English

Research Express@NCKU Volume 26 Issue 7 - May 23, 2014
[http://research.ncku.edu.tw/re/articles/e/20140523/3.html]

tmVar: A text mining approach for extracting sequence variants in biomedical literature

Chih-Hsuan Wei[1,2], Bethany R. Harris[3], Hung-Yu Kao[2,*] , Zhiyong Lu[1]

[1] National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), 8600 Rockville Pike, Bethesda, Maryland 20894, USA.
[2] Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, Republic of China.
[3] UCI Libraries, University of California, Irvine, California, USA.

hykao@mail.ncku.edu.tw

Bioinformatics, doi: 10.1093/bioinformatics/btt156

Sequence variation plays a critical role in how genes and gene products connect to human health or disease. In recent years, several text mining studies have attempted to develop and use sequence variation recognition tools to process the biomedical literature and have obtained reasonably successful results. However, most of these tools only handle formal variation mentions of a specific type of variation: substitutions. Based on this reason, we have developed a robust variation extraction method, **tmVar**[1], which can handle most variation types and informal variation mentions (i.e., those that do not conform to any standard variation nomenclature guidelines). We applied a conditional random field model to identify the variation mentions in biomedical text and the assembled components (i.e., wild type, sequence position, and mutant) of the mentions. In our benchmarking, we have compared our method against MutationFinder [2], a state-of-the-art literature mining tool on mutation information extraction. Our results show that our method coverage of variation types is better than MutationFinder.

We defined this issue as a sequence labeling problem which is identifying the location of sequence variation (e.g., p.Pro184ArgfsX19) in biomedical literature. Unlike to previous studies [2-4], we proposed a multiple-states conditional random field model to recognize the 11 components in variation formula. An example in Figure 1 shows two variation formulas (i.e., c.2708_2711delTTAG and p.V903GfsX905) have been identified. Compare to original BIO model (B: Begin, I: Inside and O: Outside), we defined 11 states for variation components (e.q., Reference sequence (A); Mutation position (P); Mutation type (T); wild type (W); Mutant (M); Frame shift (F); Frame shift position (S); Duplication time (D); SNP (R); Other inside mutation tokens (I); Outsid token (O). In our evaluation, using multiple states can effectively improve the performance for 5%.

Furthermore, tmVar can extract a wide range of sequence variants described at protein, DNA, and RNA levels according to a standard nomenclature developed by the Human Genome Variation Society (HGVS). However, about 77% variation mention in literature did not follow HGVs guideline. Therefore, we cover several important types of variations that were not considered in past studies. Using a novel CRF label model and feature set, our method achieves higher performance than a state-of-the-art method on both our corpus (92.2% vs. 78.1% in F-measure) and their own gold standard (93.9% vs. 89.4% in F-measure). These results suggest that tmVar is a high-performance method for sequence variation extraction from biomedical literature.

...	one	family	(	c	.	2708	_	2711	del	TTAG	,	p	.	V	903	G	Fs	X	905	)	...
O	O	O	O	A	I	P	P	P	T	M	O	A	I	W	P	M	F	F	S	O	O

Figure 1. An example of mutation component labels in an excerpt "... (c.2708_2711delTTAG, p.V903GfsX905) ..." in PMID: 22042570. Each cell in the top row represents a token in our processing.

Reference

1. C.-H. Wei, *et al.* , "tmVar: A text mining approach for extracting sequence variants in biomedical literature," *Bioinformatics*, vol. Published, 2013.
2. J. G. Caporaso, *et al.* , "MutationFinder: a high-performance system for extracting point mutation mentions from text," *Bioinformatics*, vol. 23, pp. 1862-1865, 2007.
3. E. Doughty, *et al.* , "Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature.," *Bioinformatics*, vol. 27, pp. 408-415, 2011.
4. L. I. Furlong, et al., "OSIRISv1.2: a named entity recognition system for sequence variants of genes in biomedical literature.," BMC Bioinformatics, vol. 2008, p. 84, 2008.