International Journal of Nanomedicine

Dovepress
open access to scientific and medical research

Open Access Full Text Article

METHODOLOGY

Feature selection and survival modeling in The Cancer Genome Atlas

This article was published in the following Dove Press journal:
International Journal of Nanomedicine
12 September 2013
Number of times this article has been viewed

Hyunsoo Kim[1]
Markus Bredel[2]

[1]Department of Pathology, The University of Alabama at Birmingham, Birmingham, AL, USA; [2]Department of Radiation Oncology, and Comprehensive Cancer Center, The University of Alabama at Birmingham, Birmingham, AL, USA

Purpose: Personalized medicine is predicated on the concept of identifying subgroups of a common disease for better treatment. Identifying biomarkers that predict disease subtypes has been a major focus of biomedical science. In the era of genome-wide profiling, there is controversy as to the optimal number of genes as an input of a feature selection algorithm for survival modeling.

Patients and methods: The expression profiles and outcomes of 544 patients were retrieved from The Cancer Genome Atlas. We compared four different survival prediction methods: (1) 1-nearest neighbor (1-NN) survival prediction method; (2) random patient selection method and a Cox-based regression method with nested cross-validation; (3) least absolute shrinkage and selection operator (LASSO) optimization using whole-genome gene expression profiles; or (4) gene expression profiles of cancer pathway genes.

Results: The 1-NN method performed better than the random patient selection method in terms of survival predictions, although it does not include a feature selection step. The Cox-based regression method with LASSO optimization using whole-genome gene expression data demonstrated higher survival prediction power than the 1-NN method, but was outperformed by the same method when using gene expression profiles of cancer pathway genes alone.

Conclusion: The 1-NN survival prediction method may require more patients for better performance, even when omitting censored data. Using preexisting biological knowledge for survival prediction is reasonable as a means to understand the biological system of a cancer, unless the analysis goal is to identify completely unknown genes relevant to cancer biology.

Keywords: brain, feature selection, glioblastoma, personalized medicine, survival modeling, TCGA

Introduction

We expect that next generation sequencing technology keeps evolving, and that the cost of sequencing will drop to a practically affordable range.[1] It may, therefore, soon be feasible to obtain whole genome gene expression profiles of individual patients from whole transcriptome shotgun sequencing (also called RNA-Seq). A critical question is whether the availability of high-content information from this new technology will be clinically useful; for example, can it help predict survival of an individual patient and personalize treatment? Contemporary approaches for survival prediction often use a few number of genes that were identified as biomarkers from intensive scientific studies with whole gene expression profiles and/or other molecular measurements. Protein interaction networks in combination with gene expression data have been used to identify biomarkers associated with cancer metastases.[2] Another approach

Correspondence: Hyunsoo Kim
Department of Pathology, The University of Alabama at Birmingham, West Pavilion P220, 619 South 19th Street, Birmingham, AL 35249-7331, USA
Tel +1 205 975 9377
Fax +1 205 934 5499
Email hyunsookim@uab.edu

is to identify subcategories of a cancer and the associated biomarkers for each category, so as to allow treating a patient based on the cosegregation of her/his cancer profile within one cancer subcategory.

Recently, subgroup-specific biomarker networks have been shown to predict glioblastoma prognosis.[3] However, what if a patient's cancer is an example of a rare case that was not identified as a major tumor subcategory/group? Such rare cases tend not to be identified within a unique group, mostly because previous studies did not consider large enough numbers of patients; yet a new patient's genomic profile may be very similar to a few patients' genomic profiles. In such cases, database pattern match, which attempts to fit an individual genomic profile to previously characterized profiles and related outcomes, might be more useful than using a biomarker-based approach since the biomarkers were chosen only to discriminate known subcategories. In this paper, we call this approach devoid of group identification a "nongroup approach." The nongroup approach, which is based on pattern match or regression instead of classification or clustering for biomarker selection, uses a large number of multitype features for pattern matching in order to identify previous cases with high genomic similarity to a particular patient.

While we acknowledge the strength of the cancer subcategory-based approach, in this study, we investigated the feasibilities of the nongroup approach for predicting survival based on machine learning of whole genome gene expression profiles or cancer pathway gene expression profiles. We present some interesting genes identified through the analyses of whole-genome gene expression profiles and cancer pathway gene expression profiles, and we will make the point that using a set of genes selected by preexisting biological knowledge might be better as an input of a feature selection algorithm for survival modeling.

