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

Abstract

Interpretation of rare sequence variants is a key challenge in clinical genetic testing. In the absence of a definitive model to ascertain variant pathogenicity interpretation is usually conducted by combining evidence from multiple sources via heuristic rules and points-based systems. In this paper, we explore a fundamentally different modeling approach – one based on probabilistic graphical models. We present initial attempts at graphical modeling of the variant interpretation task, highlighting the benefits such as transparency of modeling assumptions, explainability, sensitivity analysis, etc. while also describing challenges that are to be overcome.

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.

The current guidelines for variant interpretation suggest a five-fold classification system (*Benign, Likely Benign, Uncertain Significance, Likely Pathogenic, Pathogenic*). Such a system was originally proposed by Plon et al. (2008) for cancer genes, who also gave probability thresholds that should be met for each class. For example, a variant classified as *Pathogenic* should have more than a .99 probability of causing disease. The American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP) guidelines (Richards et al. 2015) build on this foundation to propose a classification for rare sequence variants in general. These guidelines define several criteria (i.e., predictors/features) which indicate pathogenicity or the lack of it. Given all criteria that apply to a variant, classification is arrived at by using heuristic rules.

Sherloc (Nykamp et al. 2017) is a classification system derived from the ACMG/AMP framework that uses pathogenic/benign points associated with features. The pathogenic (positive) and benign (negative) points of the features are combined using rules that are called *exclusion groups*. These rules are designed to prevent double-counting of evidence and to prioritize certain forms of evidence over others. The variant classification is given based on where the cumulative points fall among predefined thresholds for the five classes.

Related Work

While there is a large body of literature on algorithmic tools to predict the impact of sequence variants, most focus on predicting the functional or biochemical consequences of the variant on its gene product. *PolyPhen-2* (Adzhubei et al. 2010), *SIFT*(Kumar, Henikoff, and Ng 2009), *CADD* (Rentzsch et al. 2019), etc. are well-known computational predictors of this type. In contrast, clinical variant interpretation is concerned with whether a sequence variant causes disease in humans. Clinical variant interpretation necessarily includes evidence on the functional impact of the variant but considers it in the context of clinical evidence about human carriers.

The ACMG/AMP guidelines and Sherloc are heuristic classification schemes for integrating functional and clinical evidence about variants. The problem of formalizing these systems has received less attention. Tavtigian et al. (2018) explored whether the ACMG/AMP framework can be re-interpreted as a Bayesian classifier. They showed that a Naive Bayes model using ACMG/AMP features and specific parameter values has the following interesting property: when the ACMG/AMP classifier outputs a particular class, the Naive Bayes model outputs a probability of pathogenicity that meets the thresholds defined by Plon et al. (2008) for that class. However, this construction of the Naive Bayes model is not data-driven and the parameters given by Tavtigian et al. (2018) were not explicitly validated against datasets of interpreted variants.

Motivation

The post-facto re-interpretation of the ACMG/AMP framework by Tavtigian et al. (2018) is an interesting result. But a more principled approach to designing a Naive Bayes model would be to learn it directly from the data. Moreover, the strong independence assumptions made by the Naive Bayes model may not be supported by the data, and we may have to consider more expressive models that capture complex dependencies between the features. This leads us to the class of directed *Probabilistic Graphical Models (PGMs)* of which the Naive Bayes model is one instance. PGMs, also known as *Bayesian Networks* and *Belief Networks* (Darwiche 2009; Koller and Friedman 2009; Pearl 2009), provide an expressive language for representing probabilistic models. The key idea in encoding a PGM is to use a directed acyclic graph (DAG) to capture the conditional independencies between random variables, thereby obtaining an efficient, factorized representation of the joint multivariate distribution. Based on the semantics of the edges, they can also be viewed as encoding causal models. PGMs offer several advantages over black-box machine learning techniques, such as providing explicit representation of the modeling assumptions, enabling the user to extract symbolic explanations, supporting sensitivity analysis of the parameters, etc. These observations motivate us to leverage the power of PGMs for the task of variant interpretation.

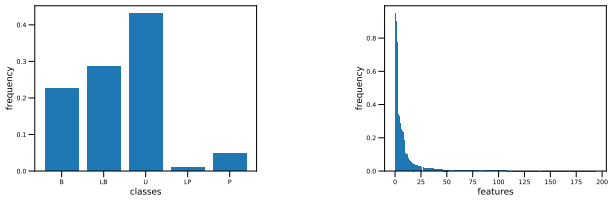
Problem Formulation

The learning task we are pursuing involves inducing PGMs for variant interpretation from data on previously interpreted variants. We use an internal dataset containing $\approx 200K$ variants that were interpreted using a framework similar to Sherloc. 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. These features are associated with pathogenic/benign points (see Table 1), and the classification is computed by using exclusion groups similar to those in Sherloc. While the dataset is not publicly available, we note that these interpreted variants are routinely submitted to *ClinVar* (Landrum et al. 2018).

