Identifying Thyroid Carcinoma Subtypes and Outcomes through Gene Expression Data

Kun-Hsing Yu, Wei Wang, Chung-Yu Wang

Abstract:

Unlike most cancers, thyroid cancer has an ever-increasing incidence rate over recent years. In order to better understand its molecular mechanisms, we acquired gene expression data from The Cancer Genome Atlas (TCGA), leveraged supervised machine learning methods to predict stages and outcomes, and utilized unsupervised machine learning methods to gain biomedical insights on the gene activity patterns. Results showed that support vector machine with Gaussian kernel could distinguish patients with different survival outcomes with 81% accuracy, and factor analysis identified important biological processes in the tumor development. With continuing effort to improve classification accuracy, we envision personalizing patient treatments based on their predicted disease outcomes with larger sample size, thereby increasing the quality of care and reducing the cost of cancer management.

Introduction:

Thyroid cancer is one of the few cancers with increasing incidence rates over recent years. The American Cancer Society estimated that in 2013, there will be about 60,220 new cases of thyroid cancer in the U.S. and about 1,850 people will die from this disease[1, 2]. The rise in incidence rates cannot be completely accounted for by improved disease detection[1]. Understanding the genetic and molecular basis of the disease will help us identify its risk factors and point to possible explanations for the trend of increasing incidence rates.

The recent availability of public gene expression data sets has created great opportunities to study the gene expression changes underlying the development and progression of the disease. In addition, gene expression data may be used to identify novel subtypes of thyroid cancer. Papillary tumor is the most common type of thyroid cancer, which generally grows slowly; however, there exist several variants of thyroid cancer that tend to grow and spread more quickly[3]. Gene expression data can help us further refine the subtyping system, making it more relevant to treatment response and clinical outcomes.

By leveraging gene expression data of a large number of thyroid cancer patients from The Cancer Genome Atlas (TCGA) Research Network[4], we aim to (1) predict the stages and survival outcomes of papillary thyroid cancer patients and identify the most predictive genes for each classification, and (2) discover patterns of gene expression which can be signatures of unknown subtypes of papillary thyroid carcinoma with clinical relevance.

Materials and Methods:

Data Sources

We obtained gene expression and clinical data from 484 patients with thyroid cancer from TCGA, a collaborative project sponsored by the National Cancer Institute and the National Human Genome Research Institute. We used the expression levels of 20,531 genes measured by RNA-sequencing in the tumor tissue as the features for classification and clustering. The main outcome labels we examined were patient survival groups (with survival time longer than or shorter than 2,000 days) and tumor stages. The distributions of these two labels are shown in Table 1.

Table 1. Stages and Survival Outcomes of Thyroid Carcinoma Patients.

Stage at diagnosis	
Stage I	N = 275
Stage II	N = 52
Stage III	N =108
Stage IV	N = 49
Survival	
Alive	N = 471
Died	N = 13
Survived > 2,000 days	N = 37
Survived ≤ 2,000 days	N = 12

Preprocessing of Feature Data

We visualized the distributions of the gene expression levels through histograms, normalized them, and excluded genes with little or no variation in expression levels from subsequent analyses.

Classifying Clinical Outcomes using Supervised Machine Learning Techniques

We classified patients with naive Bayes (NB), K-nearest neighbor (KNN), support vector machine (SVM), decision tree (DT), and regularized logistic regression (RLR). We utilized MATLAB built-in functions for most of the classifiers and the liblinear package for SVM, We implemented the Newton's method for RLR. We evaluated the performance of our classification algorithms by hold out cross-validation (CV), 10 fold CV, and leave-one-out cross validation (LOOCV; for classification tasks with less than 50 samples). The performance of each classifier was compared to the majority rule (MR), which served as the baseline classifier. To optimize our classifiers, we explored different box constraint parameter C and kernels in SVM, distribution assumptions in NB, Ks and distance measures in KNN, tree depths in DT, and lambdas in RLR.

We performed feature selection by forward selection and Wilcoxon test statistics (comparing the features in outcome group 1 vs. outcome group 2). We conducted Gene Ontology (GO) analysis to identify the enriched biological processes of the selected genes.

Clustering Patients for Subtype Discovery using Unsupervised Learning

Due to the vast number of features, we used principal component analysis (PCA) to reduce dimensions and visualize the data, and examined if the first few principal components suggested distinct clusters that correspond to clinically relevant characteristics of the patients such as stages, survival outcomes, histological types, and races.