Material and methods

TCGA glioblastoma gene expression data

Although it is generally accepted that next generation sequencing can produce more accurate data with higher sensitivity, we decided to use available microarray data in our studies because of the availability of a larger number of samples, which were downloaded from The Cancer Genome Atlas (TCGA) data portal (https://tcga-data.nci.nih.gov/tcga/). A total of 560 gene expression profiles were retrieved from the Broad Institute HT_HG-U133A platform (Affymetrix, Santa Clara, CA, USA). The total number of unique patients was 544. Each gene expression profile had gene expression data for 12,042 genes. Normal and control samples were excluded. All genes had expression data available across all samples. Samples that did not have actual gene expression values were excluded, as were samples when the corresponding patient did not have survival information. After these filtering steps, 538 tumor samples were used in downstream analyses.

Survival modeling

One can consider a classification problem that can separate a shorter survival group and a longer survival group; this approach is well established. Refined classification accuracy can be obtained from feature selection, so it is more relevant to a cancer category-based approach where it is necessary to define subcategories (classes) before this classification process. However, it is possible that the gene expression profile of a patient does not share similarities with any of the gene expression profiles of predefined disease subcategories. Our interest in this study focuses on this special case, for which we studied the following survival modeling methods.

Nearest neighbor survival prediction method

One plausible method for predicting survival with whole-genome gene expression data is the 1-nearest neighbor (1-NN) approach. Once a gene expression profile of a patient A has been established, another patient B's gene expression profile that is most similar to patient A will be identified, and patient A's survival will be predicted as the (known) survival of patient B. The 1-NN approach is based on pattern matching with Pearson's correlation coefficients. There is no concern for over-fitting in the training phase since there is no training phase. This approach depends on a large dataset, and does not perform well when the number of patients in the database is not sufficiently large, especially when detecting two patients with similar gene expression profiles. In order to assess whether the 1-NN method can capture signals in spite of high noise, we compared it with a control method (ie, random patient selection).

Random patient selection

Whenever a patient A's gene expression profile is given, the random patient selection method randomly chooses a patient C in the database, and returns her/his survival time. Although it is based on random patient selection, this method is stronger than completely random survival generation since it at least considers the distribution of survivals. When the number of patients with longer survivals is smaller, the chance to predict longer survival is also smaller; the larger the number of patients within a range of survivals, the larger the probability

of predicting a number within that range of survivals. For each test sample, the method randomly chooses a patient and uses her/his survival as a prediction. The random patient selection process excludes the patient of the test sample for fair prediction simulations.

Regression-based survival prediction

Another approach is based on regression,[4] using k-fold cross-validation (CV) in order to reduce over-fitting. However, the regression approach faces the curse-of-dimensionality due to the nature of the problem: the number of genes is much larger than the number of samples.[5] In order to handle this issue, one may try to apply two different types of dimension reduction: feature selection and feature extraction. Machine learning with feature selection does not use whole genome gene expression profiles, but uses the expression profiles of selected genes, which is more relevant to the current biomarker approach to personalized medicine. Feature extraction produces new features generated from the original features, which are not easily interpreted in biomedical language.

In the case of high-dimensional predictors with a small number of samples, the traditional Cox regression model cannot be directly applied, and some genes are highly correlated.[5] Ridge regression with L_2-penalty and the least absolute shrinkage, as well as selection operator (LASSO) with L_1-penalty can handle the collinearity problem.[6] The LASSO was applied for variable selection in the Cox model.[7] The computationally more efficient least angle regression algorithm was used to obtain the solution of the Cox model.[5,8] In order to take advantage of both L_1 and L_2 penalties, an elastic net was developed.[9] More recently, the optimal application of these penalized regression methods to genomic data has been studied,[10] which showed that elastic net with two-dimensional tuning ($\lambda_1 + \lambda_2$) can perform comparably in both ridge regression-favoring simulation data and LASSO-favoring simulation data.

Friedman et al[11] developed an efficient algorithm for LASSO and elastic net regularized generalized linear models based on cyclical coordinate descent for linear, two-class logistic, and multinomial regression models with L_1 (LASSO) and L_2 (ridge regression), and a mixture of the two norms (elastic net) in 2010. Simon et al[12] developed an efficient procedure for the regularized Cox regression model (Coxnet) based on GLMnet in 2011. We used the R package of Coxnet for computing LASSO solutions with whole genome gene expression profiles and cancer pathway gene expression profiles,[12,13] since computing efficiency was essential for our experiments to perform nested CV where an inner CV loop was used for parameter determination, and an outer CV loop was used to estimate the prediction accuracy (ie, CV rate).

