A Comparison of Classification Methods for Expression Quantitative Trait Loci

Joe R. Davis[1], Nilah M. Ioannidis[1], Kim Kukurba[1], Zach Zappala[1], Carlos D. Bustamante[1,2], Stephen B. Montgomery[1,3]

[1] Department of Genetics, [2] Biology, and [3] Pathology, Stanford University, Stanford, California, USA

Introduction

The functional impact of regulatory variation in the human genome remains largely unexplored. One group of regulatory variants called expression quantitative trait loci (eQTLs) has become the focus of recent large scale efforts to identify variants that effect gene regulation [2, 6, 9, 10]. eQTLs refer to non-coding variants that influence the expression of a nearby gene such that individuals homozygous for one allele of the gene will show decreased expression relative to individuals homozygous for the other allele. In addition to influencing expression, eQTLs are enriched for disease causing variants and variants that influence molecular phenotypes such as chromatin modifications [2, 4]. Thus, these variants are of important biological interest for addressing disease etiology.

Until recently, the study of eQTLs has focused primarily on discovery using simple linear models [2, 6, 9, 10]. Only a handful of studies have sought to build more complex models for the purpose of eQTL classification [2, 4]; these studies have been limited in the scope of models investigated. In this project, we examine four main classification methods for eQTL detection: 1) logistic regression, 2) decision trees and random forests, 3) support vector machines (SVM), and 4) neural networks. Our goal is to build a classification model to predict whether a variant influences gene expression using only publicly available genomic annotation. This model will therefore be able to classify eQTLs rapidly without the need of gene expression data, greatly improving our ability to interpret functional variation in the non-coding genome. In addition, by examining which features/annotations are most informative for eQTL classification, we can make meaningful inferences regarding the underlying biology of eQTLs.

Data

The GEUVADIS Consortium recently conducted one of the largest eQTL studies to date on 373 European derived lymphoblastoid cell lines (LCLs) [6]. Their genotype and expression data is publicly available (`http://www.ebi.ac.uk/Tools/geuvadis-das/`). We re-ran their eQTL analysis to generate a list of 5738 high confidence eQTLs with FDR < 0.05. Additionally, we chose as our control set 5086 SNPs with no significant association to gene expression.

Features for each SNP were obtained from the CADD resource, a publicly available database with over 90 annotations for nearly every base in the human genome (`http://cadd.gs.washington.edu/`) [5]. 68 features were selected from the CADD database for model training and testing. Features were selected based on missingness rate ($< 75\%$) and biological relevance. As a pre-processing step, we intersected our list of eQTLs and controls with the CADD features. We matched each variant in the CADD database for the specific alternative allele in our eQTL/control list. Our final dataset consisted of 10824 eQTLs and control SNPs with 68 features per observation.

Features

The 68 selected features cover a range of genomic annotations, from estimates of selection and conservation to measures of open chromatin and transcription factor binding. 16 features are categorical with number of factors ranging from 2 to 26, while the remaining features are continuous or integer valued. For a significant number of our 68 predictors, the missingness rate is quite high (35 predictors with missingness > 0.50). The CADD resource gives guidelines for imputation of missing values for these variables, and these guidelines were followed in construction of the final training and test sets. Missingness for some of the predictors can be informative. For example, if one of the predictors is measuring the amino acid change caused by a variant, missing data would indicate that the variant is not within a protein coding

region. Therefore, for 10 predictors with missing values, additional binary categorical predictors were created as indicators of whether the original predictor was missing or not for a given variant.

Models

Our goal is to predict using the annotations for a given regulatory variant whether that variant is an eQTL or not. Formally, for regulatory variant i ($i = 1, \cdots, m$), we define the class label:

$$y_i = \begin{cases} 1 & : \text{ if variant i is an eQTL} \\ 0 & : \text{ otherwise} \end{cases}$$

Let x_i be the feature vector for variant i. Our performance metric will be classification error. We compared four main types of classification models: 1) logistic regression, 2) SVMs, 3) tree-based models, and 4) neural networks. These models were chosen for their ability to handle the disparate feature types. Additionally, our literature review revealed that tree-based and SVM-based methods have performed well for similar problems focused on coding variation [1, 5]. Training was performed on 70% of the sample (randomly chosen), while testing occurred on the remaining fraction (30%).

Logistic Regression Models: We first investigated a logistic regression model including the main effects from 58 of the 68 features. The 10 binary features derived to account for missingness in other predictors were not included in the model. No higher order interactions were investigated. This model, termed the Full Model, is specified as follows:

$$P(y_i = 1|x_i) = \frac{1}{1 + \exp(\beta_0 + \beta^T x_i)}$$

where β_0 is the intercept term and β is the vector of coefficient estimates. We then performed feature selection on the Full Model using two methods: 1) forward-backward stepwise regression using AIC (Akaike Information Criterion) and 2) L1-regularized regression (Lasso regression) [3]. The Full Model and stepwise model were fit using the R functions glm and step, respectively. We used the R package glmnet to fit the lasso model [3]. λ, the lasso penalty parameter, was chosen via 10-fold cross-validation.

