Graphical Models For Rare Sequence Variant Interpretation.

Arun Nampally, Eugene Palovcak, Garrett Bernstein, Matthew Davis
Invitae Corp, 1400 16th Street, San Francisco, CA 94103
{arun.nampally, eugene.palovcak, garrett.bernstein, matthew.davis}@invitae.com

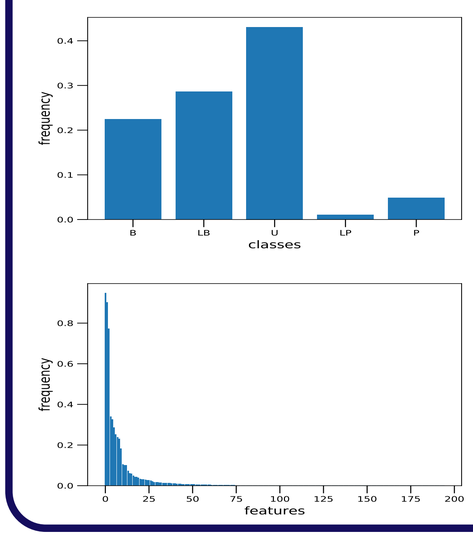
1. Introduction

Clinical genetic testing is a rapidly expanding field, powered by advances in high-throughput genomic sequencing and the increasing availability of well-curated public databases on sequence variants, population genetics, diseases, etc. One main use case is determining whether the mutations (also called variants) detected in the genomic sequence of a proband can explain disease status or predict disease risk. Given that genes influence the phenotype of an individual through highly complex processes, there exists no general model that can conclusively determine the impact of all possible mutations on the health of the individual. Barring some well-studied variants, the interpretation of most rare sequence variants is a process of weighing multiple pieces of evidence in favor/against pathogenicity.

2. Current methods

Most of the labs use variant interpretation systems that are based on ACMG/AMP guidelines. The rare sequence variants are given one of the five class labels – [benign, likely benign, uncertain significance, likely pathogenic, pathogenic] based on evidence spanning categories like population genetics, sequence observations, clinical observations, computational predictors etc. The variants are interpreted using heuristic rules and/or points based systems. While the heuristic rules and the points are based on consensus opinion among geneticists, we believe there is scope for data-driven explicit modeling.

3. Goals



We aim to learn interpret-able classification models for rare sequence variant interpretation from data on previously interpreted variants. Each variant in the dataset is described by 195 feature variables that summarize information such as variant's biochemical consequences, data on the prevalence of the variant in healthy populations, manual (human) interpretations of the variant in clinical and biomedical literature, etc.

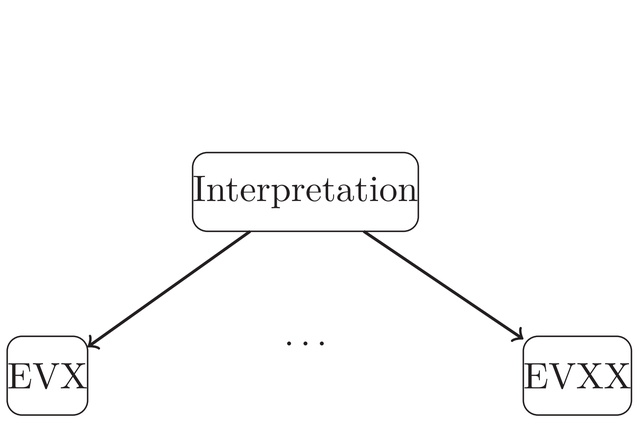
We chose to use the language of directed graphical models because they are highly interpretable and provide an explicit representation of the modeling assumptions. We were also motivated by the possibility of extracting symbolic explanations from these models, being able to perform sensitivity analysis etc.