Prediction accuracy assessment

The accuracy of a prediction was measured as the absolute difference between observed survival and predicted survival. In order to compare the 1-NN and the random patient selection methods, we defined the overall prediction error as the mean absolute difference (MAD) of survival days:

$$\mathrm{MAD} = \frac{1}{n}\sum_{i=1}^{n} |s_o(i) - s_p(i)| \quad (1)$$

where $s_o(i)$ and $s_p(i)$ are the observed and predicted survival days of *i*-th sample, and *n* is the number of predictions. Observed survivals were obtained from days to death in the TCGA clinical data. Although the Cox model-based approach is capable of handling censored data (patient followed and alive), we did not include censored cases for better comparison of methods since the 1-NN method cannot be applied to these cases. To compare the performances of 1-NN and Cox-based methods, two Pearson's correlation coefficients were used: the first correlation coefficient (r_1) between observed survival and predicted survival for the 1-NN model; and the second correlation coefficient (r_2) between observed survival and relative risks obtained from the Cox model. Since r_2 is a negative value, we compared the absolute values of r_1 and r_2.

Results

As for the 1-NN method for each sample, we found the closest gene expression profile and predicted survival. The MAD measure (Equation 1) was used to demonstrate how good the predictions were. By repeating this process for all samples, we were able to compute the MAD value. The random patient selection method can show different results with different series of random numbers. In order to avoid the bias effect of special sequences of random numbers in the random patient selection method, we repeated its prediction process 100 times and reported the average of the MAD values.

Some samples belong to the same patients, which confounds the analysis towards the higher probability to select another sample from the same patient. In order to simulate predictions in the database, we ignored the closest samples of the same patient, and instead selected the closest sample observed from a different patient.

Comparison between 1-NN and random selection methods

Table 1 shows the mean absolute survival difference values between observed survival and the survival predicted by both the 1-NN and random patient selection methods. The MAD value of the 1-NN method was 386.2, whereas the average MAD value of the random patient selection method was 455.8. The lower prediction error of the 1-NN method compared to the random selection method illustrates that the 1-NN method can readily predict patient survival based on whole genome gene expression profiles, warranting further investigation of its prediction power in relation to regression-based predictions. Figure 1 shows the histogram of the absolute difference between observed survival and the survival predicted by the 1-NN method. Of note, the 1-NN method very accurately predicted the survival for more than 80 samples.

Comparison between 1-NN and Cox-based methods

The correlation coefficient (r_1) between observed survival and predicted survival obtained from the 1-NN method was 0.18 with a *P*-value of 0.00018. The Cox-based approach was performed with nested tenfold CV, where an inner loop was used for LASSO parameter determination. The average correlation coefficient ($\overline{r_2}$) was obtained from a series of correlation coefficients (r_2) between observed survival days and relative risks. When we used whole genome gene expression profiles, the average correlation coefficient ($\overline{r_2}$) was –0.22, with the absolute value being larger than r_1 (see Table 2). A reason for the higher prediction power of the Cox-based method compared to the 1-NN method could be due to the fact that the earlier method removes many genes unrelated to survival prediction by the LASSO optimization strategy. Only 164 genes among 12,042 total genes were used as an input to build models in the CV step due to the feature selection function of LASSO regression. The 164 genes included *SLC25A20, CLEC5A, ZNF208, C13orf18, NYX, PCNXL2, RBP1, EFEMP2, HIST3H2A, ELA2B,* and *RPS28*.

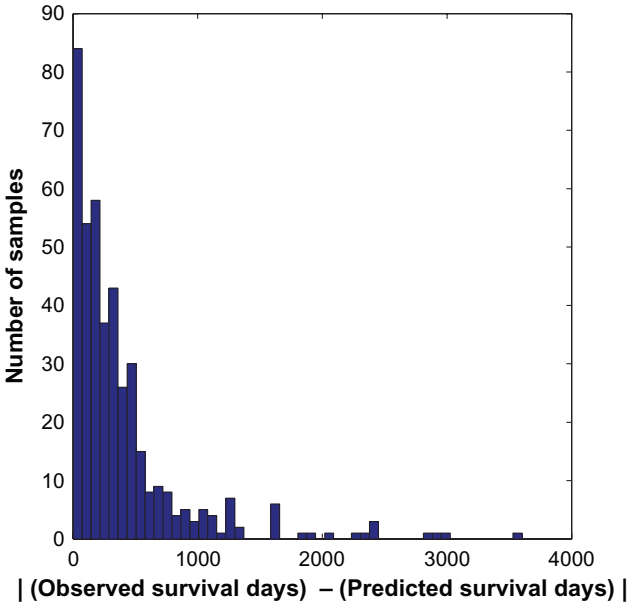

Figure 1 Histogram of absolute difference between observed survival (in days) and survival (in days) predicted by the 1-NN method.
Abbreviation: 1-NN, 1-nearest neighbor survival prediction method.

