimal model and also evaluate the prediction performance. In the three-fold cross-validation, the double resampling (as shown in figure 4) is used to perform model selection and prediction performance evaluation. The sample set was divided into three folds. When one of the three folds be the test set for evaluating the prediction performance, the other two folds would be the training set and the validation set which were used for model selection. The optimized model with the lowest validation error (i.e., the prediction error of validation set) would be used for recurrence prediction of the test set. This double resampling procedure would be repeated three times to make sure each sample has been predicted as a test sample.

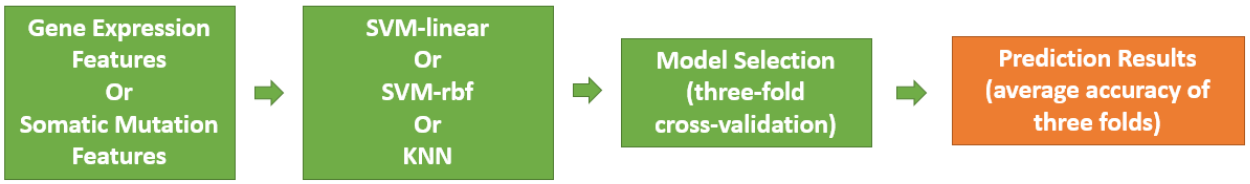

Figure 3. Experimental procedure

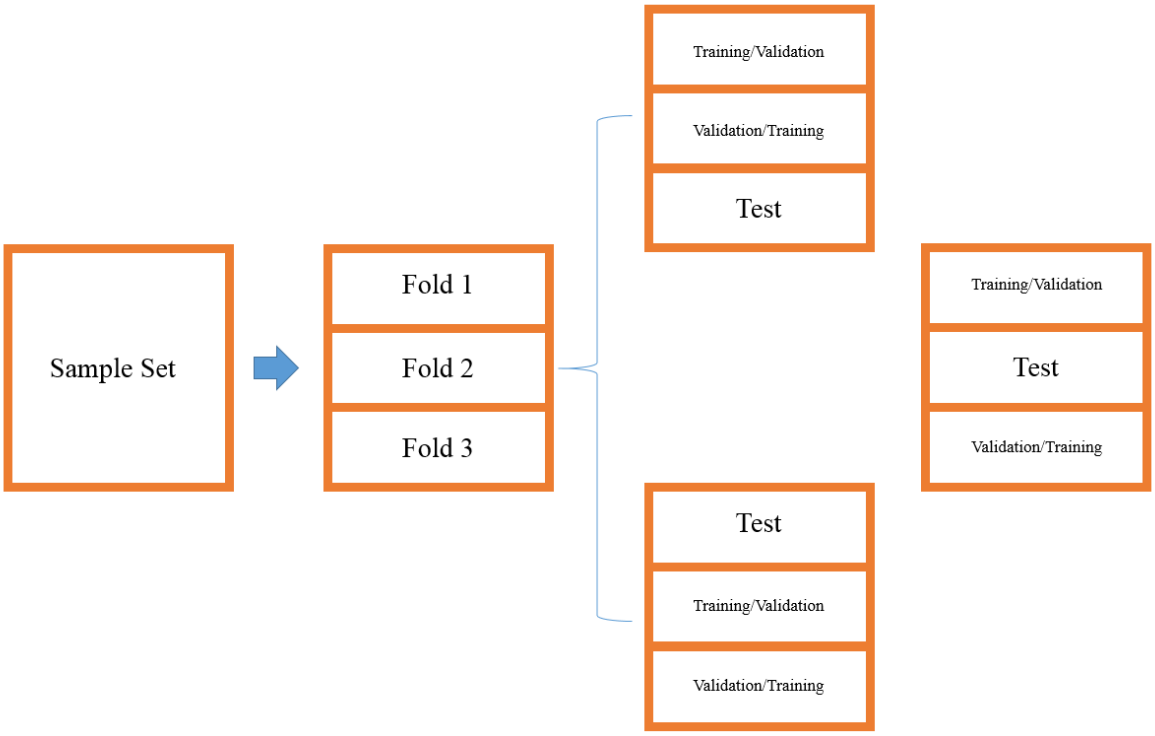

Figure 4. Double resampling

Expreimental Results

The results of experiments followed the procedure mentioned in the section of experimental design are presented in figure 5 and 6. Figure 5 and 6 show the results of recurrence prediction based on gene expression and somatic mutation respectively. For both genetic types, I apply three machine learning models (linear SVM, non-linear SVM, and KNN) to make recurrence prediction via different numbers of feature (5, 10, 20, 50, 100, 200, 500, 1000, and 2000) with the lowest p-value of statistical test between recurrence and non-recurrence samples.

