Machine Learning Approaches for Classifying Genetic Variants

Jen Johnson `18
Advised by Professor Michael Linderman
A Thesis for the Middlebury College Computer Science Department

1 What is variant classification?

Variant classification assigns a label to describe a variant's effect on an individual's disease phenotype. A pathogenic variant is disease-causing and obliges a clinical response.

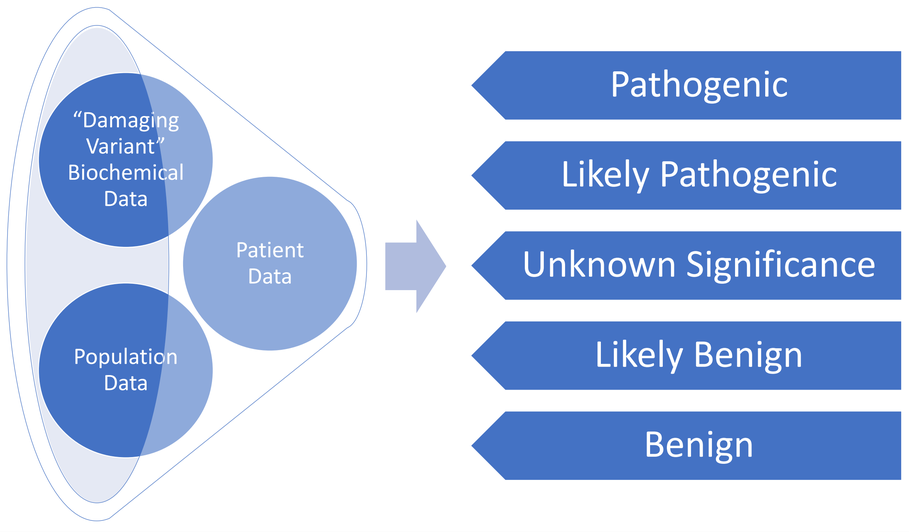
2 American College of Medical Genetics Guidelines

The ACMG guidelines [1]:

1. Evaluate and weight evidence by type. e.g.

Evidence Type	Weight of Evidence
If a variant is present at an *extremely low frequency* in the Exome Aggregation Consortium…	… this provides moderate evidence of pathogenicity.

2. Combine weighted evidence into a classification.



This rule is based on the assumption that:

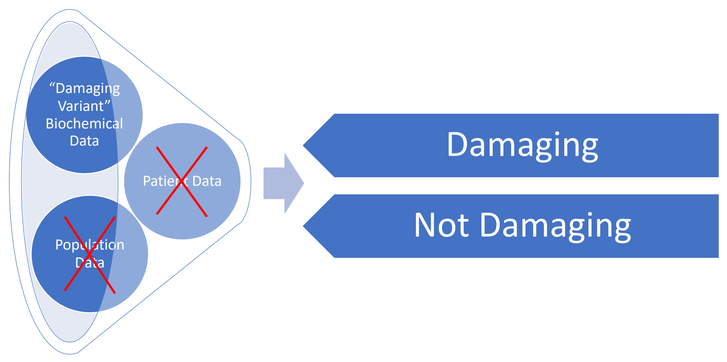
A low frequency variant provides strong evidence of pathogenicity because natural selection acts to eliminate most pathogenic variants from the population.

What is an "*extremely low frequency*?" This ambiguity allows the rules to be adapted to patient-specific data.

3 Machine Learning Approaches

Machine learning approaches can make variant classification more efficient. One existing tool is REVEL. [2] It classifies variants as "damaging," or having an effect on the protein product of a gene. However, "damaging" does not

- oblige clinical treatment
- address population nor patient data
- classify variants as "pathogenic."



My goal is to implement a hybrid classification tool that combines the efficiency of machine learning with geneticists' expertise.

4

I used the framework of Sherloc criteria [3] to redefine the "*extremely low*" allele frequency (AF) and model its relationship with pathogenicity.

Mode of inheritance. An autosomal dominant disease has a lower AF than a recessive disease because natural selection acts on *all* carriers of the variant. Therefore, the AF threshold for classifying dominant variants as pathogenic should be lower than for recessive variants.

