Kimberly McManus
December 13, 2013
CS 229 – Project Write-up

TITLE:
Distinguishing pathogenic vs. neutral nucleotide variants in Cystic Fibrosis and the CFTR gene

INTRODUCTION:
A major goal in genetics today is to predict and catalog whether specific single nucleotide variants (SNPs) in the genome are pathogenic or harmless. This is useful because it will allow doctors to immediately know if a patient is likely to get a disease, just by looking at the patient's genome sequence.

Cystic Fibrosis (CF) is one disease that these prediction methods are particularly important for, due to the severity of the phenotype and the high frequency of the disease. About 1 in 2,500 Europeans are diagnosed with Cystic Fibrosis and about 1 in 25 are carriers for the disease (Ratjen & Doring 2003). The disease is characterized by the production of thick, sticky mucus due to abnormal sodium and chloride transport across cells. It is known that Cystic Fibrosis is caused by mutations in the CFTR (Cystic Fibrosis Transmembrane Conductance Regulator) gene, which regulates this sodium and chloride transport. One particular mutation in the CFTR gene, delta508, is the disease cause in greater than 90% of patients (Bobadilla et al. 2002). However, it is known that other mutations in this gene also cause Cystic Fibrosis.

A variety of SNP classifiers currently exist, but they generally have precision and recall values around 0.8. To use these classifiers in a clinical environment, it is essential that these values are much higher. Current classifiers utilize a variety of methods for their predictions, and the goal of this project is to draw on their different strengths to develop an improved classifier to distinguish between pathogenic and neutral alleles in the CFTR gene. In the future, I hope to apply this method to other genes and diseases.

METHODS

DATA SET + PREPROCESSING
The truth data set used for this project is the results of a *Nature Genetics* paper by Sosnay et al. (2012). In this paper, the authors analyzed all SNPs in the CFTR gene with an allele frequency of at least 0.01% in ~40,000 Cystic Fibrosis patients. The authors conducted extensive clinical and functional analyses to determine whether or not each SNP was CF causing. My tests are based on 58 missense mutations. 47 of the mutations are CF causing and 11 are neutral. Missense mutations are point mutations where a single change results in a different amino acid in a protein. As different amino acids have different characteristics, the mutated amino acid may cause the whole protein to be nonfunctional. My goal is to make a classifier that is highly accurate, but relies only on computational features (i.e. does not require data from clinical or functional tests).

I am looking at various non-clinical features: allele frequency in CF patients, allele frequency in the general European population, PolyPhen-2 (Adzhubei et al. 2010), SIFT (Ng et al. 2001), PROVEAN (Choi et al. 2010), PON-P2 (PON-P2 2013), POSE (Masica et al. 2012) and PANTHER (Thomas et al. 2003). These last six features are all mutation classifiers themselves, which estimate how much impact a mutation will have on protein structure. It was my goal to draw on the power of multiple previous classifiers (known to have different strengths) to create an overall better classifier. These previous classifiers are based on various characteristics. For example, PANTHER estimates the evolutionary conservation of that amino acid in various species. It is thought that genome regions that are more conserved between different species are more important to protein function than those that are less conserved. PolyPhen-2 estimates the effect that the mutation will have based on protein structure. It considers features such as the differences in polarity and molecular weight of the two amino acids. The European allele frequencies come from 758 sequenced European chromosomes available through my thesis lab and the frequencies of the mutation in CF patients come from the primary CFTR database (www.cftr2.org). It is thought that CF causing SNPs may be at a lower frequency than neutral SNPs, as natural selection selects against the negative CF causing SNPs. (Note that it is possible that the European allele frequency data includes Cystic Fibrosis patients, but if so, likely at the level of the general population.)

As there are a very limited number of training samples, leave-out-one cross validation was utilized for the test error. There was also missing data in 0.0172 of the PANTHER results and 0.156 of the POSE results. These missing values were filled in with the median of the rest of the values.

MACHINE LEARNING

Testing performance of individual features

First, I measured the success of all previous classifiers that I am using as features (Table 1). Interestingly, these classifiers perform better on this data than they do in standard genome data (Adzhubei et al. 2010). They usually have precision and recall values in the range of 0.75-0.85. In the future, when more data is available, it will be interesting to see the results. Since these classifiers were already quite good, it was a bit difficult to improve upon them.

Initial feature selection & logistic regression

The Mann-Whitney test was used to initially determine significant features. I used this, as opposed to a t-test, because the features were generally not normally distributed. I found that all features, besides allele frequency in CF patients, were significantly different between the neutral and pathogenic mutations. I then normalized the data of the remaining seven features to a mean of 0 and a variance of 1.

The first machine learning algorithm tested was logistic regression (as implemented in R's glm package) with my seven normalized features. I utilized an ROC curve to determine the threshold value of 0.2. This resulted in no training error, but a LOOCV error of 0.086 (Table 2). It also outperforms all of the individual feature classifiers. However, this result does have high variance and is over fitting the data.

Secondary feature selection & logistic regression

In attempt to reduce this error, I conducted more advance feature selection methods. I utilized both forward and backward search to find the subset of model features that minimizes the Akaike's Information Criterion (AIC) of the model. (as implemented in R's step function). This method resulted in five features: allele frequency in European population, Polyphen-2, PANTHER, SIFT and PON-P2.

