The International Young Researchers' Conference
DOI: doi.org/10.34614/iyrc0040

Classifying allergic rhinitis subjects and identifying single nucleotide polymorphisms using a support vector machine approach

Jason Chan
Poway High School, United States
jasontc04@gmail.com

Abstract

Allergic rhinitis is a common respiratory disease that affects a large proportion of the population and is associated with a loss of work productivity and economic losses. There has been known to be a genetic link in the onset of allergic rhinitis, so we aimed to identify correlated SNPs using a novel support vector machine (SVM) method. We gathered our genetic data from a publicly available database and one-hot encoded the SNP files. Then, we created sparse matrices to reduce random access memory (RAM) and ran a SVM to classify individuals on the basis of allergic rhinitis, as well as identify key SNPs. Our model achieved moderately high accuracy/macro F1 score and identified 736 genome-wide significant SNPs. Analyzing these SNPs further, we found a common gene associated with many of the discovered allergic rhinitis-associated SNPs. This study furthered the knowledge in understanding the onset of allergic rhinitis and introduced using SVMs in analyzing the genetic implications of allergic rhinitis.

Keywords: Allergic Rhinitis, Support Vector Machine, Single Nucleotide Polymorphisms

Introduction

Allergic rhinitis, also known as hay fever or seasonal allergies, is a predominantly mild health condition that induces symptoms similar to that of the common cold (Wheatley & Togias, 2015). Inflammation of the nasal canal, as well as the lower respiratory system (due to the many functional relationships between the nasal canal and the lower respiratory tracts), is prevalent in a patient when one is experiencing symptoms (Small & Kim, 2011). Individuals with allergic rhinitis experience these symptoms when exposed to an allergen, such as pollen or dust (Varshney & Varshney, 2015). The body naturally releases histamine, which can trigger allergic rhinitis (Church & Church, 2016). Adults with allergic rhinitis account for around 10-30% of the adult population in the United States, making allergic rhinitis the fifth most common chronic disease (Tran, Vickery, & Blaiss, 2011). Furthermore, there are many economic losses associated with allergies. In 2005 alone, there was an estimated $11.2 billion cost for medical bills for allergic rhinitis (Tran, Vickery, & Blaiss, 2011). Decreased productivity in workers accounts for 3.5 million lost workdays, making allergic rhinitis the fifth costliest disease as well (Tran, Vickery, & Blaiss, 2011).

The causes of allergic rhinitis are numerous, however many experts attribute it to a combination of an individual's environment and genetics. While this paper will not delve into the environmental variables of seasonal allergies, the genetic basis of allergies is still highly influential in

determining its incidence. Several studies have displayed the extensive strength of the role of genetic inheritance. One study found that monozygotic twins showed a concordance of 45-60% probability of the onset of allergic rhinitis while dizygotic twins had a 25% rate of correlation (Tran, Vickery, & Blaiss, 2011). Atopy is a trait among individuals that helps describe the incidence of allergy-based diseases in individuals (Bellanti & Settipane, 2017). Some papers imply that atopy is both genetic and environmental, which assist in displaying the intertwined relationships of both genetic and environmental factors (Wang, 2005). The current exploration of the genetics of allergies is still unclear, as concrete genetic evidence is scarce. Several genes are known to be correlated with allergic rhinitis; however, their contribution to the disease is not fully understood. Using a novel computational method, support vector machines (SVMs), we aim to discover important loci associated with allergic rhinitis. In doing so, we hope to be able to close the gap in identifying genetic evidence for the onset of allergic rhinitis.

Review of Literature

Currently, many of the genes discovered by scientists known to influence allergic rhinitis are only weakly correlated and their explicit function in determining the onset of allergic rhinitis is largely unknown. Recently, though, through genome-wide association studies (GWASs), several loci have been discovered. A novel GWAS conducted in 2018 identified 20 new loci that could possibly be associated with allergic rhinitis (Waage et al., 2018). However, some of these results have yet to be replicated and confirmed by other researchers. GWAS is the most common method for discovering genetic variants as it is an effective way to discover single-nucleotide polymorphisms (SNPs) and different alleles throughout the entire genome of an individual (as opposed to focusing in on a specific chromosome or part of a chromosome) (Bush & Moore, 2012).