We explored novel tumor subtypes through hierarchical clustering. Gene expression levels were clustered based on Euclidean distances. As an external validation of the clustering, we investigated whether or not any of the gene expression-defined clusters correlated with known clinical characteristics.

We also employed factor analysis to explore the latent factors that explain the heterogeneity in gene expression among patients.

Results:

I. Supervised Learning

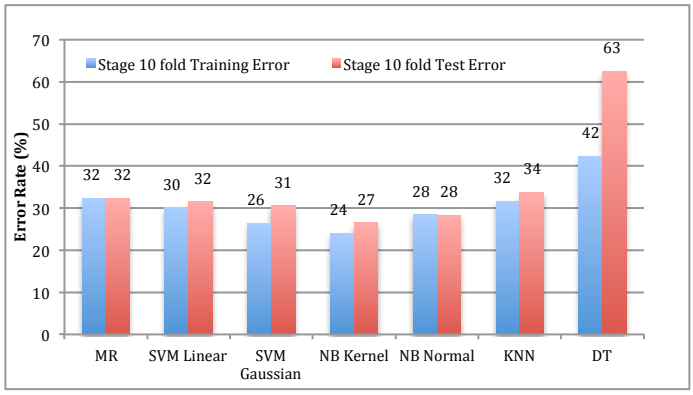
I-1. Classifying TNM Stages

We trained various classifiers to distinguish patients with different TNM stages (stage I to IV). Since multiclass classifications did not perform significantly better than our baseline classifier, we focused on binary class classification, with the first class being stages I and II and second class being stages III and IV.

Forward Feature Selection:

We selected the top 10 most effective features, since test errors plateaued after so, and demonstrated that using only the top 5 features generally had the best test performances. Training and test error of different classifiers are shown in Figure 1. Naïve Bayes classifier with 5 features achieves the best performance, with test error rate of 27%.

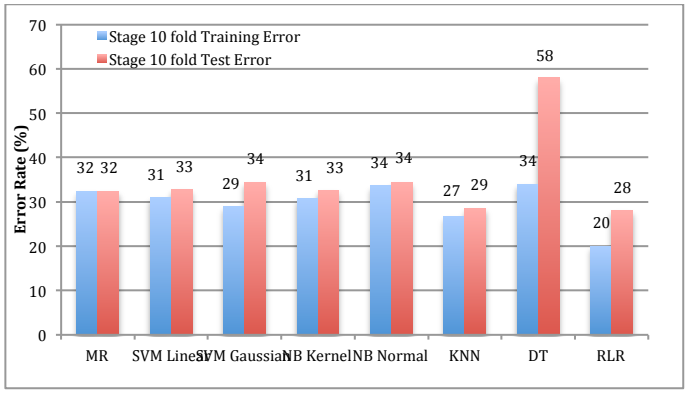
Figure 1. Training and Test Error of Binary Stage Classification with Top 5 Features from Forward Feature Selection.



Feature Selection by Wilcoxon Test Statistics:

We also selected features that showed the largest differences between the two outcome groups as measured by Wilcoxon test statistics. Figure 2 shows that the Wilcoxon test statistics offered comparable performances to forward feature selection.

Figure 2. Training and Test Error of Binary Stage Classification with Top 5 Features from Wilcoxon Test Statistics.



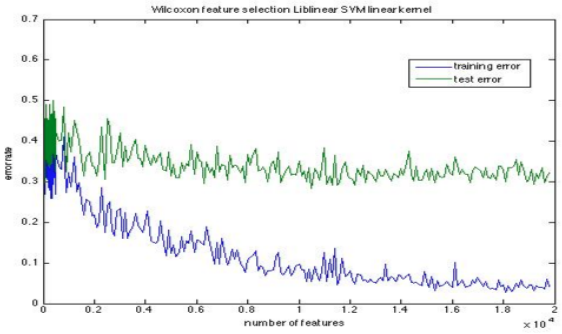
Test error rates didn't improve significantly when varying box constraint C in SVM, distribution

assumption in NB, K in KNN, or lambda in RLR. We only showed the classifiers with the best performances in the above figures.

