Journal of Biometrics & Biostatistics

Zhang et al., J Biom Biostat 2016, 7:1
DOI: 10.4172/2155-6180.1000281

Opinion

Open Access

Statistical Methods in Precision Medicine: Employing Systems Biology for Cancer Survival Prediction

Xinyan Zhang[1], Yan Li[1], Zaixiang Tang[2] and Nengjun Yi[1]*

[1]*Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35294, USA*
[2]*Jiangsu Key Laboratory of Preventive and Translational Medicine for Geriatric Diseases, Department of Epidemiology and Biostatistics, Soochow University, China*

Challenges in Predictive Modeling for Cancer Survival

Many cancer drugs showed limited therapeutic effects in fighting the tumor in a certain proportional of patients, due to the heterogeneity of tumors. Nevertheless, the current clinical pathological factors have not reached the expectation of accuracy in discriminating cancer patients. It is believed that this heterogeneity is, for a large part, genetically determined and rooted in molecular profile of the patient. Precision medicine has been initiated by White House to expand cancer genomics as a short-term goal to develop better prevention and treatment methods for more cancers. Recent high-throughput technologies can easily and robustly generate large-scale molecular profiling data, offering extraordinary opportunities to develop molecular signature or biomarkers through predictive modeling on the patients' survival or metastatic status. Notably, in analyzing these large-scale data, a potential statistical challenge arises in which the number of predictor variables greatly exceeds the sample size. The classical Cox proportional hazard model cannot simultaneously analyze a large number of and/or correlated predictors, due to the problems of non-identifiability and possibly overfitting. To date, various statistical approaches have been applied in analyzing large-scale molecular profiling data to build predictive models for cancer survival prediction and prognosis, which will be discussed in the following section.

Predictive Modeling with High-dimensional Genomic Data in the Survival Framework

For high dimensional data, the standard use of Cox proportional hazards model is highly unstable in terms of multicollinearity. Various methods have been proposed to solve these problems under this motivation. We will selectively review these methods in the following subsections. For a detailed performance comparative discussion, please refer to Bøvelstad et al.

Univariate variable selection or forward stepwise selection

Univariate method tests each gene one by one through univariate Cox regression model which is considered robust and easier to carry out. Forward stepwise selection is performed by adding genes one by one to the cox regression model until they select similar number of top genes which put correlation among predictors into consideration but also considered as a greedy approach. Score test were used for both methods instead of likelihood ratio and Wald tests because it does not require to estimate regression coefficients so that it outperforms in reducing computational time when dealing with large data sets. Both of these two methods work quite poorly in predicting cancer survival with large data sets compared to those other methods in the following sections.

Principal components analysis

Principal component analysis (PCA) is a technique to explain the variance-covariance structure through a few linear orthogonal combinations of the original genomic variables. Yeung and Ruzzo adapted this mathematic techinique in gene experssion as a clustering tool. In Bøvelstad et al.'s comparative study, they reduced the genomic covariates matrix to first chosen number of principal components which would then be used to fit the multivariable cox regression model as the predictive model. Because of that PCA cannot guarantee the association between components and survival outcome, Bair and Tibshirani and Bair et al. proposed a supervised principal components analysis which applied univariate selection to pick out sets of genes and then used PCA for the chosen genes. Although Bair and Tibshirani and Bair et al.'s results showed that the supervised PCA outperformed the unsupervised PCA, Bøvelstad et al. made a contrary conclusion based on analysis of three datasets.

Partial least squares regression

Similar to PCA, Partial least squares (PLS) regression is also a data reduction technique which constructs a set of linear combinations of genomic variables incorporating survival outcome as weights. PLS cox regression for including both clinical covariates and genomic data but only utilizing PLS in genomic data has been applied in different studies Also similar to PCA, unsupervised PLS may be problematic since it left out the association between components and survival outcome. Nguyen and Rocke developed a two-stage PLS as a supervised approach. They first determined PLS components through linear regression for survival data and then Cox regression was fitted with the resulted components. The linear regression step was replaced by Cox regression by Bastien Bøvelstad et al. also demonstrated that the pre-select stage in supervised PLS did not improve the prediction performance but lead to unstable results.