I reran logistic regression with these five features. Interestingly, the resulting model performed slightly worse and had slightly higher variance than the initial seven feature model. I also noted that the precision was lower that the recall, indicating that logistic regression incorrectly classified more neutral mutations than pathogenic mutations. I hypothesized this may be because most of the training examples are pathogenic mutations.

Weighted, Regularized Logistic Regression

To investigate this problem and attempt to lower the variance, I tried out L1 and L2 regularized logistic regression with weighting. The samples were weighted by the proportion of type in the data set (1/47 for pathogenic, 1/11 for neutral). The cost parameter was determined empirically to minimize the LOOCV error. I found that L1 regularization resulted in a lower LOOCV error than L2 regularization. Thus, the best performing settings for this method was L1 regularization with a cost of 1,000. This method resulted in a model with slightly decreased variance. However, this decrease is due to a slightly increased training error and an equivalent LOOCV error.

Other classifiers tested

Support vector machines (as implemented in R's LiblineaR package) and random forests (as implemented in R's randomForest package) were also explored. For support vector machines, I utilized the same weighting as previously. The cost parameter was chosen empirically to minimize the LOOCV error. L1 and L2 regularization were also examined to minimize LOOCV error. The SVM model that minimized the LOOCV error had a cost of 100. L1 and L2 regularization performed equivalently with this cost value. Unfortunately, the best support vector machine model had a higher training and LOOCV error than the initial logistic regression. However, the precision and recall values in this svm model are equal to each other, indicating that the model is not incorrectly classifying one class more than the other.

The last classifier tested was random forests. I hypothesize that this was a promising algorithm as it is an ensemble method and is it immune to colinearity. As most of my features are classifiers, they are correlated. Unfortunately, random forests performed worse that the other algorithms. I used the default value of 500 trees. I also weighted the classes, similar to previous algorithms. I noticed that, despite the weights, all of the neutral mutations (in the LOOCV) were misclassified. In attempt to improve upon this, I empirically tested increasing the weight of the neutral class. Despite drastic increases in the neutral class's weight, random forests continued to incorrectly classify all of the neutral mutations in the LOOCV.

Classifier	**Precision**	**Recall**	**F-measure**	**% unclassified**
PolyPhen.2	0.88	0.98	0.93	0.05
PROVEAN	0.95	0.79	0.86	0
SIFT	0.90	0.96	0.93	0
PANTHER	0.81	0.91	0.86	0
PON-P2	0.95	1	0.97	0.31
POSE	(Data not avail)	"	"	"

Table 1: This table shows the results of current classifiers (used as features in this project) on the CFTR truth dataset. % unclassified is the percent of the mutations that the classifier returned a value of "unknown".

Data	Method	Training Error	LOOCV error	Precision	Recall	F-measure
7 features chosen by Mann-Whitney test	Logistic Regression	0	0.086	0.938	0.957	0.947
5 features chosen in for./back. search	Logistic Regression	0	0.103	0.918	0.957	0.937
7 features	Logistic Regression, Weighted, C= 1,000, L1 reg.	0.0172	0.086	0.975	0.938	0.947
7 features	SVM, C=100, Weighted, L1 or L2 reg.	0.0172	0.103	0.938	0.938	0.938
7 features	Random Forest, Weighted	0.069	0.190	0.810	1	0.895

Table 2: Results from various machine learning classifiers and data subsets.

FUTURE

From this project, I learned that complex methods are not necessarily better methods. No method I tried outperformed my Wilcoxin rank-sum feature selection and standard logistic regression. With these simple techniques, I achieved increased accuracy compared to the current state-of-the-art mutation classifiers. By drawing on the different strengths of current classifiers I was able to create a classifier that was better all around.

One major caveat to these results is the small number of training examples. Since each mutation has to be validated with extensive clinical and functional assays, it is difficult to get large numbers. There are a few other CFTR mutation databases available, but these are not extensively validated. In the future, I could test my model on this data (with the caveat that some of the truth set might be wrong). I am also in the process of acquiring data for more neutral mutations, though not in time for this project. Furthermore, it would be interesting to look into databases of other mutations. As current classifiers perform better than usual on this dataset, it may be useful to look into datasets that have more room for improvement.

(Almost all of the work and thought about this project was conducted by me. The only help I received was a brief chat with Suyash Shringarpure, a post doc in my advisor's lab (Carlos Bustamante in Genetics).

REFERENCES

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. *Nature Methods* 2010. 7(4):248-249.

Bobadilla JL, Macek Jr M, Fine JP, Farrell PM."Cystic fibrosis: a worldwide analysis of CFTR mutations--correlation with incidence data and application to screening". *Human Mutation* 2002. 19 (6): 575–606.

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS ONE* 2012. 7(10): e46688.

Masica D, Sosnay PR, Cutting GR, & Karchin R. Phenotype-optimized sequence ensembles substantially improve prediction of disease-causing mutation in cystic fibrosis. *Human Mutation.* 2012 33(8):1267-74.

Ng PC, Henikoff S. Predicting Deleterious Amino Acid Substitutions. *Genome Res.* 2001.11(5):863-74.

PON-P2 Server. Retrieved 10 Nov 2013. http://structure.bmc.lu.se/PON-P2

Ratjen F, Doring G. Cystic fibrosis. Lancet. (2003); 361(9358):681-9.

Sosnay PR et al. Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. *Nature Genetics*. 2013. 45:101160-7.

Thomas PD, Campbell ML, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, and Narechania A. PANTHER: A library of Protein Families and Subfamilies Indexed by Function. Thomas *Genome Res.* 2003. 13: 2129-2141