Another recent study has shown a link in the onset of allergic rhinitis along with asthma and eczema. Ferreira et al. (2020) deployed a GWAS discovering 76 possible genetic variants associated with a grouping of the 3 traits. 18 of the 76 identified were novel findings in the field, while the other findings helped to solidify genetic evidence for allergic rhinitis, asthma, and eczema (Ferreira et al., 2020). The group also looked at the correlation of the phenotype along with the age at which the subjects attained the allergenic disease (Ferreira et al., 2020). In another study, Forno et al. (2012) focused on the early age of onset of asthma in which GWAS was able to be effectively utilized in discovering SNPs.

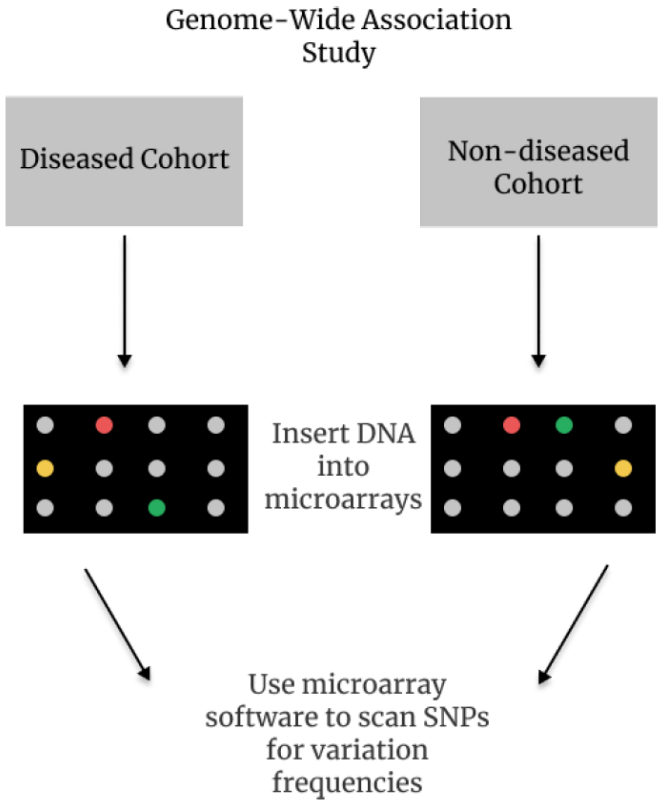

FIGURE 1. The process of a GWAS starts with identifying two cohorts-one group with the disease of interest, and one without. Next, the unknown DNA gets injected into a microarray, typically made of glass, and the microarray's nucleotide sequences act as probes to attach onto the injected DNA. Lastly, a microarray software scans the microarrays for SNP patterns in each subject's DNA chip.

However, other researchers have taken other approaches using more machine learning-based methods. So far, these methods have been utilized in detecting other diseases such as asthma and autism. Pandey et al. (2018) used a combination of machine learning techniques such as linear regression, SVM-linear, and random forest to detect nasal biomarkers to classify individuals for diagnosis of asthma. Similar machine learning methods were employed in a study to identify subjects with Autism Spectrum Disorder (ASD), as well as the genes associated with the disorder (Asif, Martiniano, Vicente, & Couto, 2018). There has yet to be a study on utilizing SVM on classifying the loci associated with allergic rhinitis.

Purpose

Our first objective was to create a model to perform binary classification on a subject given their SNP data. Even with genetic data, we aimed to produce a model with a high macro F1 score and accuracy. Additionally, we intended to use the SVM model to determine which genes influence the onset of allergic rhinitis, as well as the weight of these genes.

Data

Dataset

We retrieved our SNP data from the publicly-available database, The Personal Genome Project. Established in 2005 at Harvard University, The Personal Genome Project (PGP) spans 5 countries, each with their own respective database. As each database consisted of varying degrees of information, we extracted data only from the Harvard PGP group.

The SNP data for each individual were from various labs. We only used the data generated from the 23andme lab in order to have a consistent format. Each individual file consisted of four features: rsID, chromosome number, position, and genotype. The genotype column was composed of two alleles on the forward strand of DNA, one from each parent.

