Gene expression analysis of HCMV latent infection

Brian Hie and Seokho Hong
brianhie@stanford.edu and seokho@stanford.edu

Introduction

Human cytomegalovirus (HCMV) infects a large percentage of humans worldwide [1]. In most individuals, HCMV establishes a largely asymptomatic latent infection. It can, however, lead to serious disease in immunocompromised individuals, and HCMV infection of the fetus is a common cause of congenital defects. Because it produces relatively few symptoms, HCMV latency was largely thought to be quiescent; only recently did evidence emerge supporting a highly active latent infection [1].

To further investigate HCMV latency, our study takes advantage of relatively recent improvements in high throughput RNA-Sequencing technology, which has enabled a quantitative and reasonably confident measure of gene expression across large populations of individuals. Combining RNA-Seq data with environmental variables such as HCMV status can be used to explore the relationship between environment and gene expression. Using this gene expression data, our study applies machine learning methods that attempt to predict HCMV infection and to better understand patterns of gene expression during HCMV latency.

Data

Our analysis was performed on a data set of 908 tissue samples from 106 individuals obtained through the Genotype-Tissue Expression (GTEx) project [2]. Each sample has gene expression data in the form of RNA-Seq reads, which were mapped to genes using common gene annotations (UCSC, Ensembl, RefSeq) for a total of 33,407 genes and transcripts. Raw read counts were normalized according to RPKM (Reads Per Kilobase of gene per Million total reads). We have metadata on each individual's sex, age, race, BMI, hypertension status, and sample tissue type.

We also have information on HCMV status as determined by an ELISA antibody assay for each sample, also from the GTEx project. 54/106 individuals were HCMV+ and 451/908 tissue samples were HCMV+, giving us a relatively balanced data set.

Baseline

For a simple baseline, we predicted HCMV status using RPKM with a feature vector consisting of all 33,407 genes and a variety of different classifiers. The results are given in Table 1. All F1 scores in this study were computed using 10-fold cross validation (CV).

Table 1: Baseline classification

Classifier	F1 (10-Fold CV)
Logistic Regression	0.495
Linear SVM	0.604
Random Forest	0.491

Feature selection

We next performed filter feature selection based on significant association between gene expression and HCMV status. To do so, we correlated RPKM and HCMV status and computed a permutation based p-value on this correlation. We then played with different p-value cutoff parameters to vary the number of gene features we wanted to keep. A cutoff of $p < 0.05$ gave us a list of 226 genes, $p < 0.067$ gave us 1220 genes, and $p < 0.075$ gave us 3766 genes. Table 2 summarizes the results from classifying on the feature vectors of filtered genes.

Table 2: Classification after filter feature selection.

Number of Genes	Classifier	F1 (10-Fold CV)
226	Linear SVM	0.611
226	Nonlinear SVM (Poly2 kernel)	0.665
1220	Linear SVM	0.710
3766	Linear SVM	0.723

We reasoned that the full set of 33,407 genes caused the SVM to overfit, resulting in high training set error but low test error. Reducing the number of gene features using the above association heuristic allowed the classifier to perform much better.

The non linear SVM result on the 226 genes also shows that there is information to be captured by non-linear interactions between genes. As the number of features grows, however, using non-linear kernels (even a second-order polynomial) causes overfitting on the training set and resulted in poor results for 1220 and 3766 genes.

Noise reduction

To improve the performance of our classifier, we next tried to reduce noise in the gene expression data by regressing out effects from known factors assumed to be independent of HCMV. Known factors in this case were age, sex, race, gender, hypertension status, tissue type, and individual. Removing these known covariates would ideally then increase our power to detect HCMV related signal from the gene expression data.

We used two techniques to regress out known covariates. We first tried an Ordinary Least Squares (OLS) linear model. For each gene, we would fit the OLS model over known covariates using RPKM for that gene as the target vector. The second technique we tried was PEER (Probabilistic Estimation of Expression Residuals) [3] [4], which uses a more sophisticated Bayesian linear regression to model the relationship between known factors and gene expression. A gamma prior is placed on the weights of each factor, which attempts to drive unused factors' weights to 0. Parameter estimation for PEER is implemented based on the Expectation Maximization algorithm. For both models, residuals were computed as the difference between observed and expected expression, which would ideally contain stronger HCMV related signal. Both of these methods assume the covariates have a linear effect on gene expression.

There was a substantial improvement in the F1 scores of the Random Forest (RF) classifier when trained on residuals from both methods of noise reduction. On RPKM alone, the RF classifier had an F1 of 0.49 (Table 1). After either OLS or PEER, the RF classifier had F1 scores greater than 0.9 (Table 3). When trained on residuals, other classification algorithms performed slightly better than baseline, but less well than RF.

Table 3: Classification after noise reduction

Noise Reduction Method	Classifier	F1 (10-Fold CV)
Ordinary Least Squares	Random Forest	0.963
PEER	Random Forest	0.979
PEER	Linear SVM	0.620
PEER	Nonlinear SVM (Poly2 kernel)	0.587
PEER	Logistic Regression	0.528

Discussion

The raw RPKM data used for the baseline contained some information, but not enough for most classifiers to make use of it. Careful feature selection and tuning did increase F1 score, but the model was still very weak.

Noise reduction through regressing out known covariates slightly improves the linear SVM model that was the most successful from training on raw data. It is clear based on the results of random forests that nonlinearities are necessary for modeling with gene expression data.

Fitting SVMs with nonlinear kernels proved difficult on both the raw and processed data, mostly because they tended to overfit on the high dimensional data. Using a non-linear kernel greatly increases the dimensionality of the features, and it becomes very difficult to train a good model without overfitting.

Random forest does much better because it does not have to expand the dimensionality of the feature space to capture the nonlinearities of the data. Furthermore, random forests can capture what might only be otherwise possible with a higher order kernel, while SVMs with a kernel more complex than a second-degree polynomial overfit and simply do not perform well on this data.

Another reason why random forest outperforms nonlinear SVMs is that gene frequency data is best modeled by the decision boundaries that characterize decision trees. Genes have fairly clear, but complex relationships with a limited set of other genes. This is not modeled well by an SVM that takes all the features into consideration at the same time. Decision trees are much better at capturing local complexities given the high dimensional data.

Future Directions

The natural extension of this project would be to gain insight into which genes best indicate HCMV status. We attempted to do this with the RF classifier using feature importances, but the most important genes for the RF classifier were not significantly enriched for viral functions. Using feature importances of the decision tree model weakly identified some viral genes, but not to a satisfactory extent. Although linear models would arguably offer greater interpretability, none of them could classify the dataset accurately enough.

The best solution would be to develop a procedure for increasing the interpretability of random forests. Possible methods might include backward feature selection that narrows the gene feature set to include the most predictive genes.

References

[1] J. H. Sinclair *et al*., “Human Cytomegalovirus Manipulation of Latently Infected Cells,” *Viruses*, vol 5, no. 11, pp 2803-2824, Nov. 2013.

[2] J. Lonsdale *et al*., “The Genotype-Tissue Expression (GTEx) Project,” *Nature Genetics*, vol. 45, pp. 580-585, May 2013.

[3] O. Stegle *et al.*, “A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies,” *PLOS Comp. Bio.*, vol. 6, no. 5, pp. 1-11, May 2010.

[4] L. Parts *et al.*, “Joint Genetic Analysis of Gene Expression Data with Inferred Cellular Phenotypes,” *PLOS Genetics*, vol. 7, no. 1, pp. 1-10, Jan. 2011.

[5] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *JMLR*, vol. 12, pp. 2825-2830, 2011.