We then used a more focused gene input consisting of cancer pathway genes obtained from the Molecular Signatures Database (MSigDB) version 3.0,[14] and the Kyoto Encyclopedia of Genes and Genomes database.[15] Even though gene expression profiles of only 300 cancer genes were used as an input of LASSO optimization, the average correlation coefficient ($\overline{r_2}$) was –0.24, thus generating a better result than the same method using whole genome gene expression profiles. This result implies that the preselection of genes based on biological knowledge is still helpful even in the setting of LASSO, which is capable of sophisticated gene selection for more generalized fitting. Only 88 genes among 300 cancer genes were used for building the models in the CV step. These genes included *FZD7, MAPK8,*

Table 1 Survival prediction comparison between the 1-NN survival prediction method and the random patient selection method

Measure	Type of prediction	
	1-NN survival prediction	Random patient selection
MAD	386.2	455.8[a]

Note: [a]The average of MAD values for the random patient selection method was computed by repeating the simulation of survival prediction 100 times.
Abbreviations: 1-NN, 1-nearest neighbor survival prediction method; MAD, mean absolute difference between observed survival (in days) and predicted survival (in days).

Table 2 Survival prediction comparisons based on Pearson's correlation coefficient

Measure	Type of prediction		
	1-NN survival prediction	Coxnet with whole genome	Coxnet with cancer pathway genes
Correlation	0.18[a]	–0.22[b]	–0.24[b]

Notes: [a]Pearson's correlation coefficient between observed survival and predicted survival; [b]Pearson's correlation coefficient between observed survival and predicted relative risks for nested tenfold cross-validation, where an inner loop was used for LASSO parameter determination.
Abbreviations: 1-NN, 1-nearest neighbor survival prediction method; Coxnet, regularized Cox regression.[12]

LAMB4, NCOA4, RAC3, CCDC6, CTNNB1, CBL, ETS1, NFKB1, RARB, IL8, HIF1A, CASP3, NFKBIA, FZD8, EGF, CHUK, FGF5, BMP4, IL6, MET, TPM3, MITF, DVL3, GLI2, RB1, EGLN3, BMP2, SHH, SPI1, TRAF3, and *EPAS1,* many of which have cancer-relevant functions. Table 3 shows the functional annotation of the top 32 genes selected by the regularized Cox regression to various biological pathways,[12] including the Wnt, ERBB, nuclear factor-kappaB, and Hedgehog pathways.

Discussion

When cancer subcategories are known, it is reasonable to identify biomarkers that can discriminate between these cancer subtypes. The identification of class-separable biomarkers can be done via classification with feature selection. Even when cancer subcategories are not known, similarity comparisons using clustering algorithms can be applied to identify subcategories of cancers. However, rare subtypes of a cancer may not be captured due to small sample sizes. In this study, we focused on predicting patient survival based on gene expression profiles without grouping tumors into subtypes. The ability of this approach to predict individual patient survival represents a major advantage relative to the risk grouping of patient populations who share similar disease characteristics. Risk grouping classifies patients into distinct classes and tends to ignore the individual fate of each disease. Survival prediction and risk estimation algorithms that do not rely on cancer subclassification lend themselves to assist clinicians with difficult clinical decision-making. Such risk estimation is substantially easier to use and more adaptable to study tailored therapeutic options for individual cancer patients. Cancer subclassification and associated risk groupings provide only average predictions, limiting the ability to estimate the survival and risk of individual patients. As mentioned, cancer subtyping is inherently prone to fail in identifying and subgrouping patients with rare disease characteristics.