To create our experimental and control groups from this database, we analyzed surveys that were self-reported by PGP participants. We designed our experimental group from the 2,128 responses to the PGP Trait and Disease Survey 2012: Respiratory System. We extracted the control group from the "whole genome sequences and other data" section, rejecting any matching experimental subjects chosen from the survey. Our control and experimental groups consisted of both male and female participants, and the SNP data submission dates ranged from January 2011 to July 2020. Our final experimental group consisted of 99 files, while our control group consisted of 102 files.

Preprocessing

In order to utilize a machine learning model, we had to encode our categorical data into vectors in a machine-readable format. This process is done in many different machine learning cases when dealing with categorical data. Models such as convolutional neural networks (CNN) and logistic regression are often used with some form of categorical variable encoding.

As encoding genetic data has become popular in the machine learning field, there have been different methods that researchers have proposed to encode large amounts of data. In order for a program to categorize disorders influenced by genetics, the nucleotide sequence must be preprocessed and encoded. Nguyen et al. (2016) proposed grouping nucleotides in windows of 3 and one-hot encoding vectors of length 64. Another study introduced the usage of ordinal encoding in genetic data where the researchers compared the performance between one hot encoding and ordinal encoding (Choong & Lee, 2017). While one-hot encoding is more popular, it requires vectors and matrices to be restricted to certain dimensions (Choong & Lee, 2017). It was found that ordinal encoding required less memory and faster training speeds (Choong & Lee, 2017).

While vectorization can take many different forms, the method we used was one-hot encoding. One-hot encoding first denotes a unique integer to every possible value. In our data, these encoded values were the allele pairs. Next, a zeroed vector of length n—where n is the number

of unique allele pairings—was constructed for each SNP. Our SNP vectors each had a length of 27. In each vector, a 1 was placed in the index that was associated with the respective allele combination. Lastly, in a zeroed matrix (shape d × n) with 2,050,307 rows, each row corresponding to a unique rsID, we inserted the SNP vector in the row's index unique to the associated rsID of the SNP. We created 201 unique matrices for each individual SNP data file. These 201 matrices were split by a 80:10:10 split for the train, validation and test set respectively.

As our matrices consisted mainly of zeros, we transformed our 2-D Numpy tensors into sparse matrices. This was done by utilizing SciPy's sparse class.

Next, we concatenated each matrix together, grouping the experimental and control group separately as to segregate the two classes for ease of labeling purposes. In order to avoid concatenating our 2-D matrices (2050307 × 27) together to create a 3-D tensor (201 × 2050307 × 27), we instead compressed our 201 2-D matrices into 1-D vectors (1 × 55358289), and combined the 1-D vectors to create a single 2-D matrix (201 × 55358289), comprised of all 201 subject SNP files.

Methodology

Model

The main machine learning method we used was the support vector machine (SVM). The SVM is a machine learning classification tool used to separate classes and designate a hyperplane that maximizes the distance between two classes. This hyperplane is created through examining support vectors-data points that lie the closest to the boundary where the classes meet-and finding the maximum margin between the distinguished classes. The hyperplane is defined as follows:

$$w \times x_i + b = 0$$

We were able to find the optimal hyperplane that categorized subjects on whether or not they were likely to contract allergic rhinitis, as well as several SNPs associated with allergic rhinitis as defined by passing the genome-wide significance threshold of $p\ value = 5 \times 10^{-8}$.

Our training data consisted of a number of subjects diagnosed with allergic rhinitis, as well as a control group. The tested model was able to classify different individuals on the basis of having allergic rhinitis or not. Additionally, we aimed to create a model that outputs the highest accuracy possible. In terms of individual classification, the support vector machine found the optimal hyperplane in which to separate the two classes. These two classes were our control and experimental group. A benefit of using support vector machines is that they are effective in analyzing high dimensional data, even if the number of dimensions outnumbers the number of samples. As there are a high number of dimensions (upwards of a million) in our model, this classification task is feasible due to the nature of support vector machines.

Once our model was able to adequately classify each subject, we dissected the model to discover the most influential SNPs. This was done by observing the coefficients of each feature—the SNP—in our model as shown below:

$$\hat{Y} = \min_{\beta_0,\beta} \sum_{i=1}^{n} [1 - y_i(\beta_0 + x_i^T \beta)]_+ + \frac{\lambda}{2}\|\beta\|_2^2$$

The prediction model seeks to minimize each weight to fit the data optimally. Additionally, we used L2 regularization to further minimize the weights, and only focus on the heavy weight values. This method proved to result in a higher accuracy than L1 regularization, displaying the notion that many weights rather than fewer are crucial to categorizing allergic rhinitis. After L2 regularization with an inverse regularization strength of 1, we observed the highest coefficients from our model to determine which SNPs considerably influenced allergic rhinitis.