Diagnostics:

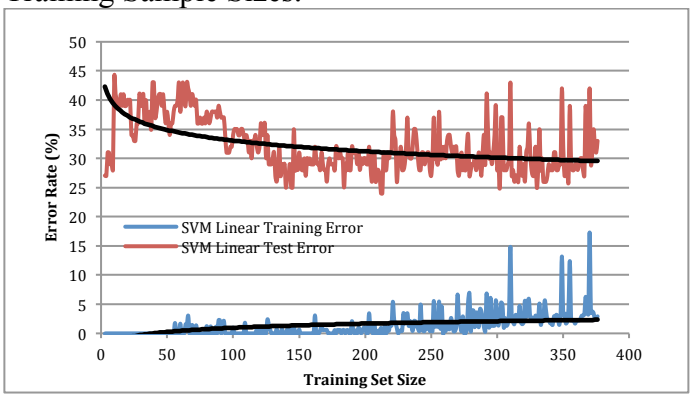
When using features selected by Wilcoxon test statistics, training error steadily decreased until feature number reached 10,000, while test error remained at a constant level when using 5,000 or more features, suggesting high variance with more than 5,000 features (Figure 3).

Figure 3. Error Rates of SVM with Linear Kernel with Varying Numbers of Features.



The plot of error rates with increasing training size (Figure 4) indicates that a larger training set could mitigate the high variance problem in our models.

Figure 4. Error Rates of SVM with Varying Training Sample Sizes.



The most informative genes in stage prediction selected by Wilcoxon test statistics are shown in Table 2. GO analysis revealed their association with protein kinase C (PKC) and vascular endothelial growth factor (VEGF) receptor signaling pathways.

Table 2. Top Genes in Stage Prediction Selected by Wilcoxon Test Statistics.

LAD1	FZD9	ST3GAL1	PVRL4	VEGFA
MCF2L	KIF5C	C1orf85	CCL17	SLC10A4
CERCAM	TMEM79	FLT4	DEPDC6	RETN
RANBP17	FAM178B	EPHA10	RBM20	PLP2

I-2. Classifying Survival Outcomes

We categorized the patients into two groups based on their survival outcomes: the ones who died within 2,000 days after diagnosis and the ones who survived at least 2,000 days. Observing that using all 20,351 features was computationally expensive and incurred overfitting, we applied forward feature selection (Figure 5) and feature selection according to Wilcoxon test statistics to select the most relevant features (Figure 6). Among all classifiers, SVM with Gaussian kernel on 10 features from Wilcoxon test statistics achieved the best performance, with 19% test error rate.

Figure 5. Training and Test Error of Survival Classification with Top 10 Features from Forward Feature Selection.

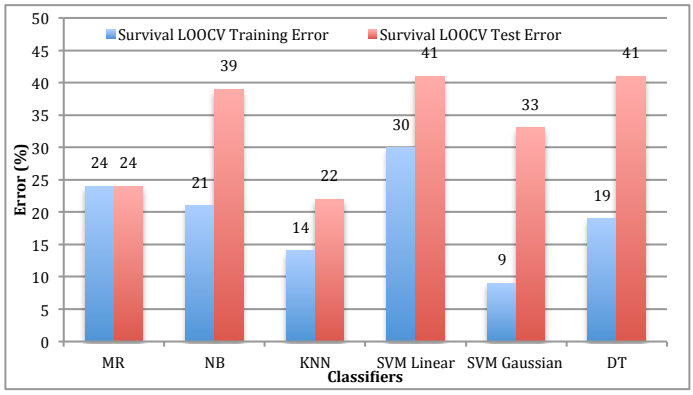
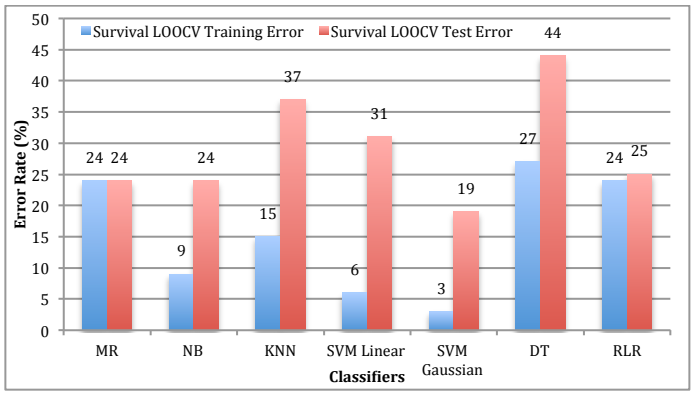


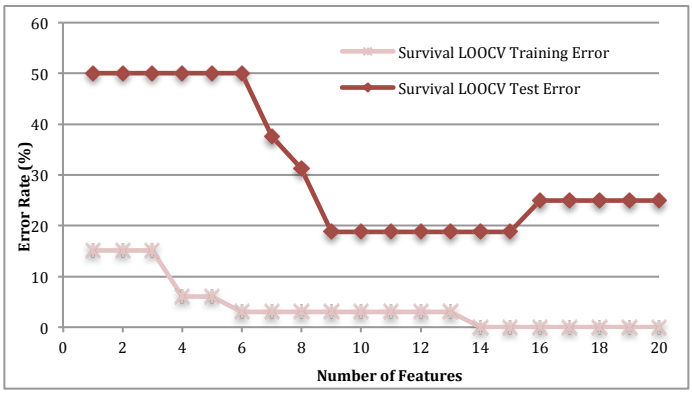
Figure 6. Training and Test Error of Survival Classification with Top 10 Features from Wilcoxon Test Statistics.



