*Bioinformatics*, 2016, 1–3
doi: 10.1093/bioinformatics/btw234
Advance Access Publication Date: 2 June 2016
Applications Note

OXFORD

Data and text mining

# SETH detects and normalizes genetic variants in text

**Philippe Thomas,[1,2,*] Tim Rocktäschel,[3] Jörg Hakenberg,[4] Yvonne Lichtblau[2] and Ulf Leser[2,*]**

[1]Language Technology Lab, DFKI Berlin, Germany, [2]Knowledge Management in Bioinformatics, Institute for Computer Science, Humboldt-Universität Zu Berlin, Unter Den Linden 6, Berlin 10099, Germany, [3]University College London, Gower Street, LondonWC1E 6BT, UK and [4]Illumina, Inc, 451 El Camino Real, Santa Clara, CA 95050, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on December 28, 2015; revised on March 26, 2016; accepted on April 18, 2016

## Abstract

**Summary**: Descriptions of genetic variations and their effect are widely spread across the biomedical literature. However, finding all mentions of a specific variation, or all mentions of variations in a specific gene, is difficult to achieve due to the many ways such variations are described. Here, we describe SETH, a tool for the recognition of variations from text and their subsequent normalization to dbSNP or UniProt. SETH achieves high precision and recall on several evaluation corpora of PubMed abstracts. It is freely available and encompasses stand-alone scripts for isolated application and evaluation as well as a thorough documentation for integration into other applications.
**Availability and Implementation**: SETH is released under the Apache 2.0 license and can be downloaded from http://rockt.github.io/SETH/.
**Contact**: thomas@informatik.hu-berlin.de or leser@informatik.hu-berlin.de

## 1 Introduction

Over the last decades, a wealth of information about genetic variations and their effect on phenotypes has been published in the scientific literature. However, finding such variations, the genes they affect, and the associated effects on phenotypes is cumbersome due to the many ways in which variations can be described. Despite efforts towards defining nomenclature guidelines for genetic variants (den Dunnen and Antonarakis, 2000), authors often use descriptive natural language instead of following a suggested formalism, which incurs ambiguity and uncertainty in which variation actually is described. This severely hinders the usage of published variation data for subsequent analyses such as biological and clinical interpretation. Several tools have been developed to facilitate the identification of genetic variants in text (for a recent evaluation we refer to Jimeno Yepes and Verspoor 2014b), but none of them is freely available, recognizes and normalizes mentions to an acknowledged nomenclature, links variations to structured databases (e.g. dbSNP or UniProt), and covers a large range of mutation types. SETH is an open source tool unifying all these features. Furthermore, SETH achieves consistent high precision and recall on several publicly available gold standard corpora.

## 2 Methods

SETH's workflow is depicted in Figure 1. It consists of two main components: (i) Recognition of genetic variants mentioned in text (see Section 2.1) and (ii) mapping of recognized variants to dbSNP or UniProt identifiers (see Section 2.2).

### 2.1 Named-entity recognition

SETH relies on four complementary modules to recognize genetic variants.

i. *Nomenclature EBNF*: Variation mentions adhering to the human mutation nomenclature, described by den Dunnen and Antonarakis (2000), are recognized by an Extended Backus–Naur (EBNF) grammar which was originally published by Laros *et al.* (2011). We adapted this grammar to also detect

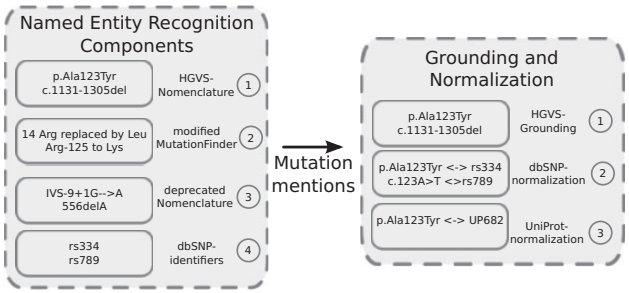


**Fig. 1.** SETH uses multiple strategies to recognize and subsequently normalize genetic variations (see text for details)

frequently observed deviations from the nomenclature ('p.Arg282 >Gln' instead of 'p.Arg282Gln').