For gene expression data, linear SVM shows the best performance (~70% accuracy) when

using 100 features (as shown in figure 5). The prediction accuracies of all three models decreased when using more than 100 features. It may be because too many irrelevant features are included in the model estimation. For somatic mutation data, non-linear SVM with RBF kernel shows the greatest performance (~95% accuracy) when using 1000 features (as shown in figure 6). This result is much better than the other two models. Especially for KNN method, the performance is inferior. No matter how many features are used in model estimation, the prediction accuracy is still about 50%. The reason for the poor performance of KNN method is due to the sparse binary features of somatic mutation data. KNN model cannot work well for sparse binary feature input because the Euclidean distance calculated based on samples with the sparse binary feature are most the same.

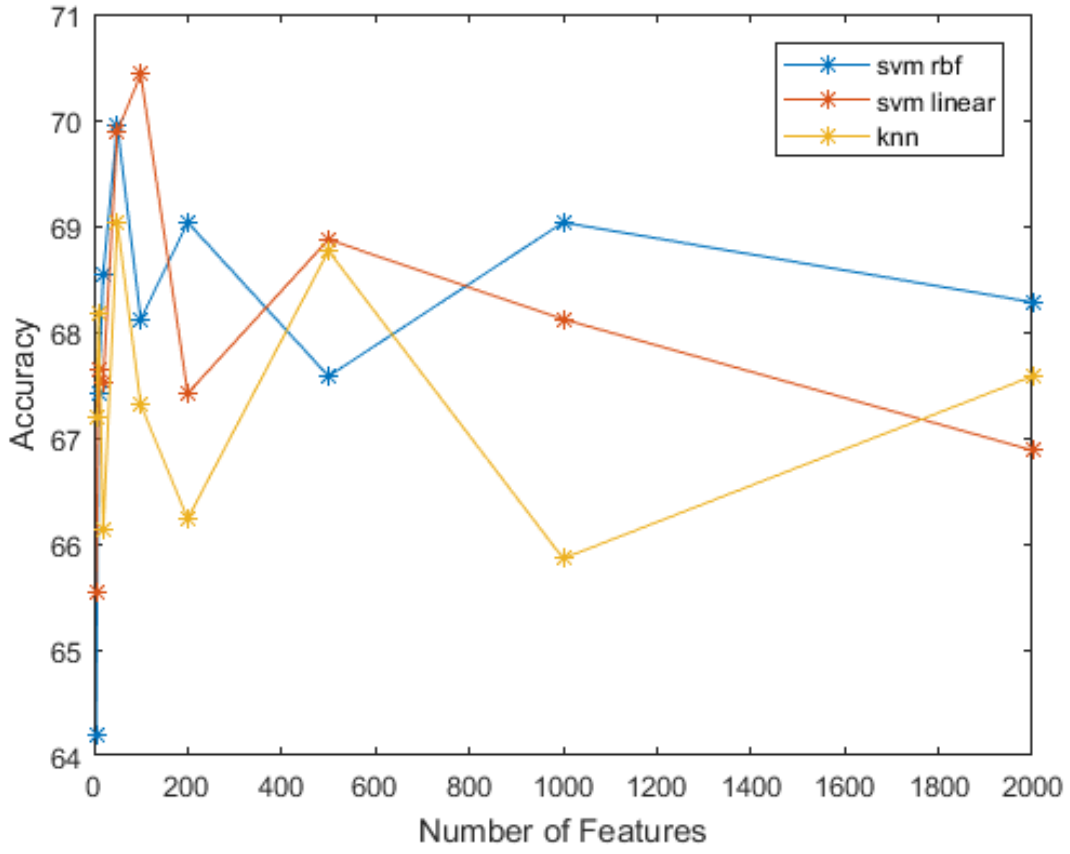

Figure 5. Prediction accuracies of three machine learning methods with different number of gene features of gene expression data

