中文

成大研發快訊　第二十六卷 第七期 - **2014**年五月二十三日
[http://research.ncku.edu.tw/re/articles/c/20140523/3.html]

tmVar: 在生醫文獻上的序列變異擷取探勘

魏至軒[1,2], Bethany R. Harris[3], 高宏宇[2,*], 陸致用[1]

[1] National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), 8600 Rockville Pike, Bethesda, Maryland 20894, USA.
[2] 國立成功大學 資訊工程學系
[3] UCI Libraries, University of California, Irvine, California, USA.

hykao@mail.ncku.edu.tw

Bioinformatics, doi: 10.1093/bioinformatics/btt156

序列變異基因與人類健康或疾病的關聯性議題上是扮演著相當重要的角色。近年來，一些文獻挖掘的研究試圖開發序列變異識別(sequence variation recognition)工具從生醫文獻擷取變異名詞，並已取得相當不錯的結果。然而，大多數的工具都只能處理多種突變類型當中的"替換"(substitutions)的正規變化。基於這個原因，我們開發出一個強大的變異擷取系統，tmVar[1]，它能夠處理大多數的變異類型(包括不符合標準的變化命名)。我們採用了條件隨機域(conditional random field)模型來識別變異中的組成元件(包括,變異點,位置,變異後元素)。在我們的效能評估中，我們跟目前最好的變異擷取工具MutationFinder [2]做評比。我們的方法所能處理的變異類型多於MutationFinder。

我們將這個問題定義成一個序列標註的問題，在生物醫學文獻中找到序列變異出現的位置（例如，p.Pro184ArgfsX19）。不同於以往的研究[2-4]，我們提出了一個多狀態條件隨機場模型(multiple-states conditional random field model)來識別變異中的11個元件。如Figure 1所示的2個變異（c.2708_2711delTTAG, p.V903GfsX905）的例子。跟普通的BIO (B: Begin, I: Inside and O: Outside)模型做比較，我們針對變異的元件定義了11個狀態，包括: 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)，在我們的效能評估中，使用多個狀態，能有效的提高5%的效果。

此外，tmVar可以擷取蛋白質，DNA，RNA的變異，根據人類基因組變異協會(Human Genome Variation Society ,HGVS)所開發的標準命名法，描述了正規的序列變異寫法。但有77%的變異並沒有根據HGVs guideline去寫，因此，我們建立了一個新的條件隨機域模型來解決這個問題，跟MutationFinder做比較，我們取得了更好的效果。這些結果表明，tmVar是一個在生物醫學文獻中提取的序列變異的高性能方法。

...	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. 變異元件擷取的例子 "... (c.2708_2711delTTAG, p.V903GfsX905) ..."(PMID: 22042570).

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.

Copyright 2013 National Cheng Kung University