We introduced a new method to discovering SNPs associated with allergic rhinitis by dissecting each feature's weight in our model to analyze the highest impact weights. We found the weights with the highest absolute value and assuming a normal distribution, we took a two-tailed z-test to find each weight's p value. This revealed not only the SNPs that positively influence the onset of allergic rhinitis, but also revealed the SNPs that negatively influenced the incidence of allergic rhinitis.

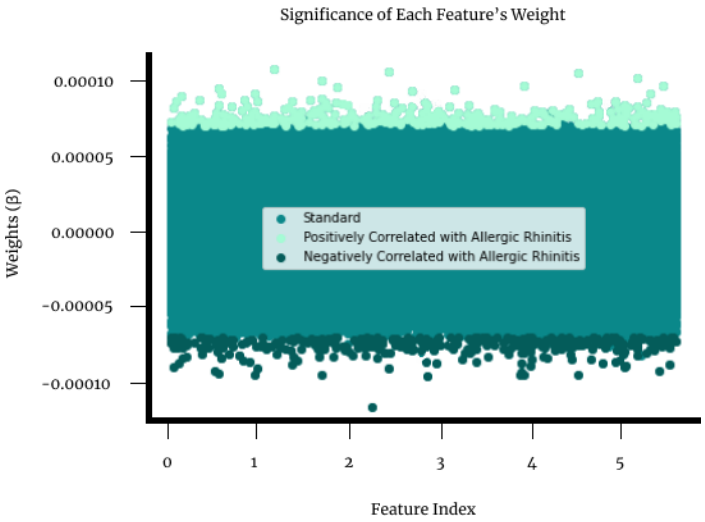

FIGURE 2. Our study utilized the SVM's weights to analyze important SNPs. We took the absolute value of each weight to find both positive and negative correlations. The figure contrasts each weight by displaying whether the identified SNP was positively or negatively correlated with allergic rhinitis, as well as to what extent.

Results and Discussion

Our final model achieved a macro F1 score of 0.74 on a relatively small test set of 23 data points (12 experimental and 11 control). Additionally, the model's accuracy was also 0.74. We experimented with several values of C, however, they each gave us the same macro F1 score and accuracy. The values we tested for C were 0.001, 0.01, 1, 100, and 1000. We also experimented with a radial basis function kernel. With values of C=10 and C=100, we received a macro F1 score of 0.73 and an accuracy of 0.74. However, with a value of C=1, our macro F1 score dropped to 0.62 with an accuracy of 0.65.

In terms of SNP discovery, we found a total of 736 genome-wide significant (p value $< 5 \times 10^{-8}$) SNPs. 381 of the significant SNPs positively correlated with allergic rhinitis, while 355 of the genome-wide significant SNPs negatively correlated with allergic rhinitis. We looked at the 20 most prominent positively influential SNPs and matched each with a gene, if applicable. Out of these 2 SNPs, four of them were associated with the gene EVI5 (rs10735781, rs6680578, rs11808092, and rs4847267). The EVI5 gene was the most prominent result in our findings in analyzing the top 20 positively influential SNPs. Other genes correlated with these 20 SNPs were PALM2AKAP2 (rs1980874), LOC105370922 (rs7163642), TMEM132C (rs4882801), ARHGAP35 (rs10425259), ZP3 (rs3789833), and OPCML (rs12418625). 10 of the examined SNPs did not have any close relation to a specific gene.

Our results show a grouping of influential SNPs on the EVI5 gene, which provides evidence towards the conclusion that there is a link between the gene and the incidence of allergic rhinitis. The other genetic variants discovered may also play a role in the pathophysiology of allergic rhinitis or a similar respiratory disease. Unfortunately, due to the nature of self-reported data, these surveys cannot be guaranteed to be completely accurate, as the individual could have withheld information or reported false information. Similarly, some participants may have chosen not to answer the survey, so in creating our control group, we cannot be 100% confident that every control individual was aseptic to allergic rhinitis. These imprecisions are a possible explanation for our moderate model accuracy.