Penalized cox models

Penalized cox models with various penalties have been developed. Ver weij and van Houwelingen suggested a L2 penalized Cox regression model also known as Ridge which does not possess the sparsity property. Tibshirani proposed the application of Lasso (least absolute shrinkage and selection operator) in Cox regression model to achieve sparsity. The elastic net is also a widely used penalization approach proposed by Zou and Hastie. Furthermore, various extensions of Lasso have been proposed and widely applied in different studies. In penalized Cox framework, a penalty are added to the log-likelihood

***Corresponding author:** Nengjun Yi, Department of Biostatistics, Ryals Public Health Bldg 327, 1665 University Blvd, University of Alabama at Birmingham, Birmingham, AL 35294-0022, USA, Tel: (205) 934-4924; Fax: (205) 975-2540; E-mail: nyi@uab.edu

Received February 15, 2016; **Accepted** February 18, 2016; **Published** February 25, 2016

Citation: Zhang X, Li Y, Tang Z, Yi N (2016) Statistical Methods in Precision Medicine: Employing Systems Biology for Cancer Survival Prediction. J Biom Biostat 7: 281. doi:10.4172/2155-6180.1000281

function and estimates the parameters β by maximizing the penalized log-likelihood. The partial log likelihood function in penalized cox model can be summarized as: $pl_{pen}(\beta) = pl(\beta) - P_{\lambda}(\beta)$, where Cox partial log-likelihood is $pl(\beta) = \sum_{i=1}^{n} d_i \log[\exp(X_i\beta) / (\sum_{j \in R(t_i)} \exp(X_j\beta))]$ and $P_{\lambda}(\beta)$ is the penalty function which differs among methods and λ denotes the tuning parameter and normally determined by cross-validation. By imposing constraints on the parameter coefficients, the penalty functions can reduce modeling biases and improve prediction accuracy to provide meaningful estimates, even when highly correlated predictors are involved. Small penalties generate large models with smaller bias but potentially higher variance; while large penalties result in less variance but selection of fewer predictors.

Bayesian hierarchical cox models

Another efficient approach to handling high-dimensional data is hierarchical modeling, that is, the regression coefficients in the model are themselves modeled and are normally handled in the Bayesian framework. Various prior distributions can be used. Many penalized regressions can re-expressed as a Bayesian hierarchical model, just expressing the penalty term as a prior distribution of the parameters. For example, the ridge penalty can be expressed as a normal distribution, and the lasso penalty can be expressed as a double-exponential distribution.

Integrating pathway information

Besides the methods discussed above, some statistical methods incorporating higher-order information of functional units in cancer, i.e. pathways, have also been investigated. Abraham et al. adopted a gene set statistic to provide stability of prognostic signatures instead of individual genes. Huang et al. converted the gene matrix to a pathway matrix through “principal curve”, similar to PCA. Both of these two methods did not incorporate outcome when generating the pathways scores from the individual genes. Some other sophisticated group-wise statistical methods have been developed using an “all-in-all-out” idea meaning when one predictor in a group is chosen, then all variables in that group are chosen. Eng et al. proposed a method to reduce the computational complexity by incorporating a binary outcome to stand for decreased or increased risk score in each pathway as well.

Concluding Remarks

With the emergence of technologies, genomic data has become easily feasible and often encounters the problem with large number of predictors much exceeding the number of subjects. Among our selective review, both univariate and forward selection are easy and robust with certain limitations. Principal component analysis and partial least squares have outperformed in prediction but are both lack of the ability to detect a specific gene through variable selection. Penalized regressions are popular variable selection methods as their efficiency in computation and the ability to detect significant genes. The limitations of penalized regression are obvious as well. The lasso put L1-penalty on the coefficients and can shrink many coefficients exactly to zero, thus performing variable selection, but it preform ineffective when no significant differences among predictors. On the other hand, ridge performs well for evenly distributed coefficients for the predictors. The research in using Bayesian methods for cancer survival prediction is limited. Furthermore, in order to achieve better prediction of cancer treatment, pathway information incorporated and clinical pathological factors combination may be extremely crucial which require some novel statistical methods to be applied.