Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240747

Comparing Sequence-Based and Literature-Based Pathogenicity Scoring Methods for Human Variants

Luc MOTTIN[a,b,1], Nona NADERI[c], Anaïs MOTTAZ[a,b], Pierre-André MICHEL[a,b], Gerieke BEEN[d], Lennart JOHANSSON[d], Morris SWERTZ[d], Andrew STUBBS[e], Emilie PASCHE[a,b], Julien GOBEILL[a,b] and Patrick RUCH[a,b]
[a] *HES-SO\HEG Genève, Information Sciences, Geneva, Switzerland*
[b] *SIB Text Mining, Swiss Institute of Bioinformatics, Geneva, Switzerland*
[c] *Department of Computer Science, Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, Orsay, France*
[d] *University of Groningen, University Medical Centre Groningen, Groningen, Department of Genetics, Genomics Coordination Centre, The Netherlands*
[e] *Department of Pathology and Clinical Bioinformatics, Erasmus University Medical Centre, Rotterdam, The Netherlands*
ORCiD ID: Luc Mottin https://orcid.org/0000-0002-9614-9293, Nona Naderi https://orcid.org/0000-0002-1272-7640, Anaïs Mottaz https://orcid.org/0000-0003-0080-9451, Pierre-André Michel https://orcid.org/0000-0002-7023-1045, Gerieke Been https://orcid.org/0009-0002-6667-5951, Lennart Johansson https://orcid.org/0000-0002-4914-3737, Morris Swertz https://orcid.org/0000-0002-0979-3401, Andrew Stubbs https://orcid.org/0000-0001-9817-9982 , Emilie Pasche https://orcid.org/0000-0002-9118-5762, Julien Gobeill https://orcid.org/0000-0001-9809-7741, Patrick Ruch https://orcid.org/0000-0002-3374-2962

Abstract. Assessing the pathogenicity of genetic variants is a critical aspect of genomic medicine and precision healthcare. Over the last decades, the identification of genetic variants and their characterization has become simpler (advent of high-throughput sequencing technologies, analysis, and visualization support tools, etc.). However, the quality of assessments to distinguish benign from pathogenic variants is critical to inform clinical decision-making and improve patient outcomes. In this article, we investigate the relationships using correlation tests between the characterization of genetic variants in the literature and their pathogenicity scores computed by two state-of-the-art assessment tools (SIFT and PolyPhen-2).

Keywords. Text-mining, Variant pathogenicity, SIFT, PolyPhen-2

1. Introduction

The assessment of variant pathogenicity is crucial in human genetics as it aims to differentiate between functionally neutral mutations and those contributing to disease

[1] Corresponding Author: Luc Mottin, HEG Genève, Rue de la Tambourine 17, 1227 Carouge, Switzerland; E-mail: luc.mottin@hesge.ch.

pathology [1-3]. Variations at the DNA level are primarily represented by single nucleotide polymorphisms (SNPs), which are therefore an important driver of research in the fields of oncology and rare diseases. Genome-Wide Association Studies (GWAS) have identified numerous disease-associated SNPs that provide insights into the underlying molecular pathological mechanisms and serve as biomarkers for disease susceptibility, prognosis and prediction of treatment outcomes and enable tailored therapeutic strategies based on individual genetic profiles. This is especially relevant in oncology, but other medical areas can benefit from information about sequence variations such as orphan diseases [4].

Yet, the sheer volume of SNPs represents a challenge in identifying variants responsible for specific traits. The prioritization of SNPs based on their functional significance emerges as a strategic approach, leveraging biological knowledge to distinguish between neutral variants and those of likely functional importance. To face this challenge, computational tools like SIFT (Sorting Intolerant From Tolerant) and PolyPhen-2 (Polymorphism Phenotyping v2) have emerged [5,6]. By identifying potentially damaging variants in protein-coding regions, these tools provide insights into disease susceptibility and may inform clinical decision-making with the objective to optimize personalized therapeutic efficacy while minimizing adverse effects.

More comprehensive characterization of variants in human genetics relies heavily on evidence gathered from scientific literature, leveraging the vast repository of knowledge accumulated over decades of research (over 36 million citations in MEDLINE, plus clinical trials, full-text articles and their supplementary data). With the aim to automatically extract actionable insights beyond pathogenicity predictions, we could sift through vast quantities of textual data using Natural Language Processing approaches. One of the first steps to determine how automated literature approaches can help predict the pathogenicity of variants is to see whether there is a correlation between the predicted pathogenicity and how it is across the scientific literature. In this study, we therefore sought to test whether there is a correlation between the literature prevalence of variants, and their pathogenicity scores. If so, this would open up the way to designing better curation-support tools, likely to combine together sequence-based pathogenicity scores and evidence automatically derived from the literature.

2. Methodology

For this study, we used data from four sources: a list of reference variants from VariBench [7], their pathogenicity scores computed with the SIFT and PolyPhen-2 tools, and a comprehensive set of scientific literature and clinical trials documents provided by SIB Literature Services (SIBiLS) [8] and searched through Variomes [9].

