Bioinformatics, 38(18), 2022, 4449–4451
https://doi.org/10.1093/bioinformatics/btac537
Advance Access Publication Date: 29 July 2022
Applications Note

OXFORD

Data and text mining

tmVar 3.0: an improved variant concept recognition and normalization tool

Chih-Hsuan Wei [1], Alexis Allot [1], Kevin Riehle[2], Aleksandar Milosavljevic[2] and Zhiyong Lu [1,*]

[1]National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, MD 20894, USA and [2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on April 7, 2022; revised on July 7, 2022; editorial decision on July 8, 2022; accepted on July 27, 2022

Abstract

Motivation: Previous studies have shown that automated text-mining tools are becoming increasingly important for successfully unlocking variant information in scientific literature at large scale. Despite multiple attempts in the past, existing tools are still of limited recognition scope and precision.

Result: We propose tmVar 3.0: an improved variant recognition and normalization system. Compared to its predecessors, tmVar 3.0 recognizes a wider spectrum of variant-related entities (e.g. allele and copy number variants), and groups together different variant mentions belonging to the same genomic sequence position in an article for improved accuracy. Moreover, tmVar 3.0 provides advanced variant normalization options such as allele-specific identifiers from the ClinGen Allele Registry. tmVar 3.0 exhibits state-of-the-art performance with over 90% in F-measure for variant recognition and normalization, when evaluated on three independent benchmarking datasets. tmVar 3.0 as well as annotations for the entire PubMed and PMC datasets are freely available for download.

Availability and implementation: https://github.com/ncbi/tmVar3

Contact: zhiyong.lu@nih.gov

Introduction

Genomic variants are an essential part of precision medicine which aims to provide personalized treatments based on an individual's genetic profile. To better understand the mechanism of genetic diseases, (semi-)automatically collecting and assimilating published knowledge about sequence variants in scientific literature becomes an increasingly important task. A recent study (Lee *et al.*, 2021) reviewed a number of existing software tools previously developed for such a task (Caporaso *et al.*, 2007; Cejuela *et al.*, 2017; Cheng *et al.*, 2020; Thomas *et al.*, 2016; Wei *et al.*, 2017). Most of these tools use regular expressions based on the Human Genome Variation Society (HGVS) nomenclature and frequent variant forms in text. We previously developed tmVar (Wei *et al.*, 2013), which uses a machine learning-based approach to optimally recognize variant components (wild type, mutant and position). More recently, a new function was added to tmVar so it performs variant normalization by linking recognized variant mentions to standard concept identifiers (Wei *et al.*, 2017). Specifically, tmVar 2.0 normalizes variant mentions to dbSNP (https://www.ncbi.nlm.nih.gov/snp/) RS identifiers (RS IDs). Relying on the variant linking results of tmVar, several downstream text mining applications were successfully developed (Allot *et al.*, 2018; Nie *et al.*, 2019).

Despite these efforts, existing variant extraction tools are still limited in (i) recognizing variant types of a broader scope such as incomplete variants (e.g. V600) and related concepts (e.g. genomic regions). Such concepts are found to play important roles in connecting variants with disease and drug information in the same article, (ii) linking mentions to specific alleles. Note that dbSNP RS IDs (e.g. rs113488022) only record the polymorphism at a specific position but do not differentiate specific (e.g. T>A versus T>C) registered in the ClinGen Allele Registry (CAR) (Pawliczek *et al.*, 2018). We herein propose a new comprehensive variant extraction system that specifically addresses all these challenges. The improved tmVar 3.0 system achieves consistent high precision and recall on several publicly available gold standard corpora and is freely available for the scientific community.

Methods

We first expanded the recognition scope of tmVar to cover more difficult cases that were rarely addressed by existing tools, such as incomplete variant mentions (e.g. Cys326; guanine to cytosine), copy number variants (e.g. chr19:54 666 173–54 677 766 bp del), reference sequence (RefSeq) (e.g. NM_203475.1), chromosomal locations (e.g. chromosome 5 q 33) and genomic regions (e.g. chr7:156 583 796–156 584 569) as shown in Table 1.

