Cilantro: A Divisive Flavor, Genetically Speaking

Gavin Mischler
The Johns Hopkins University

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

Foundations of Computational Biology and Bioinformatics Spring 2020

Summary

Cilantro, despite being a popular herb in cooking, tastes terrible to some people, reminding them of soap. This investigation analyzes the genetic variants that are related to this phenotype and uncovers several potential genetic features which are highly predictive of it. Supervised learning methods are utilized to prune the numerous genetic variants within genes of interest down to a small number of highly predictive features. Statistical testing verifies the relation of these genetic features to the phenotype in the study population, and a validation cohort collected from a different source provides more evidence for the link between these genetic variants and the phenotype.

Background

Although cilantro is used in many cuisines, its taste can be highly divisive, since some people say it tastes like soap, or even that it smells like bed bugs. This smell is attributed to the natural aldehydes present in the herb. Aldehydes are also produced in the soap-making process, as well as by some insects. It is believed that some people experience these aldehydes in the same way due to altered taste- and olfactory-receptors in their mouth and nose, which normally are responsible for distinguishing these subtly different aldehydes [1]. Two studies have identified evidence of a total of 4 genetic variants which may be highly linked to the trait [1, 2]. However, further work is needed since the genetic mechanisms responsible for this trait are not completely understood. Also, due to the interconnected genes, it is unlikely these variants act alone. There are likely more genetic features which are predictors of the phenotype, and which may be important factors in determining a person's likelihood of thinking cilantro tastes like soap.

Methods

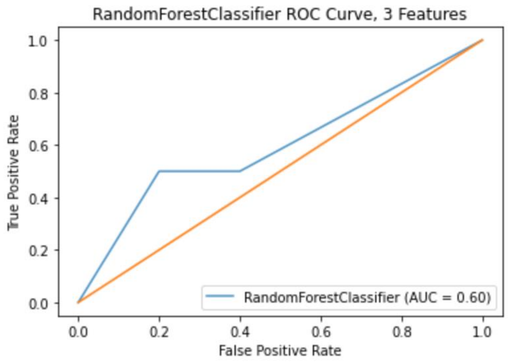
Publicly available genomic data was downloaded from openSNP to create a study population of 20 subjects who think cilantro tastes like soap, and 23 who don't. This genomic sequencing data was annotated using OpenCRAVAT, and then information about the genetic variants in each subject's genome was extracted. First, information about the existence of the 4 variants from the literature in the patients were extracted. Only 3 were present, so these 3 features were used to train a predictive model of the phenotype.

Next, information about a much larger set of variants from an extensive list of taste- and olfactory-receptor genes [3, 4] was extracted from each subject. Paired with zygosity information for each of these variants, this led to 7107 genetic "features" for each subject. A 1-norm regularized linear SVM model was used to identify important features from this large set. 1-norm regularization leads to models which use a sparse set from the original feature space to perform prediction. The strength of the regularization was continuously increased (leader to sparser models), until the model's accuracy began to drop. At this threshold, the 8 features being used by the model were identified for further study. Statistical tests confirmed the importance of these features in the study population.

A validation cohort of genetic data from classmates, professors, and their families, was gathered. A model trained on the 8 features identified previously was tested for its predictive abilities in the validation cohort.

Results

First Predictive Model



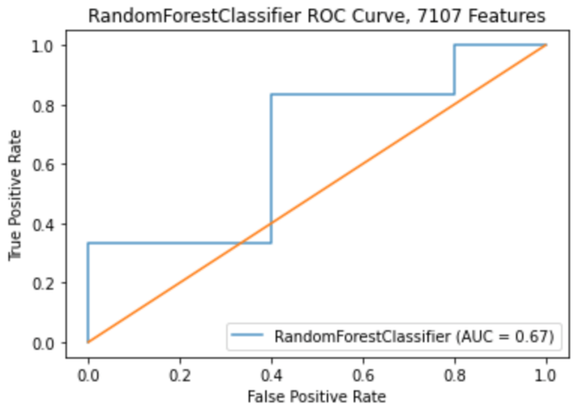
Using the presence of each of the 3 genetic variants from the literature as a feature was not very predictive of the phenotype in the study cohort, with the best model, a Random Forest Classifier, only achieving an AUC of 0.6.

Full Variant Feature Extraction

43 Subjects

rsID: Indicator Matrix for 2369 Genetic Variants in the Olfactory- and Taste-Receptor Genes	rsID_het: Indicator Matrix for Whether Each of the 2369 Variants was Heterozygous or Not	rsID_hom: Indicator Matrix for Whether Each of the 2369 Variants was Homozygous or Not

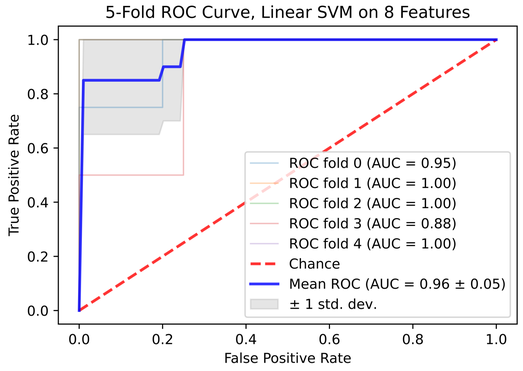
7107 Features



These extra features improved the predictive ability of the classifier, which now achieved an AUC of 0.67, but it still could be improved greatly by pruning unnecessary features.