2.1. List of Variants

VariBench provides benchmark datasets for variation interpretation, aiding in the development and evaluation of computational prediction methods. These datasets are curated and cover diverse variation types and effects, supporting method training, testing, and post-publication comparisons. In this study, we use the training subset of VariBench data provided by Gene-Aware Variant INterpretation (GAVIN) [10], which represents a dataset of 17,490 variants, all SNPs, with around half of them being predicted to be deleterious, as illustrated on Figure 1.

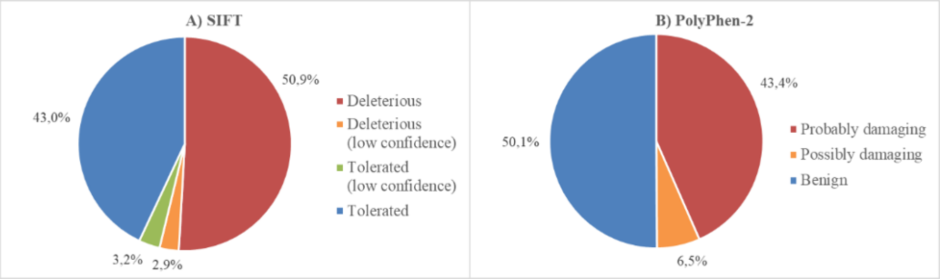

Figure 1. Distribution of pathogenicity assessments in the VariBench dataset (training set split) according to A) SIFT, and B) PolyPhen-2.

2.2. Variants' Pathogenicity Score

SIFT and PolyPhen-2 are two commonly used algorithms for predicting the effect of a SNP on protein function. SIFT is a sequence homology-based tool designed to predict the phenotypic effect of amino acid substitutions in proteins. Obtaining related protein sequences, the program leverages the inherent conservation patterns observed across the protein family to construct position-specific scoring matrices. Then, it applies a predetermined cutoff to predict the tolerance or intolerance of substitutions. The SIFT score ranges from 0 to 1, with a score close to zero indicating that the substitution is predicted to be deleterious or damaging to the protein function while a score closer to 1 suggests that the substitution is likely to be tolerated or benign. PolyPhen-2 is designed to predict the functional impact of amino acid substitutions in proteins. Leveraging eight sequence-based and three structure-based predictive features, PolyPhen-2 automatically selects informative features characterizing wild-type and mutant alleles. It has been trained and tested on two datasets (HumDiv and HumVar) and utilizes a Naïve Bayes classifier to predict the functional significance of allele replacements based on individual features. PolyPhen-2 score ranges from 0 to 1, where a higher score indicates a higher likelihood that the substitution is damaging to the protein function, in opposition to the SIFT score where 1 is indicative of benignity.

2.3. Compendium of Literature

To estimate the prevalence of each variant in the literature, we used Variomes (https://variomes.text-analytics.ch/), a search engine supporting the curation of genomic variants with the biomedical literature. Variomes uses four articles' collections provided by SIBiLS (https://sibils.text-analytics.ch/): MEDLINE, PMC Open Access subset, Clinical Trials (CT) and a set of Supplementary Data. These collections are daily updated and automatically annotated to identify biomedical entities such as drugs, diseases, genes, and species. The set of supplementary data consisted of 4.2M files from ~800k PMC references retrieved through the query « [(gene AND variant) OR polymorphism* OR mutat*] ». The files were mainly text/tables and images (*i.e.* 28.5% of .txt files and 28.5% of .jpg images, as detailed in [11]). To efficiently search for variants, Variomes relies on SynVar [12], a dedicated system to expand the query and retrieve documents mentioning the variant in many formats, including protein, transcript, and genome levels, as well as syntactic variations as found in the literature. The literature prevalence of each variant was therefore estimated as the number of documents in each collection containing at least one mention of the given variant or one of its synonyms.

2.4. Correlation

The correlations between the frequency measures and SIFT and Polyphen-2 scores were computed using Spearman's correlation from the SciPy package, given the non-normal distribution of these scores. Of the 17,490 variants from VariBench, the SIFT scores were retrievable for only 10,123, which constitutes the final dataset used in this study. To prevent any confusion when comparing correlations with SIFT and PolyPhen-2 scores, which have inverse pathogenicity scales, we present our results in the form of absolute correlation coefficient values.

3. Results

Table 1 displays the absolute rank's correlation coefficients for 10,123 variants. We observe that PolyPhen-2 and SIFT scores are inversely correlated (0.744) which is expected given the definition of their score as mentioned in the Methodology section. Furthermore, when comparing the number of hits from the literature with either SIFT or PolyPhen-2, we see that the highest correlations were obtained when using PMC, with a coefficient of 0.112 and 0.152 respectively, and when using MEDLINE, with of a coefficient of 0.112 and 0.141 respectively. Although relatively low, this suggests that variants found in full-text articles and abstracts are more likely to be pathogenic than variants found in CT (with correlation coefficients of 0.055 and 0.027) or in supplementary data files (with non-significant correlation coefficients of 0.007 and 0.024) [12]. This is expected because supplementary materials share more similarity with raw clinical data (*e.g.*, as stored in cohorts).