Allele number. A measure of the amount of data.

$$\text{Allele Frequency} = \frac{\text{Num Observations}}{\text{Allele Number}}$$

A large allele number provides more confidence in the AF estimate.

5

Probabilistic Soft Logic (PSL). A statistical relational learning tool for inferring knowledge from a logic network. [4]

- Atoms are AF and mode of inheritance observations from variants.
- Predicates, e.g. whether two variants have similar allele frequencies, can take soft truth values [0,1] to encode uncertainty.
- Rules encode relationships between predicates. Their relational structure allows knowledge to be inferred from similar entities in the network.

I created synthetic variant data to serve as a **proof of concept that PSL and machine learning can be applied to variant classification.**

6 PSL Variant Classification

I developed a PSL rule for AF that incorporates allele number indirectly.

$$\text{HasCat}(A, C) \;\&\; \text{HasSimilarAF}(A, B) \;\&\; (A \mathrel{!}= B) \gg \text{HasCat}(B, C)$$

Variant A has Class C — Variant A has a similar allele frequency to Variant B — Variant A is not Variant B — Variant B has Class C

→ HasCat(A, Benign)
→ HasCat(A, Pathogenic)

- I assigned all known HasCat predicates the min-max scaled value of the allele number and therefore a value from [0,1].
- I assigned HasSimilarAF predicates a value of 1 if the 2 variants A and B are in the same allele frequency category, defined by Nykamp *et al.* [3]

I developed a parallel rule for mode of inheritance data.

$$\text{HasCat}(A, C) \;\&\; \text{HasSimilarAD}(A, B) \;\&\; (A \mathrel{!}= B) \gg \text{HasCat}(B, C)$$

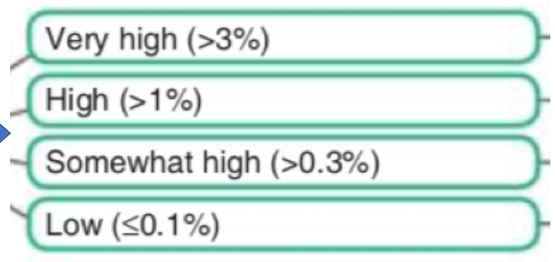
AD: autosomal dominant, as opposed to autosomal recessive.

7 Results With Synthetic Data

Weight Learning

Rule	Weight
Allele Frequency	1.052
Mode of Inheritance	0.027

The AF rule has a much higher weight, as was expected.



Inference Performance

Performance Metric	Value
False Positive Rate	25
False Negative Rate	50
Specificity	75
Sensitivity	50

False Positive Rate: Proportion of benign variants classified as pathogenic. The FPR is a critical metric for evaluating a classifier's accuracy and ensuring that a "pathogenic" label continues to obligate care.

8 Conclusion

I implemented a proof of concept model that PSL and machine learning can be applied to variant classification as a hybrid approach that is both efficient and can encode the complexity of biological data.

Future Work:

- Use real variant data collected from ClinVar and genomAD. Can PSL could extract the same trends in rule weights from background noise present in real data?
- Reflect the allele frequency indirectly with soft truth values in the HasSimilarAF predicate.

Selected References
1. Richards, Sue, Nazneen Aziz, Sherri Bale, *et al*. 2015. "Standards and Guidelines for the Interpretation of Sequence Variants." *Genetics in Medicine* 17 (5): 405–23.
2. N. M. Ioannidis, J. H. Rothstein, V. Pejaver, *et al*. 2016. "REVEL: an ensemble method for predicting the pathogenicity of rare missense variants." *American Journal of Human Genetics.*, 99(4):877–885.
3. Nykamp, Keith, Michael Anderson, Martin Powers, *et al*. 2017. "Sherloc: A Comprehensive Refinement of the ACMG–AMP Variant Classification Criteria." *Genetics in Medicine* 19 (10): 1105–17.
4. Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2017. "Hinge-loss Markov random fields and probabilistic soft logic." *Journal of Machine Learning Research.*

Acknowledgements

I would like to thank Professor Linderman and the CS department for their guidance throughout this project.