SubLasso：a feature selection and classification R package with a fixed feature subset

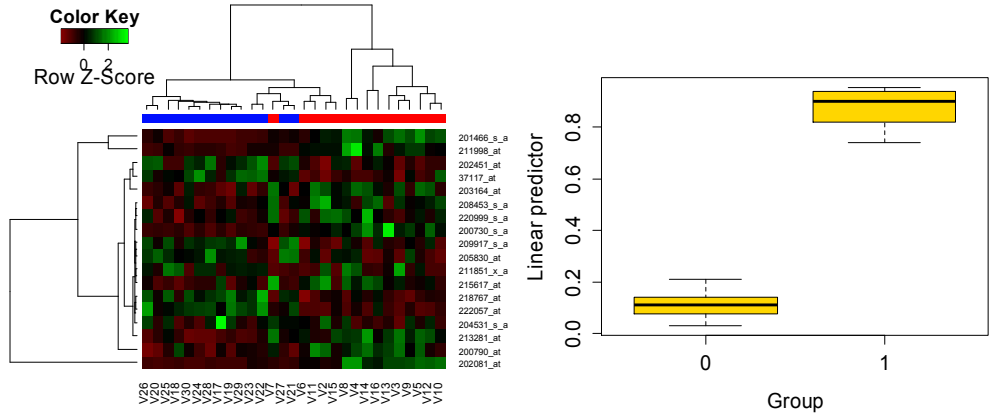
Youxi Luo[1,3,*], Qinghan Meng[1,2,*], Ruiquan Ge[1,2], Guoqin Mai[1], Jikui Liu[1], Fengfeng Zhou[1,#]

1. *Shenzhen Institutes of Advanced Technology, and Key Lab for Health Informatics, Chinese Academy of Sciences, Shenzhen, Guangdong, 518055, China.*
2. *University of Chinese Academy of Sciences, Beijing, 100049, China*
3. *School of Science, Hubei University of Technology, Wuhan, Hubei, 430068, China.*

* These authors contribute equally to this work.

Corresponding author: Fengfeng Zhou, phone: +86-755-86392200; fax: +86-755-86392299; e-mail: FengfengZhou@gmail.com, or ff.zhou@siat.ac.cn. Web site: http://www.healthinformaticslab.org/ffzhou/ .

Supplementary Figure 1



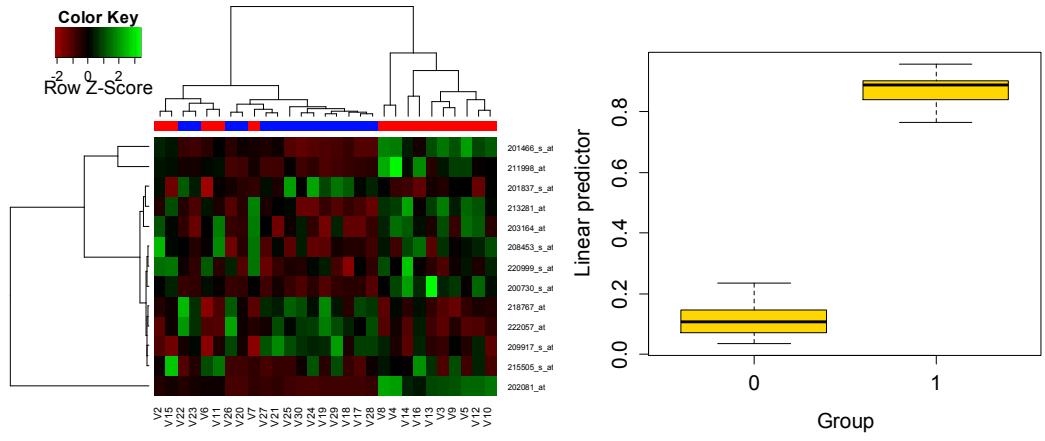
(a) (b)

```
1.239365e-02*Exp("211851_x_at") + 2.742123e-03*Exp("204531_s_at")
+ 6.162963e-04*Exp("37117_at") + -1.269584e-03*Exp("200730_s_at") + -5.761154e-05*Exp("200790_at") + -1.385047e-04*Exp("201466_s_at")
+ -5.562896e-05*Exp("202081_at") + 2.501702e-04*Exp("202451_at") + -3.131031e-04*Exp("203164_at") + 7.980953e-04*Exp("205830_at")
+ -1.441833e-03*Exp("208453_s_at") + 2.553688e-03*Exp("209917_s_at") + -1.289445e-05*Exp("211998_at") + -1.757298e-04*Exp("213281_at")
+ -7.394639e-03*Exp("215617_at") + 2.221148e-03*Exp("218767_at") + -3.249635e-03*Exp("220999_s_at") + 8.816857e-04*Exp("222057_at")
+ 0.0738978
```