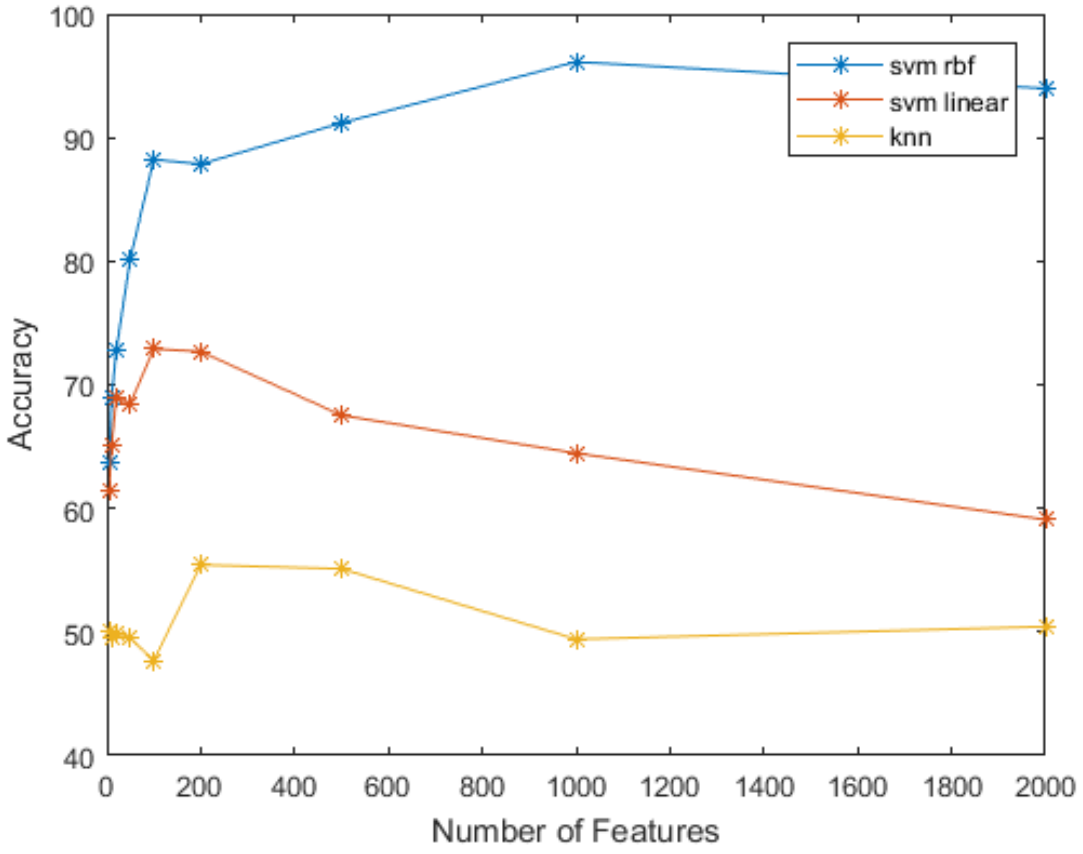

Figure 6. Prediction accuracies of three machine learning methods with different number of gene features of somatic mutation data

Gene Ontology Enrichment Analysis

According to the experimental results, using different numbers of features as input can affect the prediction performance. The linear SVM shows the best performance when using features of the top 100 genes selected from gene expression data and the non-linear SVM shows the best performance when using features of the top 1000 genes selected from somatic mutation data. These two sets of genes (100 genes selected from gene expression and 1000 genes selected from somatic mutation) may relate to some important biological mechanisms so the machine learning models can generate good prediction performance based on them.

To know the biological mechanisms related to these two sets of genes, I use the website tool [7] to perform Gene Ontology (GO) analysis for each of them. Table 1 and 2 show the GO analysis of the top 100 genes from gene expression data and the top 1000 genes from somatic mutation data respectively. For the top 100 genes selected from gene expression

data, there are several biological processes related to them such as chromosome segregation, cell division, and cell cycle [8-10]. These three processes are related to cancer development. For the top 1000 genes selected from somatic mutation data, the related biological processes are cell adhesion, biological regulation, and cellular process. These mechanisms are also important to cancer and related to the treatment of cancer [11-13]. From the GO analysis, the biological mechanisms related to these two sets of genes can be found. Those biological mechanisms (process) are indeed related to cancer development and treatment outcome.

Table 1. GO analysis using the top 100 genes of gene expression data

	Homo sapiens (REF)	upload_1 (▽ Hierarchy NEW!)					
GO biological process complete	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
regulation of chromosome segregation	96	6	.44	13.56	+	7.54E-06	1.06E-02
mitotic nuclear division	136	7	.63	11.17	+	4.32E-06	6.70E-03
↳nuclear division	275	11	1.27	8.68	+	8.15E-08	6.31E-04
↳organelle fission	305	11	1.41	7.82	+	2.22E-07	8.60E-04
↳mitotic cell cycle process	647	15	2.98	5.03	+	3.35E-07	1.04E-03
↳mitotic cell cycle	703	15	3.24	4.63	+	9.31E-07	1.80E-03
↳cell cycle	1355	21	6.25	3.36	+	8.79E-07	1.95E-03
↳cell cycle process	1023	20	4.72	4.24	+	4.45E-08	6.90E-04
nuclear chromosome segregation	259	10	1.19	8.38	+	4.48E-07	1.16E-03
↳chromosome segregation	302	11	1.39	7.90	+	2.02E-07	1.04E-03
cell division	488	12	2.25	5.33	+	3.10E-06	5.33E-03
negative regulation of cell cycle	542	12	2.50	4.80	+	8.77E-06	1.13E-02
↳regulation of cell cycle	1132	17	5.22	3.26	+	1.71E-05	2.04E-02
microtubule-based process	581	12	2.68	4.48	+	1.73E-05	1.91E-02