TABLE 1. We examined 44 of the most prominent SNPs, including negatively correlated SNPs as well. The highest p value was from a negatively correlated SNP with a value of 1.34E-19. We display the rsID of the SNP, the nucleotide bases of the SNP and each SNPs p value. Additionally, the last column displays whether the corresponding weight from the SVM was a positive or negative influence.

rsID	genotype	p value	+/-
rs303727	TA	1.34E-19	Negative
rs10821080	T	6.15E-17	Positive
rs10735781	ID	1.56E-16	Positive
rs1980874	I	3.03E-16	Positive
rs7163642	GT	2.76E-15	Positive
rs6680578	TC	6.08E-15	Positive
rs4882801	G	4.99E-14	Positive
rs10425259	AT	6.89E-14	Positive
rs7828114	CT	9.02E-14	Negative
rs12913	AT	1.07E-13	Negative
rs2211938	AT	1.15E-13	Positive
rs6730045	GG	1.43E-13	Negative
rs1040411	CC	1.58E-13	Negative
rs4633807	TA	1.86E-13	Positive
rs9951150	GT	2.25E-13	Positive
rs2354178	GG	2.54E-13	Negative
rs1801274	GT	3.92E-13	Negative
rs7526587	AA	5.13E-13	Positive
rs7158744	GA	5.32E-13	Negative
rs958898	TT	7.24E-13	Negative
rs11808092	CA	9.56E-13	Positive
rs2415290	TT	1.12E-12	Positive
rs550915	CA	1.41E-12	Negative
rs6604026	GC	2.48E-12	Negative
rs337277	D	3.17E-12	Negative
rs4847267	A	3.75E-12	Positive
rs6733711	GG	4.12E-12	Negative
rs11120923	CA	4.93E-12	Negative
rs4078690	T	7.36E-12	Negative
rs198178	CC	7.76E-12	Negative
rs9612352	ID	8.49E-12	Negative
rs11764618	TG	8.78E-12	Negative
rs1522679	AT	9.29E-12	Negative
rs6583565	DD	9.57E-12	Positive
rs6896806	--	1.03E-11	Negative
rs3789833	C	1.11E-11	Positive
rs6690126	AC	1.17E-11	Positive
rs12418625	AC	1.30E-11	Positive
rs9450450	AA	1.32E-11	Positive
rs587771	T	1.35E-11	Negative
rs4502845	TG	1.44E-11	Negative
rs11666652	CC	1.49E-11	Negative
rs2825493	TC	1.88E-11	Positive
rs1483578	GG	2.03E-11	Negative

Nevertheless, the novel use of an SVM in gene detection took advantage of machine-learning capabilities to assign SNPs to features of our model and predict associations. While the current optimal method to identify loci in modern genetic studies is genome-wide association studies, the use of an SVM has benefits over a GWAS. Primarily, SVMs are able to function moderately accurately with a relatively small sample size. Most GWAS are performed with thousands of experimental subjects, with an equal sized control group. However, this study managed to draw conclusions with only 201 subjects.

Naturally, the performance of our SVM model and the validity of our identified SNPs would improve with a larger sample size. However, due to the limited size of our database, we were unable to increase the batch size for our model. Despite this limitation, we were still able to prove the effectiveness of a SVM.

Conclusion

Our model achieved a moderately high accuracy and was able to identify crucial SNPs that characterized the onset of allergic rhinitis. While our macro F1 score may be modest in the machine learning field, we think that our SVM model is robust due to the multifactorial nature of the genetic basis for disease. Our work can be expanded further and would benefit from a larger sample size to increase the accuracy of our model and our findings. Nonetheless, our SNP findings may lead to understanding the implications of the pathways and functions that each gene serves in the development of allergic rhinitis. Lastly, our introduction to using SVMs and machine-learning methods that rival GWAS studies may allow for researchers to discover certain SNPs without having to collect such a large cohort of subjects.

Acknowledgments

I would like to thank my mentor, Jerry Qu, for guiding me through the research process. Additionally, I would like to thank The Summer Stem Institute for giving me this opportunity to perform extensive data science research.