To better support variant-related text mining research (e.g. mining variant-disease associations), tmVar 3.0 groups variants from the same

Table 1. The mutation types extracted by tmVar 3.0 and examples

Type	Example	tmVar 3.0	tmVar2.0	SETH
SNP	Rs763780	✓	✓	✓
DNA mutation	c.1976A>T	✓	✓	✓
DNA allele	1976A	✓		
DNA change	A>T	✓	✓	✓
Protein mutation	p.Gln659Leu	✓	✓	✓
Protein allele	glutamine at codon 659	✓		
Protein change	methionine to threonine	✓	✓	✓
Other mutations	306 base pair insertion	✓		
Copy number variant	Chr15: 31 833 000–37 477 000 bp deletion	✓		
RefSeq	NM_203475.1	✓		
Chromosome	10q11.12	✓		
Genomic region	Chr10: 46 123 781–51 028 772	✓		✓

genomic sequence position even in different form/types (e.g. DNA and protein variants/alleles). For instance, in PMID: 20 577 006, we group the variants (i.e. P799L and P799) belonging to rs121912637. In this article, P799 co-occurs with disease *metatropic dysplasia* in the same sentence, but not P799L. In this case, grouping the two variant mentions makes it easier to link P799L to the correct disease.

Third, tmVar 3.0 provides alternative options to normalize a particular variant. In addition to providing RS IDs that record all the possible allele changes on a specific genomic position, we offer three allele-specific options for improved precision in the normalization results: (i) CAR Canonical Allele Identifier (CA ID) (e.g. CA16602736) and (ii) the combination of an RS ID and the specific allele [e.g. rs113488022(T>A)]. CA ID is a granular identifier and can specify the specific allele of the genomic sequence position. To map the variants to CA IDs, we expanded the mapping table of the variant normalization. In addition to the existing records (i.e. variant, corresponding gene and RS ID), we appended the CA IDs to the table. With the variant in the raw text and the corresponding gene recognized by our gene tagger [e.g. GNormPlus (Wei *et al.*, 2015)], the RS ID and CA ID can be searched directly using the mapping table. Furthermore, we observed that more than half of the variant mentions cannot be linked to an existing record in dbSNP or CAR databases. In such cases, tmVar 3.0 finds the corresponding gene of the variant in the text and normalizes it with the variant as the third option (e.g. BRAF: c.1799T>A). The percentages of the normalized variants in the entire PubMed/PMC are 25.43% to CA IDs, 50.23% to RS IDs and 33.80% to corresponding genes. Not all the variants can be normalized or mapped to a corresponding gene, since gene information is lacking in some articles.

Finally, in tmVar 3.0, we improved our recognition algorithm on some previously difficult edge cases such as variants described in natural language (e.g. 'nine-nucleotide deletion starting at position 1952'), or with a missing space between the gene and variant (e.g. 'BRAFV600E').

Results

The newly improved tmVar 3.0 system is evaluated on three separate benchmarking datasets [i.e. OSIRIS (Bonis *et al.*, 2006), Thomas (Thomas *et al.*, 2016) and our revised tmVar corpus (Wei *et al.*, 2013)]. In the new tmVar corpus, we annotated all of the relevant variants (e.g. alleles) and mapped every variant to either the RS ID or the corresponding gene. The evaluation results of tmVar 3.0 on variant recognition and normalization are shown in Table 2 and compared with the previous tmVar version (2.0) and SETH (Thomas *et al.*, 2016), a previous state-of-the-art method producing normalized dbSNP RS IDs. As can be seen, tmVar 3.0 achieves consistently higher accuracy (over 90% in F-measure) than SETH and tmVar 2.0 on the three public corpora. To facilitate the use of tmVar results at PubMed scale, we have processed the entire PubMed/PMC open access and incorporated the results in the NCBI web server PubTator (Wei *et al.*, 2016). The annotations are also freely available via FTP.