Table 2. GO analysis using the top 1000 genes of somatic mutation data

	Homo sapiens (REF)	upload_1 (▽ Hierarchy NEW!)					
GO biological process complete	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
cell adhesion	869	76	42.04	1.81	+	2.38E-06	6.13E-03
↳biological adhesion	874	77	42.28	1.82	+	1.67E-06	5.18E-03
biological regulation	11969	657	579.05	1.13	+	1.30E-06	5.02E-03
cellular process	14968	809	724.14	1.12	+	3.28E-09	1.69E-05
Unclassified	3676	108	177.84	.61	-	2.80E-09	2.17E-05

Discussions and Conclusions

In this project, I successfully applied machine learning technology to predict prostate cancer recurrence. Especially for the non-linear SVM with RBF kernel, it can achieve an outstanding prediction performance (accuracy is about 95%) when using somatic mutation data. For using gene expression data, linear SVM shows the best performance (accuracy is about 70%) among the three methods.

The results also suggest that the number of gene features is important to prediction performance. If the number of gene features is too small which cannot provide enough information for recurrence prediction. However, if the number of gene features is too large may include a lot of irrelevant information which can decrease the prediction accuracy. Through the experiments in this project, the optimal number of features for different genetic types and machine learning models can be found. For using gene expression data, the top 100 genes can provide the best performance. On the other hand, the top 1000 genes can provide the best prediction accuracy for using somatic mutation data.

The top 100 genes selected from gene expression and the top 1000 genes selected from somatic mutations can allow machine learning methods to generate good recurrence prediction based on them. It may suggest that they are important factors for the prostate cancer recurrence. The GO analysis can somehow support this argument. Because the biological processes related to these two gene sets such as cell cycle, cell division, chromosome segregation, cell adhesion and biological regulation are highly related to cancer development.

References

1. Cancer Genome Atlas Research Network. "The molecular taxonomy of primary prostate cancer." *Cell* 163.4 (2015): 1011-1025.
2. Mistry M, Parkin D M, Ahmad AS, et al. Cancer incidence in the United Kingdom: projections to the year 2030. Br J Cancer. 2011; 105: 1795-1803.
3. Kwak, Jin Tae, et al. "Improving prediction of prostate cancer recurrence using chemical imaging." *Scientific reports* 5 (2015).
4. Zupan, Blaž, et al. "Machine learning for survival analysis: a case study on recurrence of prostate cancer." *Artificial intelligence in medicine* 20.1 (2000): 59-75.
5. Win, Shoon Lei, et al. "Cancer recurrence prediction using machine learning." *International Journal of Computational Science and Information Technology (IJCSIT)* 6.1 (2014).
6. Zhang, Shengping, et al. "Improvement in prediction of prostate cancer prognosis with somatic mutational signatures." *Journal of Cancer* 8.16 (2017): 3261.
7. Ashburner, Michael, et al. "Gene Ontology: tool for the unification of biology." *Nature genetics* 25.1 (2000): 25-29.
8. Jallepalli, Prasad V., and Christoph Lengauer. "Chromosome segregation and cancer: cutting through the mystery." *Nature Reviews Cancer* 1.2 (2001): 109-117.
9. Preston-Martin, Susan, et al. "Increased cell division as a cause of human cancer." *Cancer research* 50.23 (1990): 7415-7421.
10. Hartwell, Leland H., and Michael B. Kastan. "Cell cycle control and cancer." *Science-AAAS-Weekly Paper Edition* 266.5192 (1994): 1821-1828.
11. Cavallaro, Ugo, and Gerhard Christofori. "Cell adhesion and signalling by cadherins and Ig-CAMs in cancer." *Nature Reviews Cancer* 4.2 (2004): 118-132.
12. Hayes, John D., and David J. Pulford. "The glut athione S-transferase supergene family: regulation of GST and the contribution of the lsoenzymes to cancer chemoprotection and drug resistance part II." *Critical reviews in biochemistry and molecular biology* 30.6 (1995): 521-600.
13. Kondo, Yasuko, et al. "The role of autophagy in cancer development and response to therapy." *Nature Reviews Cancer*5.9 (2005): 726-734.