(c)

Supplementary Figure 1. The *k*-fold cross validation results of SubLasso classification model based on BRCA1 and 16 other probesets (*k*=5 by default). (a) Unsupervised hierarchical clustering and the heatmap of the features selected by SubLasso. Each row has the expression levels of one selected feature, and each column is a sample. Blue and red in the horizontal color bar represent the samples with *Y*=1 and 0, respectively. (b) Box plot of the trained Lasso linear score with the mean, 25th and 75th percentiles (c) The linear discriminating function generated by SubLasso, where Exp(*x*) is the expression level of the probeset "*x*".

Supplementary Figure 2



(a) (b)

Supplementary Figure 2. The *k*-fold cross validation results of SubLasso classification model

without preselected features (*k*=5 by default). (a) Unsupervised hierarchical clustering and the heatmap of the features selected by SubLasso. Each row has the expression levels of one selected feature, and each column is a sample. Blue and red in the horizontal color bar represent the samples with Y=1 and 0, respectively. (b) Box plot of the trained Lasso linear score with both Box plot of the trained Lasso linear score with the mean, 25th and 75th percentiles.

Materials and methods

Binary classification model

Let Y be a binary category variable, and Y=1 and 0 be two types of samples. Each sample X=<X_1, X_2, …, X_p> is a p dimensional vector. P=P(Y=1|X_1, X_2, …, X_p) describes the probability of Y=1 on condition of X_1, X_2, …, X_p. The following logistic model was used in the SubLasso package:

$$Ln(P/(1\text{-}P))=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p \quad (1)$$

Assuming that $\hat{w}=\left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2,..., \hat{\beta}_p\right)$ is the estimation of w=(β_0, β_1, β_2, …, β_p), the following formula can be used to predict the probability of Y=1.

$$\hat{P}=1/\left(1+e^{-(\hat{\beta}_0+\hat{\beta}_1X_1+\hat{\beta}_2X_2+\ldots+\hat{\beta}_pX_P)}\right) \quad (2)$$

Given a threshold α, Y is predicted to be 1 if $\hat{P}>\alpha$, otherwise Y=0.

SubLasso algorithm

In the formula (1), β_0 is a constant, and β_i is the coefficient of the feature i, where i=1, 2, …, p. A given feature i will not be selected, if its coefficient β_i is close or equal to zero. For a given high-dimensional binary classification problem, most of the dimensions/features do not have correlations with the class label $Y\in\{0, 1\}$. A widely used Lasso algorithm adds a L_1 penalty to the maximum-log-likelihood optimization of the formula (1), as in the following formula (3).

$$\hat{w}_{Lasso}=\left(\hat{\beta}_0, \hat{\beta}_{Lasso}\right)=\arg\max_{w}\left\{L(\beta_0, \beta; X, y)-\lambda\|\beta\|_1^1\right\} \quad (3)$$

L(β_0, β; X, y) is the log-likelihood function and λ is the penalty manipulating parameter that controls the degree shrinkage. A feature is selected by the proposed algorithm, if its coefficient is not zero.

SubLasso is a modified Lasso algorithm, allowing the user to define a subset of features that must be selected in the final model. This model restriction comes from the biologists, who want to know whether a given well-known disease biomarker together with a few other features may facilitate a highly accurate disease classification/prediction model. The major efforts of large-scale disease diagnosis studies are on the association evaluation of individual sample features. Although a number of statistically phenotype-associated features were identified for many diseases, their

population ratios are still not high enough for the prediction/diagnosis of the investigated diseases, *e.g.* the most prevalent breast cancer biomarker BRCA1 appears in only 0.25% of the general population (Whittemore, et al., 2004). SubLasso is proposed for the investigation of the existence of accurate disease classification models based on the known biomarkers, *e.g.* BRCA1 for the breast cancer patients (Whittemore, et al., 2004).