Since we want to learn the PGMs based on applicability/inapplicability information of the features instead of subjective points, we binarize the dataset before using it. We observe that the dataset is highly skewed in terms of the class distribution. While some features are present in almost every variant, most are rarely used (see Fig. 1). We also note that the dataset does not distinguish between instances where a feature was considered and found to be inapplicable and instances where the domain expert did not consider the feature during interpretation. Using this dataset, we seek to learn Bayesian networks with Bernoulli distributed feature variables and a class variable with a Categorical distribution.

Models

In this section, we describe the various PGMs we explored for the variant interpretation task. While a causal graphical model is highly valuable in this context, the complex nature of the biological processes involved makes it very challenging to define such a model. Therefore, we focused on exploring models that maximize classification performance even if they do not represent a fully causal account of the data generating process.



(a) class distribution (b) frequency of features

Figure 1: Summary of the dataset. (a) Class distribution is imbalanced with pathogenic and likely pathogenic variants forming a small fraction. (b) Small set of features are consistently applied, while rest are rarely applied.

Feature	Points	Description
EV0023	2.5	Protein function disrupted, strong evidence
EV0208	2	Initiator codon variant, loss of function not established
EV0297	1	Absent in gnomAD
EV0049	4	Strong segregation with disease
EV0214	-5	Recessive: MAF very high in gnomAD

Table 1: Sample features used in the dataset. The features can be either pathogenic or benign and have points that indicate the strength of the evidence they represent.

Naive Bayes Model

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.

Tree Augmented Naive Bayes Model

The next model we considered is the *Tree Augmented Naive Bayes (TAN)* model (Friedman, Geiger, and Goldszmidt 1997). 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. In a TAN model, the feature variables can have one more parent node apart from the class variable. These ex-

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

Table 2: Evaluation metrics. TAN and PCNB model significantly outperform NB model by modeling the dependencies between feature variables.

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

Table 3: Per-class f1 scores. The models perform well on commonly observed classes and have weak performance for *likely pathogenic* class, which has least amount of data.

tra edges are added from an "augmenting tree" over feature variables. The augmenting tree is computed using *conditional mutual information* between dataset features (see Friedman, Geiger, and Goldszmidt (1997) for details).

PC Naive Bayes Model

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 (Spirtes et al. 2000) to learn a DAG among the feature variables. The PC algorithm can learn an equivalence class of DAGs that satisfies the conditional independencies in a dataset. The algorithm starts with a fully connected undirected graph and progressively removes edges between nodes based on conditional independence tests.

In practice, we found the DAG returned by the PC algorithm to have a sparse set of dependencies between feature variables. To have a more extensive set of dependencies between features, we used the conservative skeleton output by the algorithm, then added back edges between feature variables. We used clinical genetics domain knowledge to choose these edges. For this model, we also combined related binary feature variables into categorical variables that represent higher-level genetic concepts. This was done primarily to facilitate manual construction and downstream interpretation of the model by geneticists, but also to enforce mutual exclusion between certain sets of evidence features. For example, the Sherloc system has binary evidence features for "1 case report", "2 case reports", "3 case reports" which we combined into one "case reports" variable modeled by a categorical distribution.

Implementation

We leveraged the *Pyro* probabilistic programming language (Bingham et al. 2019) to learn the *conditional probability distributions* (CPDs) of the nodes in our models (structure learning for TAN and PCNB models was implemented separately). The use of probabilistic programming allows us to move beyond discrete graphical models without implementing custom parameter learning solutions. For example, we can easily extend our models to have both discrete and continuous nodes or have nodes that are deterministically tied to their parents.

We implemented Bayesian parameter estimation of CPDs through the use of uninformative Beta priors over feature variables and uninformative Dirichlet priors over the class variable. Although Pyro allows us to search for the posteriors in an approximating family of distributions, we used the exact family of distributions to compute the posteriors over CPDs.

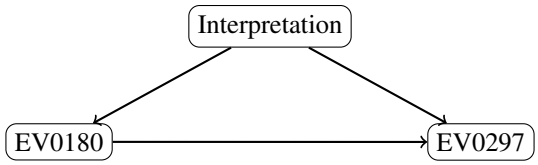


Figure 2: Example correlation in TAN model. By conditioning *EV0297* on *EV0180*, the model captures correlation between population databases.

Results and Discussion

The models described earlier were compared using classification accuracy and weighted f1-score on a hold-out test set. We used an 80/20 train/test split of the dataset for this purpose. The evaluation metrics are listed in Table 2. The breakdown of the f1-scores by class is shown in Table 3.