Table 1. Spearman's correlation coefficients between two pathogenicity scores (SIFT and Polyphen-2) and variant prevalence in different literature repositories. In bold the correlations with *P*<.01.

	MEDLINE	PMC	Clinical Trials	Supp. data	PolyPhen-2
SIFT	**0.112**	**0.112**	**0.055**	0.007	**0.744**
PolyPhen-2	**0.141**	**0.152**	**0.027**	0.024	**1**

4. Discussion

We have shown that the prevalence of variants in the literature is an indication of the pathogenicity of variants, even if relatively weak. A next step would be to verify whether it only reproduces the signal of the pathogenicity score or whether it has an additional value and improves the prediction beyond the existing score, potentially improving variant prioritization strategies. These results may be further validated by comparison with more recent pathogenicity assessment methods such as CAPICE [13].

Moreover, it would be beneficial to investigate the contrasting provenance of additional data files, particularly between image modality and tabular data. Such an analysis could leverage more granular data to highlight medium or strong correlations depending on sample size. Additionally, exploring the position of data within the full-text, whether structured according to the argumentation section (Introduction, Methods, Results, Discussion, Conclusion) or clinical templates such as PICO (Population, Intervention, Comparison, Outcome), affords an alternative and more specific means of observing correlations. We already know that the identification of these positions can be automatized [14], which can streamline the process.

5. Conclusion

With respect to the literature datasets, we found that the abstracts (MEDLINE) and the full-texts (PMC) were best suited for estimating pathogenicity. The results suggest that supplementary data may contain all experimental data, and therefore any observed human polymorphisms for a given study, while only pathogenic variants or actionable ones are retained for inclusion and discussion in the full-text part of the publication.

References

[1] Duzkale H, Shen J, McLaughlin H et al. A systematic approach to assessing the clinical significance of genetic variants. Clinical Genetics, 2013, 84(5), 453-63. doi: 10.1111/cge.12257.

[2] Gunning AC, Fryer V, Fasham J et al. Assessing performance of pathogenicity predictors using clinically relevant variant datasets. Journal of Medical Genetics, 2021, 58(8), 547-555. doi: 10.1136/jmedgenet-2020-107003.

[3] Ciesielski TH, Sirugo G, Iyengar SK et al. Characterizing the pathogenicity of genetic variants: the consequences of context. NPJ Genomic Medicine, 2024, 9(3). doi: 10.1038/s41525-023-00386-5.

[4] Mascia C, Francesca F, Paolo U et al. The openEHR Genomics Project. Proceedings of the 30th Medical Informatics Europe conference (MIE 2020), Studies in health technology and informatics, 2020, 270. doi: 10.3233/SHTI200199.

[5] Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Research, 2001, 11(5), 863-74. doi: 10.1101/gr.176601.

[6] Adzhubei I, Schmidt S, Peshkin L et al. A method and server for predicting damaging missense mutations. Nature Methods, 2010, 7, 248–249. doi: 10.1038/nmeth0410-248.

[7] Sasidharan Nair P, Vihinen M. VariBench: a benchmark database for variations. Human Mutation, 2013, 34(1), 42-9. doi: 10.1002/humu.22204.

[8] Gobeill J, Caucheteur D, Michel PA et al. SIB Literature Services: RESTful customizable search engines in biomedical literature, enriched with automatically mapped biomedical concepts, Nucleic Acids Research, 2020, 48(1), 12–16. doi: 10.1093/nar/gkaa328.

[9] Pasche E, Mottaz A, Caucheteur D et al. Variomes: a high recall search engine to support the curation of genomic variants. Bioinformatics, 2022, 38(9), 2595-2601. doi: 10.1093/bioinformatics/btac146.

[10] van der Velde KJ, de Boer EN, van Diemen CC et al. GAVIN: Gene-Aware Variant INterpretation for medical sequencing. Genome Biology, 2017, 18(6). doi: 10.1186/s13059-016-1141-7.

[11] Pasche E, Mottaz A, Gobeill J et al. Assessing the use of supplementary materials to improve genomic variant discovery. Database (Oxford), 2023, baad017. doi: 10.1093/database/baad017.

[12] Mottaz A, Pasche E, Michel PA et al. Designing an Optimal Expansion Method to Improve the Recall of a Genomic Variant Curation-Support Service. Studies in Health Technology and Informatics, 2022, 294, 839-843. doi: 10.3233/SHTI220603.

[13] Li S, van der Velde KJ, de Ridder D et al. CAPICE: a computational method for Consequence-Agnostic Pathogenicity Interpretation of Clinical Exome variations. Genome Medicine, 2020, 12(75). doi: 10.1186/s13073-020-00775-w.

[14] Ruch P, Baud RH, Chichester C et al. Extracting Key Sentences with Latent Argumentative Structuring. Connecting Medical Informatics and Bio-Informatics - Proceedings of MIE 2005, Studies in Health Technology and Informatics, 2005, 116, 835-840. Access: http://ebooks.iospress.nl/volumearticle/10410.