Diagnostics:

We investigated the optimal number of features for our classifiers. Results showed that underfitting occurred when using less than 9 features in SVM with Gaussian kernel, and overfitting was evident when using more than 16 features. Using 9 to 16 features attained the lowest test error for SVM with Gaussian kernel, our best performing classifier (Figure 7).

Figure 7. Classification Error Rates of SVM with Gaussian Kernel for Survival with Varying Numbers of Features.



Top features for survival prediction are shown in Table 3. GO analysis revealed their association with immune response and intercellular adhesion molecules.

Table 3. Top Genes in Survival Prediction Selected by Wilcoxon Test Statistics.

TERT	SLC7A3	TRIM21	IQSEC1	RAP1GDS1
CAMP	CHKB	NXF1	SCFD2	FLJ11235
MAPK9	CLEC4M	SNORA40	PSME1	TUBB2C

When varying box constraint C in SVM from 0.0002 to 100 or varying lambda in RLR from 0.001 to 200, test error rates didn't improve significantly. Test accuracy for KNN also remained at similar levels for different Ks and when using different distance measures such as Euclidean, Manhattan, Chebychev, Minkowski, cosine, or correlation.

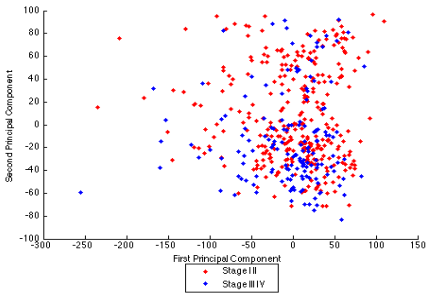
II. Unsupervised Learning

PCA on Whole Genome Expression Profiles

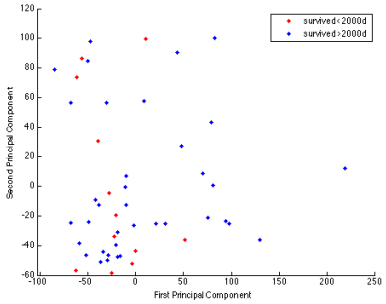
We plotted the first two principal components of the features and examined if subjects with different stages (Figure 8a) or survival outcomes (Figure 8b) form separate clusters on the plots. Although patients with different stages and survival outcomes showed significant differences in the median values of the first two principal components (Wilcoxon p-values in the range of 3.6×10^{-9} to 0.01), no distinct clusters were formed according to these two labels, which suggested that the greatest inter-individual variation in overall gene expression levels could not be readily explained by the difference in tumor stages or survival outcomes. Similar results were observed for histologic types and races (data not shown).

Figure 8. PCA on Whole Genome Expression Profiles. (a) PCA on Tumor Stages (b) PCA on Survival Outcomes.

(a)



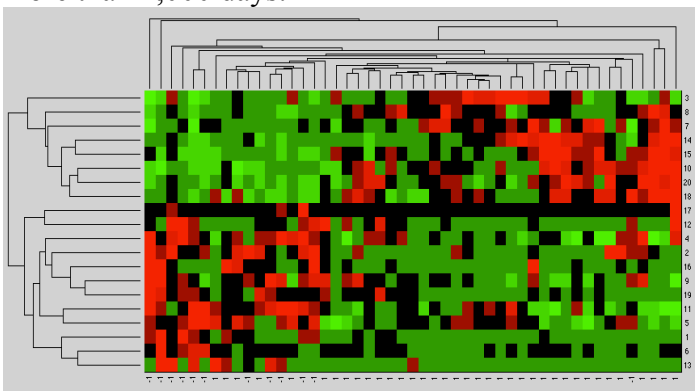
(b)



Hierarchical Clustering

To understand the clusters formed by hierarchical clustering, we examined if patients with different stages, histologic subtypes, or survival outcomes (the three most important clinical characteristics) tend to form distinct clusters with unique gene expression patterns. Results showed that hierarchical clustering didn't suggest subtypes related to survival outcomes, stages, or histologic subtypes. Although clustering was performed for the whole gene set, for illustration purposes, we present in Figure 9 the clustering results of 20 genes most predictive of survival.