Table 3 Annotation to biological pathways of the top 32 genes (among 300 preselected cancer genes) used for building the models selected by Coxnet[12]

Pathway	Genes selected by Coxnet
Wnt pathway[a]	*FZD7, CTNNB1, FZD8, DVL3*
JNK pathway	*MAPK8, RAC3*
Apoptosis	*CASP3*
ECM receptor interaction	*LAMB4*
ERBB pathway[a]	*CCDC6, ETS1, IL8, EGF, FGF5, MET, TPM3*
HIF pathway	*HIF1A, EGLN3, EPAS1*
AKT pathway	*CBL*
NFkB pathway[a]	*NFKB1, NFKBIA, CHUK, TRAF3*
Retinoic acid receptor	*RARB*
Hedgehog pathway[a]	*BMP4, GLI2, BMP2, SHH*
Inflammation	*IL6*
Resistance to chemotherapy	*MITF*
Cell cycle	*RB1*
Gene expression during myeloid and B-lymphoid cell development	*SPI1*

Note: [a]Pathways with more than three genes selected by Coxnet.
Abbreviations: Coxnet, regularized Cox regression; JNK, C-Jan N-terminal kinase; ECM, extracellular matrix; HIF, hypoxia-inducible factor; NFkB, nuclear factor-kappa B; IL, interleukin; EGF, epidermal growth factor.

We compared four different survival prediction methods: (1) 1-NN survival prediction method; (2) random patient selection method, (3) Cox-based regression with LASSO optimization; with nested CV using whole-genome gene expression profiles; and (4) the same Cox-based regression method using gene expression profiles of cancer pathway genes. The 1-NN method used whole genome gene expression profiles for pattern matching, whereas the Cox-based regression method selected some genes for predicting relative risks based on LASSO optimization. We showed that the 1-NN survival prediction method was better than the random patient selection method, although it does not include a feature selection step. This 1-NN method may thus represent a valuable approach to capture the genome of a tumor that was closest to that of a tumor that was not categorized into a subtype due to its low frequency in previous studies. This is related to the issue of determining the number of clusters when a similarity comparison based on a clustering algorithm is used for cancer subtype identification. In general, small clusters tend to be ignored in more or less subjective decisions on tumor subtypes. The current 1-NN method determined the closest gene expression profile based on Pearson's correlation coefficient. We also tested Spearman's rank correlation and Hoeffding's D measure, but they did not show better results in terms of the MAD.

There is ongoing controversy as to the input of feature selection algorithms. If the feature selection is optimal, one may conclude that larger features should generate better results since the ideal feature selection would select the best set of genes. However, the practical situation is usually more complicated than the ideal situation. For example, the model parameter should be estimated by CV, but CV does not guarantee the identification of the actual best parameters; instead, it estimates good parameters that are close to the best parameters, primarily because the number of CVs and the step size of a grid parameter search are limited by available computing resources. We showed that the Cox-based regression method performed better when using 300 cancer

pathway genes that were preselected based on relevance to cancer biology rather than whole genomes (12,042 genes) as an input of the LASSO-based regression algorithm. This result implies that using preexisting biological knowledge for survival prediction is not only reasonable, but also beneficial – unless the target problem is to identify completely unknown cancer genes from the survival prediction.

Acknowledgments

This study was funded by a faculty start-up grant in the Division of Informatics, Department of Pathology, University of Alabama at Birmingham (UAB) School of Medicine. We are indebted to Noah Simon at Stanford University for discussions about his R package, Coxnet.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Scholz MB, Lo CC, Chain PS. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Curr Opin Biotechnol*. 2012;23(1):9–15.
2. Arias CR, Yeh HY, Soo VW. Biomarker identification for prostate cancer and lymph node metastasis from microarray data and protein interaction network using gene prioritization method. *Scientific World Journal*. 2012;2012:842727.
3. Xiang Y, Zhang CQ, Huang K. Predicting glioblastoma prognosis networks using weighted gene co-expression network analysis on TCGA data. *BMC Bioinformatics*. 2012;13 Suppl 2:S12.
4. Cox DR. Regression models and life-tables. *J R Stat Soc Series B Stat Methodol*. 1972;34(2):187–220.
5. Gui J, Li H. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*. 2005;21(13):3001–3008.
6. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Series B Stat Methodol*. 1996;58(1):267–288.
7. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med*. 1997;16(4):385–395.
8. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat*. 2004;32(2):407–499.
9. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol*. 2005;67(2):301–320.
10. Waldron L, Pintille M, Tsao MS, Shepherd FA, Huttenhower C, Jurisica I. Optimized application of penalized regression methods to diverse genomic data. *Bioinformatics*. 2011;27(24):3399–3406.
11. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1–22.
12. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J Stat Softw*. 2011;39(5):1–13.
13. R Development Core Team. R: A language and environment for statistical computing [webpage on the Internet]. Vienna: R Foundation for Statistical Computing. Available from: http://www.R-project.org/. Accessed: Jan 1, 2013.
14. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–15550.
15. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*. 2012;40(Database issue):D109–D114.