References

Asif, M., Martiniano, H. F., Vicente, A. M., & Couto, F. M. (2018). Identifying disease genes using machine learning and gene functional similarities, assessed through gene ontology. *PLOS ONE, 13*(12). doi:10.1371/journal.pone.0208626

Bellanti, J. A., & Settipane, R. A. (2017). The atopic disorders and atopy … "strange diseases" now better defined! *Allergy and Asthma Proceedings, 38*(4), 241-242. doi:10.2500/aap.2017.38.4074

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS Computational Biology, 8*(12). doi:10.1371/journal.pcbi.1002822

Choong, A. C., & Lee, N. K. (2017). Evaluation of Convolutionary neural networks modeling of DNA sequences using ordinal VERSUS one-hot encoding method. *2017 International Conference on Computer and Drone Applications (IConDA)*. doi:10.1109/iconda.2017.8270400

Church, D. S., & Church, M. K. (2016). Allergic rhinitis: Impact, diagnosis, treatment and management. *Clinical Pharmacist, 8*. doi:10.1211/cp.2016.20201509

Ferreira, M. A., Vonk, J. M., Baurecht, H., Marenholz, I., Tian, C., Hoffman, J. D., . . . Koppelman, G. H. (2020). Age-of-onset information helps Identify 76 genetic variants associated with allergic disease. *PLOS Genetics, 16*(6). doi:10.1371/journal.pgen.1008725

Forno, E., Lasky-Su, J. A., Ramsey, C. D., Brehm, J., Klanderman, B. J., Ziniti, J. P., . . . Celedon, J. C. (2010). Genomewide Association study of age of onset in childhood asthma. *B93. GENOME-WIDE APPROACHES FOR THE IDENTIFICATION OF ASTHMA GENES: NEW METHODS, NEW PHENOTYPES, NEW GENES*. doi:10.1164/ajrccm-conference.2010.181.1_meetingabstracts.a3729

Nguyen, N. G., Tran, V. A., Ngo, D. L., Phan, D., Lumbanraja, F. R., Faisal, M. R., . . . Satou, K. (2016). DNA sequence classification by convolutional neural network. *Journal of Biomedical Science and Engineering, 09*(05), 280-286. doi:10.4236/jbise.2016.95021

Pandey, G., Pandey, O. P., Rogers, A. J., Ahsen, M. E., Hoffman, G. E., Raby, B. A., . . . Bunyavanich, S. (2018). A nasal Brush-based classifier of Asthma identified by machine learning analysis of NASAL RNA sequence data. *Scientific Reports, 8*(1). doi:10.1038/s41598-018-27189-4

Small, P., & Kim, H. (2011). Allergic rhinitis. *Allergy, Asthma & Clinical Immunology, 7*(S1). doi:10.1186/1710-1492-7-s1-s3

Tran, N. P., Vickery, J., & Blaiss, M. S. (2011). Management of rhinitis: Allergic and non-allergic. *Allergy, Asthma and Immunology Research, 3*(3), 148. doi:10.4168/aair.2011.3.3.148

Varshney, J., & Varshney, H. (2015). Allergic rhinitis: An overview. *Indian Journal of Otolaryngology and Head & Neck Surgery, 67*(2), 143-149. doi:10.1007/s12070-015-0828-5

Waage, J., Standl, M., Curtin, J. A., Jessen, L. E., Thorsen, J., Tian, C., Schoettler, N., 23andMe Research Team, AAGC collaborators, Flores, C., Abdellaoui, A., Ahluwalia, T. S., Alves, A. C., Amaral, A., Antó, J. M., Arnold, A., Barreto-Luis, A., Baurecht, H., van Beijsterveldt, C., Bleecker, E. R., … Bønnelykke, K. (2018). Genome-wide association

and HLA fine-mapping studies identify risk loci and genetic pathways underlying allergic rhinitis. *Nature genetics*, *50*(8), 1072–1080. https://doi.org/10.1038/s41588-018-0157-1

Wang, D. (2005). Risk factors of allergic rhinitis: Genetic or environmental? *Therapeutics and Clinical Risk Management, 1*(2), 115-123. doi:10.2147/tcrm.1.2.115.62907

Wheatley, L. M., & Togias, A. (2015). Allergic rhinitis. *New England Journal of Medicine, 372*(5), 456-463. doi:10.1056/nejmcp1412282