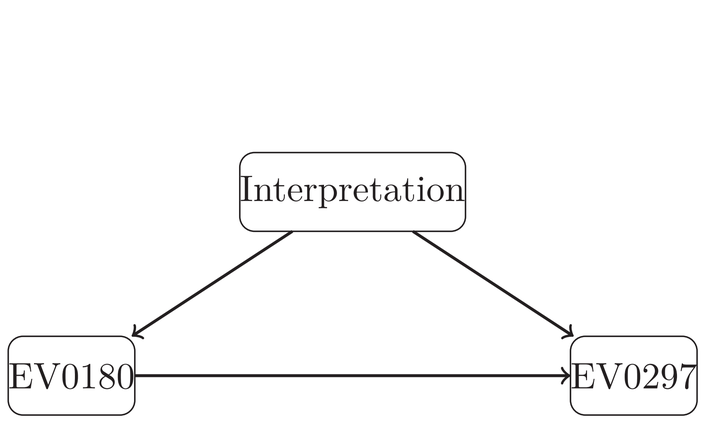
4. Models

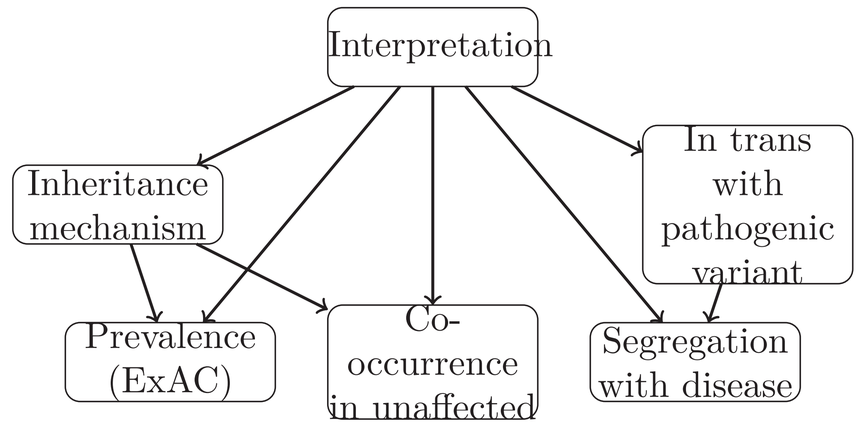
- The first model we considered was the Categorical Naive Bayes model. We trained this model as a baseline PGM against which other models could be compared. We know that the strong independence assumptions made by the model are violated by the features in our dataset. In particular, the features that belong to the same exclusion group exhibit strong correlation and are therefore not independent given the class variable. As an example, there are several features related to minor allele frequency (MAF) thresholds (low, medium, high, etc.) that belong to the same exclusion group. A high MAF obviously rules out features for other thresholds.

- The next model we considered is the Tree Augmented Naive Bayes (TAN) model. The TAN model preserves the appealing properties of the Naive Bayes model (such as computational efficiency and the Markov blanket of the class variable including all features) while relaxing the strong independence assumptions. It allows extra edges between features based on conditional mutual information.

- The last model we considered is also in the spirit of the Naive Bayes model and the TAN model and retains the class variable as the parent of all feature nodes. However, it allows a richer set of dependencies between feature variables than the one extra parent allowed by the TAN model. For this purpose, we used the PC algorithm to learn a DAG among the feature variables. We augmented the DAG returned by the PC algorithm by adding edges based on clinical genetics domain knowledge.







5. Results

- The baseline Naive Bayes model clearly makes strong conditional independence assumptions that are not supported by the dataset and has the lowest performance.

- Relaxing the strong independence assumptions in the TAN model leads to a significant improvement in the performance. The TAN model captured interesting correlations between features such as "coverage in ExAc" and "absent in gnomAD".

- The use of higher level features and encoding of genetics domain knowledge in PCNB model lead to a slight improvement in performance. We believe that complex reasoning involving "exclusion groups" may be main factor causing saturation of performance.

Model	accuracy	f1 score
NB	0.8387	0.8390
TAN	0.9330	0.9330
PCNB	0.9497	0.9502

Class	NB	TAN	PCNB
benign	0.80	0.93	0.94
likely benign	0.79	0.92	0.95
uncertain significance	0.91	0.96	0.97
likely pathogenic	0.44	0.51	0.56
pathogenic	0.75	0.90	0.86

6. Conclusions

- Our experiments demonstrated that classifiers based on graphical models can perform well at the task of variant interpretation.

- By encoding our classifiers as PGMs were able to derive highly interpretable and transparent models.

- Further improvement in performance would require a more detailed encoding of the domain knowledge underpinning variant interpretation.

- Datasets that are annotated in an independent fashion will be a key enabler of automated learning of graphical models.