ii. *MutationFinder:* Short phrases describing single nucleotide substitutions are recognized using a modified version of MutationFinder (Caporaso *et al.*, 2007). We included patterns detecting nonsense mutations ('Ala15Ter', 'A15X'), DNA substitutions ('-650A > T'), and mutations using ambiguous amino acid letter codes ('Asx' for Aspartate or Asparagine).

iii. *Nomenclature RegExes:* Using a set of regular expressions, SETH recognizes deletions ('852 + 123delT'), insertions, substitutions ('IVS123-12A- >T'), and frameshift mutations ('A123fsX1') following nomenclature recommendations from (Ad Hoc Committee on Mutation Nomenclature, 1996). Regular expressions have been developed by implementing a test-driven approach integrating variation mentions from eleven different proposals for a mutation nomenclature.

iv. *Literal mentions:* Literal mentions of dbSNP identifiers ('rs334', 'rs334:A > C') are recognized using regular expressions.

SETH merges mutations found by more than one NER-module into a single mention. This strategy helps to resolve ambiguous cases by favoring matches against the most rigorous and recent nomenclature over others. For example, MutationFinder might detect a substring of a variation, while the EBNF grammar detects the full mutation mention. All recognized mutation mentions are categorized into one of the following: substitution, deletion, insertion, duplication, insertion-deletion (insdel), inversion, conversion, translocation, frameshift, short-sequence repeat or literal dbSNP mention.

### 2.2 Named-entity normalization

Following the initial recognition of mutation mentions, SETH links variations to a unique identifier of the dbSNP (for single nucleotide variations) or UniProt database (regarding the protein containing the mutation) if possible. To disambiguate between multiple possible identifiers, this step requires additional information about genes or proteins mentioned in the same article. SETH currently is configured to work with PubMed abstracts for which it uses precomputed results from the gene recognition tool GNAT (Hakenberg *et al.*, 2011) and NCBI's gene2pubmed (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/). However, the modular implementation of SETH makes it simple to replace the gene recognition tool to be used, in order to adapt to other types of texts, such as patents or clinical trial reports.

The normalization procedure uses the heuristics discussed by Thomas *et al.* (2011). These heuristics account for frequent differences between textual and structured database information by performing a physical validation. This validation technique compares detected variations with the DNA coding sequence (CDS) or protein sequence in the respective dbSNP or UniProt database, taking into account systematic differences like post-translational modifications or inversed DNA strands.

SETH rewrites mentions of variations using a HGVS-compliant representation, substituting, for instance, 'ΔF508' with 'p.Phe508del'. This is useful for mapping large repositories of mutation mentions into the latest mutation nomenclature, helping to decrease ambiguity.

**Table 1.** SETH performance for named entity recognition and normalization to dbSNP identifiers on various corpora. Performance of NER is calculated using exact matching

| | Corpus | *P* | R | $F_1$ | No. doc | No. entities |
|---|---|---|---|---|---|---|
| NER | Caporaso *et al.* -dev | 0.98 | 0.83 | 0.90 | 305 | 550 |
| | Caporaso *et al.* -test | 0.98 | 0.82 | 0.89 | 508 | 907 |
| | Wei *et al.* -dev | 0.93 | 0.80 | 0.86 | 334 | 961 |
| | Wei *et al.* -test | 0.95 | 0.77 | 0.85 | 166 | 470 |
| | Corpus of this work | 0.98 | 0.86 | 0.91 | 630 | 895 |
| NEN | Furlong *et al.* | 0.94 | 0.69 | 0.79 | 105 | 261 |
| | Thomas *et al.* | 0.96 | 0.58 | 0.72 | 296 | 527 |