Given a user-defined subset of features $S = \{X'_1, X'_2, ..., X'_k\}$ and their coefficients $\beta^S = (\beta'_1, \beta'_2, ..., \beta'_k)$, the solution of SubLasso may be approximated as:

$$\hat{w}_{SubLasso} = \left(\hat{\beta}_0, \hat{\beta}^S, \hat{\beta}^{-S}_{Lasso}\right) = \arg\max_{w} \left\{ L(\beta_0, \beta^S, \beta^{-S}; X, y) - \lambda \left\| \beta^{-S} \right\|_1^1 \right\} \quad (4)$$

where β^{-S} is the complementary set of β^S, and coordinate descent algorithm ((Friedman, et al., 2010)) is used to solve the above formula (4).

In the SubLasso algorithm, the user-defined feature subset *S* may be set to null, if the user does not want to fix any features. The formula (3) will be solved in this case.

Classification performance measurements

The K-fold cross validation strategy is used to evaluate the classification model, and the classification performance is measured by *sensitivity* (*Sn*), *specificity* (*Sp*), *accuracy* (*Acc*), and *Matthew's correlation coefficient* (*MCC*) (Li, et al., 2013; Matthews, 1975). They are defined as follows:
$Sn=TP/(FP+FN)$, $Sp=TN/(FP+TN)$, $Acc=(TP+TN)/(TP+FN+FP+TN)$, and

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}.$$

TP and *FN* are the numbers of samples with *Y*=1 that are predicted as *Y*=1 and *Y*=0, respectively. *FP* and *TN* are the numbers of samples with *Y*=0 that are predicted as *Y*=1 and *Y*=0, respectively.

Results and Discussion

Installation and example demonstration

SubLasso may be installed using the menu function "Install Packages" and selecting the "Install from: Package Archive File (.zip; .tar.gz)" from a downloaded installation file. The current version is 1.2, and its help information may be accessed by typing "??SubLasso" in the command line of the software R or RStudio. SubLasso has been tested in the software R version 3.0.3 and RStudio version 0.98.501.

Two datasets, Golub_Merge (Golub, et al., 1999) and Colon (Alon, et al., 1999), are embedded in the package SubLasso, and the example R codes are provided in the module SubLasso::SubLasso. The boxplot and heatmap of the prediction results are provided for intuitive demonstrations.

Input data

SubLasso needs a data matrix, of which each row has the data of a given feature in all the samples, and each column has the data of a given sample's all features. Another information is the class label of each sample, and this should be 1 or 0. It is suggested that 1 and 0 are the labels of disease and control samples, respectively.

The list of pre-selected features may be provided as the optional third parameter. It is required that features in this list belong to the set of row names, which is provided in the first row, otherwise an error message “Some subset variables out of range” will be returned. A null set is provided as the default and ordinary Lasso solution will be generated.

The optional fourth parameter is the fold number for the cross validation, whose default value is 5.

Output information

The microarray-based gene expression profile of breast cancer with the GEO ID GSE9574 (Tripathi, et al., 2008) is used as an example dataset for demonstration of how SubLasso performs. The dataset GSE9574 has two groups of samples, the breast cancer (Y=1, n=14) and the control individuals (Y=0, n=15). The two probesets of BRCA1, *i.e.* 211851_x_at and 204531_s_at, are chosen as the pre-selected features, and all the other parameters of SubLasso are set as default values. The CEL files are downloaded from the GEO database, and normalized by the standard QC procedure and MAS 5.0 algorithm.

Sixteen other features are selected for the optimized classification model, and the classification performance is demonstrated by Figure 1. Firstly, the heatmap and the unsupervised hierarchical clustering plot in Figure 1 (a) show that the features selected by SubLasso may reasonably discriminate the cancer samples from the controls, even using the unsupervised clustering method. Only one control sample is mislabeled as the disease type. The constructed classification model proposed a perfect discrimination between the two classes of samples, as shown in Figure 1 (b). The classification performance measurements (*Sn*, *Sp*, *Acc*, *MCC*) reach (1, 1, 1, 1). Although the SubLasso model without preselected features also achieves (*Sn*, *Sp*, *Acc*, *MCC*)=(1, 1, 1, 1), as shown in Figure 2 (a) and (b), there are 5 false positives for the unsupervised clustering. This suggests that the first model based on BRCA1 seems to have more intra-class consistency in the expression levels than the second one. Unfortunately, if we don't have the knowledge of the BRCA1's association with breast cancer, the statistical optimization itself doesn't find BRCA1 in the second *de novo* model, as in Figure 2.