Feature Pruning

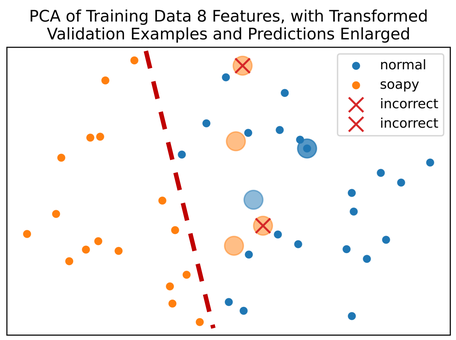
Iteratively, a 1-norm regularized linear SVM was trained on 70% of the data, then tested on the remaining 30%. If it could achieve an AUC of 1.0, the regularization strength was increased. This was done until it no longer achieved an AUC of 1.0, at which point the 8 features it was using from the full feature space were identified and used as features to train models. On 5-fold cross-validation, these 8 features achieved an average AUC of 0.96, showing high predictive ability.



A chi-squared test to examine the dependence between the phenotype and each of the 8 features was performed on the training cohort. P-values are shown below, along with the prevalence in the training and validation populations. The validation cohort was not large enough for chi-squared tests to yield any significant results.

Feature	Train Cohort - Normal Prevalance	Train Cohort - Soapy Prevalance	Train Cohort - pvalue	Val Cohort - Normal Prevalence	Val Cohort - Soapy Prevalence
rs8181529_het	0.696	0.05	6.16E-05	0	0
rs10749643_het	0.522	0.05	0.00247	0.333	0.25
rs10742809_het	0.609	0.15	0.00585	0.667	0.5
rs4237768	0.304	0.65	0.04998	0	0.5
rs238882_het	0.261	0.65	0.02413	0	0
rs7941509_hom	0.130	0.55	0.00926	0	0.5
rs10985704_het	0.391	0.8	0.01640	0.333	0
rs8181529_hom	0.174	0.7	0.00148	0	0.25

A model trained on the 8-feature data from the original subjects was used to predict the phenotype of the 7 validation cohort subjects. PCA components were computed from the 8-feature data of the original cohort, and both this cohort and the validation cohort was transformed to the same PCA space, (small circles are training subjects, large circles are validation subjects). Circles are colored by true phenotype. The 2 incorrectly predicted validation subjects have a red X over their point.



The dashed line indicates a perfect linear separation between the 2 classes (soapy and normal) in the training cohort.

Discussion and implications

The variants identified in the literature were not very informative about the phenotype for the population studied. The predictive ability improved when genetic variants from many other taste- and olfactory-receptor genes were added, potentially indicating inter-relationships between many other genes and the phenotype. However, using all these features was not the best approach, since the low AUC still indicated there were many features which did not help the prediction.

Pruning the huge feature space down to a more manageable number while maintaining very good accuracy with a linear predictive model demonstrated that a select few genetic features were highly correlated with the soapy-cilantro taste. The chi-squared statistical testing further reinforced this, since all 8 features identified had a p-value of less than 0.05 on the study population.

The validation cohort provided promising, but potentially confounded, results. Although 4 of the 7 validation subjects were labeled as "soapy", none of them think cilantro tastes like soap. Rather, they merely hate the flavor, while those who were labeled as normal like the flavor. The predictive model based on the 8 features misclassified 2 of the "soapy" subjects as being normal. On one hand, this could just be because they aren't really "soapy", so it makes sense that the model predicted them as normal. On the other hand though, if they had all been labeled as normal, the algorithm still would have predicted 2 incorrectly.

The PCA figure might often some insight though, since it shows a 2-dimensional subspace where the 2 classes (soapy and normal) are perfectly linearly separable in the training cohort, as evidenced by the small orange and blue data points being separated by the dashed red line. Interestingly, when the validation data points were projected onto these same 2 PCA components, all the points fell on the side of the normal taste. This may indicate that these subjects should all be normal, since none of them taste cilantro like soap, meaning that the points may be mislabeled. This means that the model has found features which are indicative only of cilantro tasting like soap, not cilantro tasting bad. This implies that the soapy-cilantro phenotype has a uniquely genetic basis, which is different from cilantro tasting bad either from a genetic perspective, or because cilantro tasting bad is more of a learned association which cannot be predicted well by genetics. Either way, the genetic basis of taste warrants further investigation to validate the 8 genetic features identified here.

References

1] Eriksson, Nicholas, et al. "A genetic variant near olfactory receptor genes influences cilantro preference." *Flavour* 1.1 (2012): 22.

[2] Knaapila, Antti, et al. "Genetic analysis of chemosensory traits in human twins." *Chemical senses* 37.9 (2012): 869-881.

[3] Bachmanov, Alexander A., and Gary K. Beauchamp. "Taste receptor genes." *Annu. Rev. Nutr.* 27 (2007): 389-414.

[4] Olender, Tsviya, Doron Lancet, and Daniel W. Nebert. "Update on the olfactory receptor (OR) gene superfamily." *Human genomics* 3.1 (2008): 87.