Table 2. tmVar 3.0 performance comparison with tmVar 2.0 and SETH on three public benchmarking datasets: tmVar 3.0, OSIRIS (Bonis *et al.*, 2006) and Thomas (Thomas *et al.*, 2016) for variant recognition (NER) and normalization tasks

Corpus	Task	Method	Precision (%)	Recall (%)	F-score (%)
tmVar	NER	tmVar 3.0	94.01	88.86	91.36
		tmVar 2.0	98.22	80.64	88.57
		SETH	97.92	68.77	80.79
	Normalization	tmVar 3.0	96.99	91.71	94.28
		tmVar 2.0	94.49	77.25	85.00
		SETH	86.51	69.91	77.33
OSIRIS	NER	tmVar 3.0	98.62	84.98	91.30
		tmVar 2.0	99.53	83.00	90.52
		SETH	96.43	74.70	84.19
	Normalization	tmVar 3.0	97.72	84.58	90.68
		tmVar 2.0	97.20	80.62	88.14
		SETH	94.21	69.38	79.91
Thomas	NER	tmVar 3.0	92.26	91.30	91.78
		tmVar 2.0	82.46	97.04	89.16
		SETH	84.43	69.39	76.18
	Normalization	tmVar 3.0	91.01	90.32	90.67
		tmVar 2.0	89.94	88.24	89.08
		SETH	95.58	57.50	71.80

Conclusion

We introduce tmVar 3.0, an improved open-source software tool with a broader scope and better accuracy for variant concept recognition and normalization, compared to its predecessors. tmVar 3.0 can recognize most of the variants even when the variants are described with partial information (e.g. amino acid change without the sequence position) or with natural language. tmVar 3.0 groups different mentions of the same variant together based on the context for improved normalization performance. As a result, tmVar 3.0 achieves superior variant recognition and normalization. In the future, we would like to further enhance and expand tmVar by better linking variants with other closely related concepts such as drugs and diseases.

Funding

This work was supported by the National Institutes of Health Intramural Research Program, National Library of Medicine and in part by the NIH NHGRI Clinical Genome Resource (ClinGen) grant U24 HG009649.

Conflict of Interest: none declared.

References

Allot,A. *et al.* (2018) LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. *Nucleic Acids Res.*, **46**, W530–W536.

Bonis,J. *et al.* (2006) OSIRIS: a tool for retrieving literature about sequence variants. *Bioinformatics*, **22**, 2567–2569.

Caporaso,J.G. *et al.* (2007) MutationFinder: a high-performance system for extracting point mutation mentions from text. *Bioinformatics*, **23**, 1862–1865.

Cejuela,J.M. *et al.* (2017) nala: text mining natural language mutation mentions. *Bioinformatics*, **33**, 1852–1858.

Cheng,C. *et al.* (2020) DeepVar: an end-to-end deep learning approach for genomic variant recognition in biomedical literature. *Proc. AAAI Conf. Artif. Intell.*, **34**, 598–605.

Lee,K. *et al.* (2021) Recent advances of automated methods for searching and extracting genomic variant information from biomedical literature. *Brief. Bioinform.*, **22**, bbaa142.

Nie,A. *et al.* (2019) LitGen: Genetic literature recommendation guided by human explanations. *Pac. Symp. Biocomput.* 2020, **25**, 67–78.

Pawliczek,P. *et al.*; Clinical Genome (ClinGen) Resource. (2018) ClinGen allele registry links information about genetic variants. *Hum. Mutat.*, **39**, 1690–1701.

Thomas,P. *et al.* (2016) SETH detects and normalizes genetic variants in text. *Bioinformatics*, **32**, 2883–2885.

Wei,C.-H. *et al.* (2013) tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics*, **29**, 1433–1439.

Wei,C.-H. *et al.* (2015) GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. *Biomed Res. Int.*, **2015**, 918710.

Wei,C.-H. *et al.* (2016) Beyond accuracy: creating interoperable and scalable text-mining web services. *Bioinformatics*, **32**, 1907–1910.

Wei,C.-H. *et al.* (2017) tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine. *Bioinformatics*, **34**, 80–87.