Another merit of a Lasso classification model is that its linear discriminating function provides an intuitive description of each feature's positive or negative association with the class labels. Figure 1 (c) gives the linear Lasso function, and the two features of BRCA1 have the top two positive weights among all the selected features. So the increased expression levels of BRCA1 will increase the risk of developing ER1 breast cancer, based on the above data and the literature (Hosey, et al., 2007). The feature 215617_at (gene SPATS2L, spermatogenesis associated, serine-rich 2-like) has the maximal negative weight in the linear function. Although there is no direct support from the literature, this observation may be explained by that breast cancer occurs much more frequently in females than males.

Prediction for new data

SubLasso also provides a prediction for new samples. A prediction module predict.SubLasso() is provided for this purpose, and its parameters are similar to the other classification models. The example R codes and results are provided in the help document of SubLasso, and may be tested in the softwares R or RStudio.

Example R codes for direct deployment on new datasets

Besides the informatics users, experienced biologists may also easily explore their datasets with known biomarkers using SubLasso. For this purpose, we provide an example dataset GSE9574 (Tripathi, et al., 2008) and the training/testing R scripts. The dataset was generated from the gene expression profile of breast cancer patients and the control samples. The first line of the dataset file is the samples' GEO IDs, and the second line gives the class labels of cancer or control for the sample in each column. The rest lines are the expression levels of the probesets normalized by the RMA algorithm (Irizarry, et al., 2003). The R script "exSubLasso.R" provides the functionalities from loading the dataset file, training the SubLasso model, to saving the trained models. In case that the user may want to firstly explore the SubLasso model without fixing any biomarkers, the training R code is also provided. Both the SubLasso models without and with a pre-fixed feature subset are saved as external files, so that the user may reload a model of interest for prediction of new data. A heatmap and a boxplot are generated for visualizing the model performance of the selected features. The R script "testSubLasso.R" loads both of the two aforementioned models for the prediction of new data. Basically, the user may normalize their dataset into the above file format, and explore the dataset directly using the two R scripts.

Conclusions

This study proposed a user-friendly feature selection and classification SubLasso model as a R package, where a user may decide some features in the final model. This is necessary for the

current research community with quite many disease-associated biomarkers. Although these biomarkers are discerned as statistically significantly associated with a given disease type by large-scale genome-wide screening, they alone may still have very low population ratio, *e.g.* BRCA1 for breast cancer (Hosey, et al., 2007; Whittemore, et al., 2004), and are difficult to be used as diagnosis evidence. As far as we know, this is the first bioinformatics tool for this purpose, and we believe that SubLasso will be a valuable tool for both biologists and bioinformaticians to carry out the hypothesis-driven OMICs data exploration. For the convenience of users, we also provide an R script and example input data files, so that the users may put their own data in the same format and get the model accuracy evaluation instantly.

References

Alon, U.*, et al.* (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc Natl Acad Sci U S A*, **96**, 6745-6750.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software*, **33**, 1-22.

Golub, T.R.*, et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, **286**, 531-537.

Hosey, A.M.*, et al.* (2007) Molecular basis for estrogen receptor alpha deficiency in BRCA1-linked breast cancer, *J Natl Cancer Inst*, **99**, 1683-1694.

Irizarry, R.A.*, et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Biostatistics*, **4**, 249-264.

Li, K.*, et al.* (2013) Screening features to improve the class prediction of acute myeloid leukemia and myelodysplastic syndrome, *Gene*, **512**, 348-354.

Matthews, B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochim Biophys Acta*, **405**, 442-451.

Tripathi, A.*, et al.* (2008) Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients, *Int J Cancer*, **122**, 1557-1566.

Whittemore, A.S.*, et al.* (2004) Prevalence of BRCA1 mutation carriers among U.S. non-Hispanic Whites, *Cancer Epidemiol Biomarkers Prev*, **13**, 2078-2083.