The baseline Naive Bayes model clearly makes strong conditional independence assumptions that are not supported by the dataset and has the lowest performance. Since we do not have a one-to-one correspondence between the features of our dataset and the ACMG/AMP criteria, we were unable to verify whether the learned parameters of our model fall in the space of feasible parameters identified by Tavtigian et al. (2018).

Relaxing the strong independence assumptions in the TAN model leads to a significant improvement in the performance. For example, one edge acquired by the TAN model over the Naive Bayes model is shown in Fig 2. The code EV0180 stands for "Insufficient coverage in ExAC" whereas EV0297 stands for "Absent in gnomAD" (*ExAC* and *gnomAD* are population databases). By conditioning the distribution of EV0297 on EV0180, the model takes into account the fact that a variant having low coverage in one database is likely to have low coverage or be missing from another database. The ability of the TAN model to capture correlations such as these contributes to its significantly higher performance. We observed that the augmented edges do not have a strong correspondence to the exclusion groups and the choice of the root of the augmenting tree did not have a significant impact on the model performance.

While the TAN model performed strongly, its structure

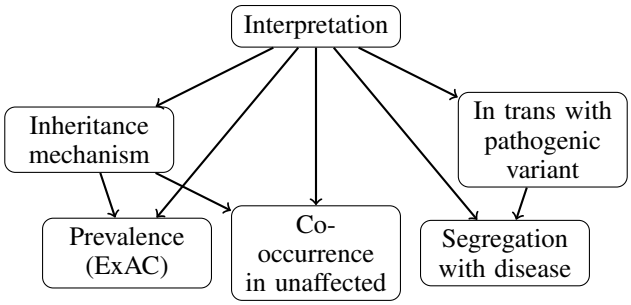


Figure 3: Part of PCNB model. High-level genetic concepts are used as nodes and edges reflect domain knowledge about their relationships.

was derived automatically from the data without leveraging domain knowledge from geneticists. In contrast, early models we constructed based entirely on domain knowledge performed very poorly, even worse than the Naive Bayes model. This is likely because Naive Bayes and TAN classifiers make the "interpretation" variable the parent of every feature variable, ensuring that every feature contributes directly to the classification (i.e. they are in its Markov blanket). We reasoned that so long as this aspect of the model structure was kept intact, there might be some flexibility in specifying the set of augmenting edges between feature variables. This led us to use the PC algorithm to assist in semi-automated structure learning. While this model did not significantly outperform the TAN model, it achieved similar performance while using feature variables that represent familiar high-level genetic concepts like "inheritance", "prevalence", and "segregation with disease". For example, in Fig 3, the "Inheritance mechanism" variable takes values such as "autosomal dominant" or "autosomal recessive". Knowing the inheritance mechanism of a variant should change our expectations about its prevalence in a population database like ExAC: pathogenic variants operating through a recessive mechanism may be more common in the ExAC population than those operating through a dominant mechanism. Including an edge between "inheritance mechanism" and "prevalence (ExAC)" enables the model to learn the expected distribution of population prevalence for a variant conditioned both on its interpretation and its inheritance mechanism. This moves us closer to a model that more fully represents the causal domain knowledge.

The classifications of the variants in the dataset were computed using a extensive set of exclusion groups. For example, EV0023 excludes EV0208 in one group and EV0208 excludes EV0297 in a lower priority group. When EV0208 and EV0297 are applicable to a variant, only the points of EV0208 are used for classification, but if EV0203 is also applicable, then the points of EV0203 and EV0297 are used. We believe that this kind of complex reasoning used in the heuristics may be a contributing factor to the saturation of PCNB model performance. While it is straightforward to model a single exclusion group through the use of *Context Specific Independence* (CSI) (Boutilier et al. 1996) (more specifically a tree-structured CPD), modeling multi-level exclusion logic is much more challenging and may be a key hurdle to achieving higher performance classifiers.

We end this section by highlighting some unique strengths of the graphical modeling approach as compared to other black-box learning techniques.

- **decision support:** We were able to compile the learned models to *arithmetic circuits* (Darwiche 2002, 2003) using the ACE[1] compiler. These circuits provide, among other things, the ability to compute the distributions resulting from all single variable changes to an instantiation. This can be a useful feature to the user by highlighting the unassigned features which cause greatest change in the classification probability.
- **testing semantics of evidence criteria:** The features used to annotate the dataset are grouped into pathogenic and benign categories based on whether they are indicative of pathogenicity or not. The use of PGMs allowed us to interrogate their semantics, by considering the change in class distribution while leaving all other features unassigned. For the TAN model, we were able to identify several *benign* features for which applying the feature causes a decrease in the probability of the variant being *benign* or *likely benign*. An example is the feature "EV0211: Variant present in a clinically useful locus-specific database". Perhaps, this is due to the fact that variants collected in locus specific databases are generally pathogenic. While this surprising behavior should be considered under the assumption that TAN model is a correct encoding of the variant interpretation task, it was nevertheless an interesting observation. Note that black-box techniques are typically unable to handle such queries involving partial instantiation of feature variables.