SVMs: We considered three different SVM models defined by the choice of kernel function: 1) linear, 2) radial, and 3) sigmoid. These kernel functions are defined as follows:

$$\begin{aligned} k_{\text{linear}}(x_i, x_j) &= x_i^T x_j \\ k_{\text{radial}}(x_i, x_j) &= exp(\gamma \|x_i - x_j\|_2^2) \\ k_{\text{sigmoid}}(x_i, x_j) &= tanh(\gamma x_i^T x_j) \end{aligned}$$

By default, we set the coefficient $\gamma = \frac{1}{\dim(x)}$. Models were fit using the R package e1071 [8].

Tree-based Models: We considered two tree-based models: 1) a single classification tree and 2) a random forest. For both models, the objective function minimized during model fitting was the training classification error. The classification tree was fit using the R package rpart [11]. The full tree was fit and then pruned by minimization of the 10-fold cross-validation error. The final, pruned tree contained 13 splits with a complexity parameter of 0.00261. The random forest was fit using the R package randomForest [7]. The forest was grown to 500 trees; there were no restrictions on tree size. The number of predictors sampled in each split was set to the default, namely $\sqrt{\dim(x)}$.

Neural Networks: Single hidden layer neural networks were fit using the R package nnet [12]. The learning decay parameter was set to 0 (default) and the maximum number of fitting iterations was set to 1000. In all, 20 models were trained, with models having $h = 1, \cdots, 20$ nodes in the hidden layer. Models were fit by minimization of the cross-entropy or deviance:

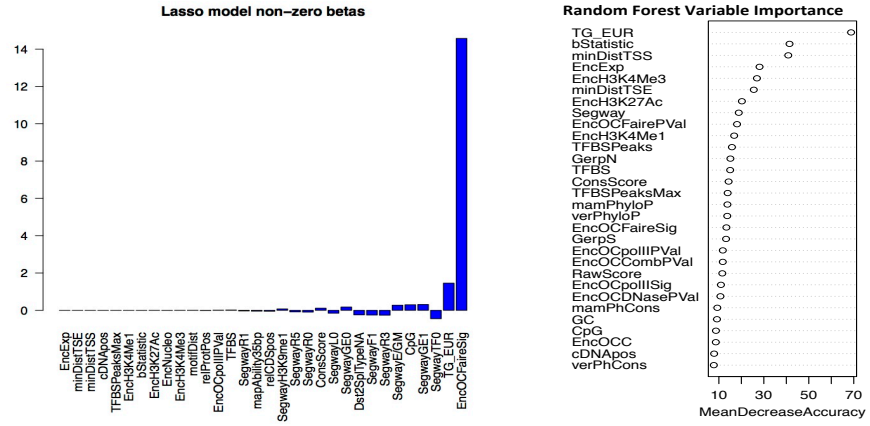
$$\underset{\theta}{\text{argmin}} \sum_{i=1}^{m} -y_i \log f(x_i; \theta) - (1 - y_i) \log(1 - f(x_i; \theta))$$

where $f(x_i;\theta)$ are the predicted probabilities from the network for observation i and θ is the vector of network parameters. The model that minimized testing error (i.e. the model with $h = 7$ hidden nodes) was chosen for comparison to the remaining three model types.

Results

The table below gives the training and test error for all 9 models. In terms of training error, the neural network with 7 hidden nodes performs the best, followed by the tree-based models and SVMs with radial and linear kernels. However, training error can be a misleading metric if over-fitting occurs. On test error, the tree-based models perform the best followed by SVMs with radial and linear kernels. Surprisingly, the lasso model performs quite well, given that it does not attempt to model higher order interactions unlike models from the other three types. The neural network performs extremely poorly compared to the other models and given it's performance on the training data. This results argues strongly for over-fitting in the neural network. In general, all models have a performance on the test set around $30 - 35\%$.

Model	Training (7577)	Test (3247)
Full additive logistic	0.3561	0.3631
Stepwise logistic	0.3558	0.3671
Lasso logistic	0.3212	0.3366
SVM Linear	0.3109	0.3302
SVM Radial	0.2746	0.3234
SVM Sigmoid	0.3780	0.4090
Decision Tree	0.2736	0.2937
Random Forest	0.2773	0.2913
Neural Net	0.1334	0.3723



.

Figure 1: Lasso regression coefficient estimates (left) and Random Forest variable importance (right)

We were also interested in determining features highly predictive of eQTL status. Figure 1 shows

the parameter estimates for features selected by the lasso regression (left) and the variable importance (mean decrease in accuracy) plots for the random forest (right). Both models select very similar features as highly predictive; these features are generally enriched for annotations regarding regulatory potential (ex. TFBS or number transcription factor binding sites disrupted), distance to nearest gene (ex. minDistTSS or absolute value of the minimum distance to transcription start site), degree of conservation/selection (ex. GerpN, a measure of sequence conservation), and specific chromatin states (ex. Segway, a chromatin segmentation map).