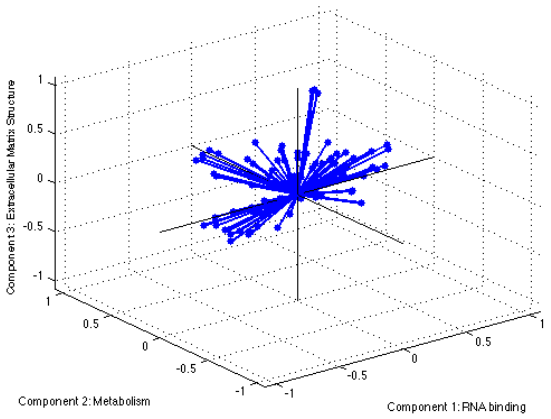
Figure 9. Hierarchical Clustering by Patients (columns) and Genes (rows). '-1' represents survival time less than 2,000 days and '1' represents survival more than 2,000 days.



Factor Analysis

Factor analysis revealed three major factors that best explained the inter-individual differences in gene expression. Factor loadings of genes with the most expression variability are shown in Figure 10. Gene Ontology (GO) analysis showed significant enrichment in RNA binding in the first factor, metabolism in the second, and extracellular matrix structure in the third factor.

Figure 10. Factor Loadings of Genes with the Most Expression Variability. Based on GO analysis, the three factors are related to RNA binding, metabolism, and extracellular matrix structure respectively.



Discussion

Our work demonstrated that supervised machine learning techniques could predict survival outcomes of thyroid cancer patients with 81% accuracy, while unsupervised methods such as PCA and factor analysis could help reduce dimensions. However, clustering on the whole genome expression profiles was not directly related to known clinical characteristics.

Our best classifiers could distinguish patients with different survival with 10 to 15 genomic features. We observed trends of overfitting when adding more features, and trends of underfitting when using fewer features. We further noticed that most of the informative genes for survival prediction were related to immune response and intercellular adhesion molecules. Our results were consistent with clinical observations that poor host immunity and extensive tumor invasion are generally poor prognostic factors of cancers[5]. The results suggest that we could further develop prognostic markers of thyroid cancer based on the expression levels of the selected genes.

Genomic information also appeared to have some predictive value for stages. The informative genes for stages were enriched in PKC and VEGF pathways, both of which are important pathways in tumor progression. However, we could only distinguish early stages thyroid cancer from the late stages with 73% accuracy, which is not much higher than the baseline accuracy. One possible explanation is that genomic profiles might not change significantly as tumor progresses in stages. Based on our diagnostics, more training samples are needed for better classification performance.

We identified the major contributors of the genomic differences among thyroid cancer patients through factor analysis. By investigating the gene components of the latent factors, we found RNA binding, metabolism, and extracellular matrix structure were the themes of these factors. Interestingly, these factors are all closely related to tumorigenesis or tumor progression[6], which provide us insights on the biological processes underlying the heterogeneity among thyroid cancer patients.

In summary, through machine learning methods, we are able to classify patients with different survival outcomes, identify the genes related with prognosis, and characterize the genomic signatures of thyroid cancer. With a larger number of samples, we could achieve better prediction accuracy. Eventually, we hope to personalize treatment plans based on their predicted disease outcomes, thereby improving the quality of care and reducing the cost of cancer management.

Reference

1. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2013. CA Cancer J Clin. 2013 Jan;63(1):11-30.
2. American Cancer Society. (Jan. 17, 2013). What are the key statistics about thyroid cancer? Retrieved Dec. 12, 2013, from http://www.cancer.org/cancer/thyroidcancer/detailedguide/thyroid-cancer-key-statistics
3. Chrisoulidou A, Boudina M, Tzemailas M, Doumala E, Pashalia KI, Patakiouta E, Pazaitou-Panayiotou K. Histological subtype is the most important determinant of survival in metastatic papillary thyroid cancer. Thyroid Research 2011, 4:12-16.
4. The Cancer Genome Atlas. (2013). Retrieved Dec. 12, 2013, from http://cancergenome.nih.gov/.
5. Dunn GP, Bruce AT, Ikeda H, Old LJ, Schreiber RD. Cancer immunoediting: from immunosurveillance to tumor escape. Nat Immunol. 2002 Nov;3(11):991-8.
6. Croce CM. Oncogenes and cancer. N Engl J Med. 2008 Jan 31;358(5):502-11.