Conclusion

In this paper we sought to formalize the heuristic approaches to variant interpretation by using the theory of graphical models. Our experiments demonstrated that classifiers based on graphical models (as opposed to heuristics rules and subjective scores) can perform well at the task of variant interpretation. By encoding our classifiers as PGMs were able to derive highly interpretable and transparent models, whose assumptions (in terms of conditional independencies) could be read off the corresponding DAG. Further improvement in performance would require a more detailed encoding of the domain knowledge underpinning variant interpretation. Since the dataset of interpreted variants we used for training was generated by heuristic variant classifiers, our models are likely biased towards their heuristic rules. We believe that high quality variant datasets annotated in an independent fashion will be a key enabler in automated learning of graphical models for this task.

Acknowledgements

We would like to thank Jeanne Leisk for clarifying the details of the Sherloc system and Luc Cary for helping with the user interface development.

[1] http://reasoning.cs.ucla.edu/ace/

References

Adzhubei, I. A.; Schmidt, S.; Peshkin, L.; Ramensky, V. E.; Gerasimova, A.; Bork, P.; Kondrashov, A. S.; and Sunyaev, S. R. 2010. A method and server for predicting damaging missense mutations. *Nature methods* 7(4): 248–249.

Bingham, E.; Chen, J. P.; Jankowiak, M.; Obermeyer, F.; Pradhan, N.; Karaletsos, T.; Singh, R.; Szerlip, P.; Horsfall, P.; and Goodman, N. D. 2019. Pyro: Deep universal probabilistic programming. *The Journal of Machine Learning Research* 20(1): 973–978.

Boutilier, C.; Friedman, N.; Goldszmidt, M.; and Koller, D. 1996. Context-specific Independence in Bayesian Networks. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, UAI'96, 115–123. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN 1-55860-412-X.

Darwiche, A. 2002. A logical approach to factoring belief networks. In *Proc. 8th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR-02)*.

Darwiche, A. 2003. A Differential Approach to Inference in Bayesian Networks. *J. ACM* 50(3): 280–305. ISSN 0004-5411.

Darwiche, A. 2009. *Modeling and reasoning with Bayesian networks*. Cambridge: Cambridge University Press.

Friedman, N.; Geiger, D.; and Goldszmidt, M. 1997. Bayesian Network Classifiers. *Machine Learning* 29(2-3): 131–163.

Koller, D.; and Friedman, N. 2009. *Probabilistic graphical models: principles and techniques*. The MIT Press.

Kumar, P.; Henikoff, S.; and Ng, P. C. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* 4(7): 1073.

Landrum, M. J.; Lee, J. M.; Benson, M.; Brown, G. R.; Chao, C.; Chitipiralla, S.; Gu, B.; Hart, J.; Hoffman, D.; Jang, W.; et al. 2018. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research* 46(D1): D1062–D1067.

Nykamp, K.; Anderson, M.; Powers, M.; Garcia, J.; Herrera, B.; Ho, Y.-Y.; Kobayashi, Y.; Patil, N.; Thusberg, J.; Westbrook, M.; et al. 2017. Sherloc: a comprehensive refinement of the ACMG–AMP variant classification criteria. *Genetics in Medicine* 19(10): 1105.

Pearl, J. 2009. *Causality*. Cambridge: Cambridge University Press.

Plon, S. E.; Eccles, D. M.; Easton, D.; Foulkes, W. D.; Genuardi, M.; Greenblatt, M. S.; Hogervorst, F. B.; Hoogerbrugge, N.; Spurdle, A. B.; Tavtigian, S. V.; et al. 2008. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Human mutation* 29(11): 1282–1291.

Rentzsch, P.; Witten, D.; Cooper, G. M.; Shendure, J.; and Kircher, M. 2019. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research* 47(D1): D886–D894.

Richards, S.; Aziz, N.; Bale, S.; Bick, D.; Das, S.; Gastier-Foster, J.; Grody, W. W.; Hegde, M.; Lyon, E.; Spector, E.; et al. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in medicine* 17(5): 405–423.

Spirtes, P.; Glymour, C. N.; Scheines, R.; and Heckerman, D. 2000. *Causation, prediction, and search*. MIT press.

Tavtigian, S. V.; Greenblatt, M. S.; Harrison, S. M.; Nussbaum, R. L.; Prabhu, S. A.; Boucher, K. M.; and Biesecker, L. G. 2018. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genetics in Medicine* 20(9): 1054–1060.