Discussion

Tree-based models achieved the best performance of all models considered, followed closely by SVMs with radial and linear kernels. The neural network model performed rather poorly, showing strong evidence of over-fitting. Surprisingly, the lasso logistic regression model performed quite well. This model is far simpler than the tree-based and SVM models and is also more tractable and interpretable. For example, from the random forest model we know the feature TG_EUR (average allele frequency in European populations) is strongly predictive, but interpreting it's exact influence on eQTL classification is difficult given the higher order interactions implicit in the model. However, from the lasso model, we see that increasing TG_EUR increases the probability of a variant being an eQTL. Moreover, given that the lasso model only accounts for main effects, its strong performance suggests that the influence of higher order interactions is relatively small. 87 parameters were estimated in the full logistic model compared to 54 for the stepwise model and 32 for the lasso model. Thus, many main effects also appear to be of little predictive value.

The performance of our models was bounded below by 29% miclassification error on the test set. This low performance is likely partly due to the resolution of our features. Most of the features have a resoltion of 10-1000 base pairs; we are, instead, trying to predict class labels for single base pair variants. The high degree of missingness in the feature set could also be limiting model performance. Lastly, the problem could be confounded by the underlying biology. The number of non-coding variants in the human genome is on the order of millions to tens of millions, yet the number of detected eQTLs is on the order of tens of thousands. They are relatively quite rare, and we simply may not have observed enough of them to determine how they differ from the background non-coding variation.

Finally, using the coefficient estimates from the logistic regression models and the variable importance measures from the tree based models, we were able to quantify the relative influence of each of the features on eQTL state. From these measures, we can make some interesting biological conclusions. For example, minDistTSS was found to be highly informative in the random forest model. It was also found to have a negative predictive effect ($\hat{\beta}_{\text{minDistTSS}} = -7.21e-6$) in the lasso model, indicating that variants nearer the site of transcription are more likely to be eQTLs, a result previously confirmed by other studies [2, 10]. More broadly, we can infer that annotations associated with gene regulation, chromatin activity, transcription factor binding, and conservation are highly predictive of eQTL state, while annotations associated with protein function (ex. SIFTval, a measure of deleteriousness of a protein-coding variant) are not predictive. These results provide evidence that the biological mechanisms influencing eQTLs, and possibly other regulatory variants, are not shared by protein-coding variation. The development of predictive models specifically for non-coding variations is, therefore, all the more pertinent.

Conclusion and Future Directions

We have presented one of the first studies to apply a diverse range of machine learning algorithms to the novel task of eQTL prediction. We compared four general algorithms, and found that performance was optimized by tree-based models followed closely by SVMs and lasso logistic regression. Using these models, we were also able to make meaningful inferences regarding the underlying biology of eQTLs, highlighting there distinction from protein-coding variation.

Unfortunately, our classification error for all models was relatively high. Future work should focus on error analysis to understand the source of this high error. Additionally, other large resources of regulatory annotations could be used to increase the feature set, hopefully including novel features that improve prediction. We can also use other databases of eQTLs to increase training and test set size [2, 9, 10]. More long term, we want to explore these models in the context of other regulatory variation, e.g. variants influencing transcript processing, chromatin modifications, and chromatin activity. We want to expand our models to produce a resource for annotation of the entire non-coding genome, and, thereby, improve our ability to interpret personal genomic variation.

References

[1] Ivan a Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–9, April 2010.

[2] Alexis Battle, Sara Mostafavi, Xiaowei Zhu, James B Potash, Myrna M Weissman, Courtney McCormick, Christian D Haudenschild, Kenneth B Beckman, Jianxin Shi, Rui Mei, Alexander E Urban, Stephen B Montgomery, Douglas F Levinson, and Daphne Koller. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome research*, 24(1):14–24, January 2014.

[3] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 2010.

[4] Daniel J Gaffney, Jean-Baptiste Veyrieras, Jacob F Degner, Roger Pique-Regi, Athma a Pai, Gregory E Crawford, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. Dissecting the regulatory architecture of gene expression QTLs. *Genome biology*, 13(1):R7, January 2012.

[5] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310–5, March 2014.

[6] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter ACt Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 2013.

[7] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[8] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2014. R package version 1.6-4.

[9] Stephen B Montgomery, Micha Sammeth, Maria Gutierrez-Arcelus, Radoslaw P Lach, Catherine Ingle, James Nisbett, Roderic Guigo, and Emmanouil T Dermitzakis. Transcriptome genetics using second generation sequencing in a caucasian population. *Nature*, 464(7289):773–777, 2010.

[10] Joseph K Pickrell, John C Marioni, Athma A Pai, Jacob F Degner, Barbara E Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–772, April 2010.

[11] Terry Therneau, Beth Atkinson, and Brian Ripley. *rpart: Recursive Partitioning and Regression Trees*, 2014. R package version 4.1-8.

[12] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.