## 3 Evaluation

We assessed the performance of SETH on a series of publicly available corpora for named entity recognition and dbSNP-normalization (see Table 1). All these corpora are annotated on abstract level and performance of named entity recognition was evaluated using an exact matching strategy. We also annotated a text corpus of 630 publications with mutation mentions by sampling papers from the journals *The American Journal of Human Genetics* and *Human Mutation* published between 1992 and 2012. To ensure a high variability of textual variation expressions, we uniformly sampled 30 publications per year. This corpus is freely available from the SETH project web page. Regarding identification of variations, SETH achieves $F_1$ scores over 85% across several corpora, demonstrating its robustness and independence from particularities of the curation processes in these different corpora. In their own evaluation on two publicly available test corpora, tmVar (Wei *et al.*, 2013) outperforms SETH by 4–5% points for NER. However, SETH has been compared with four freely available tools for mutation recognition in intrinsic and extrinsic experiments by Jimeno Yepes and Verspoor (2014b). In their work, the authors evaluated MutationFinder (Caporaso *et al.*, 2007), EMU (Doughty *et al.*, 2011), OMM (Naderi and Witte, 2012), tmVar (Wei *et al.*, 2013) and SETH. In the intrinsic set-up, SETH outperformed all competitors in terms of $F_1$ on the Variome full-text corpus (Jimeno Yepes and Verspoor, 2014a) using exact matching for evaluation. In the extrinsic set-up, where only coverage was measured, SETH achieved the best result for the COSMIC database (Bamford *et al.*, 2004).

Finally, the performance of normalization to human dbSNP identifiers was assessed on two publicly available corpora (Furlong *et al.*, 2008; Thomas *et al.*, 2011). The first corpus provides annotations for gene name, which were used in our normalization experiments. For the second corpus from Thomas *et al.* we used the precomputed GNAT and gene2pubmed results.

## 4 Conclusion

We introduced SETH, an open-source tool for recognizing and normalizing mentions of variations in natural language text. By rewriting

variants using the HGVS nomenclature and normalization to dbSNP, SETH facilitates indexing and information retrieval of heterogeneous occurrences of the same biological variant across the literature. SETH achieves state-of-the-art performance and is the only freely available tool performing physical validation not only for protein sequence mutations but also for DNA variations. Variation mentions recognized by SETH in PubMed and PubMed Central can be browsed in GeneView (Thomas *et al*., 2012) at http://bc3.informatik.hu-berlin.de/.

## Funding

The development of SETH was supported by the BMBF-funded projects: ColoNet, VirtualLiver, and OncoPath.

*Conflict of Interest*: none declared.

## References

Ad Hoc Committee on Mutation Nomenclature. (1996) Update on nomenclature for human gene mutations. *Hum. Mutat*., **8**, 197–202.

Bamford,S. *et al*. (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br. J. Cancer*, **91**, 355–358.

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

den Dunnen,J.T. and Antonarakis,S.E. (2000) Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum. Mutat*., **15**, 7–12.

Doughty,E. *et al*. (2011) Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature. *Bioinformatics*, **27**, 408–415.

Furlong,L.I. *et al*. (2008) OSIRISv1.2: a named entity recognition system for -sequence variants of genes in biomedical literature. *BMC Bioinformatics*, **9**, 84.

Hakenberg,J. *et al*. (2011) The GNAT library for local and remote gene mention normalization. *Bioinformatics*, **27**, 2769–2771.

Jimeno Yepes,A. and Verspoor,K. (2014a) Literature mining of genetic variants for curation: quantifying the importance of supplementary material. *Database (Oxford)*, **2014**, bau003.

Jimeno Yepes,A. and Verspoor,K. (2014b) Mutation extraction tools can be combined for robust recognition of genetic variants in the literature. *F1000Res*, **3**, 18.

Laros,J.F.J. *et al*. (2011) A formalized description of the standard human variant nomenclature in Extended Backus-Naur Form. *BMC Bioinformatics*, **12(Suppl 4)**, S5.

Naderi,N. and Witte,R. (2012) Automated extraction and semantic analysis of mutation impacts from the biomedical literature. *BMC Genomics*, **13(Suppl 4)**, S10.

Thomas,P. *et al*. (2011) Challenges in the association of human single nucleotide polymorphism mentions with unique database identifiers. *BMC Bioinformatics*, **12b(Suppl 4)**, S4.

Thomas,P. *et al*. (2012) GeneView: a comprehensive semantic search engine for PubMed. *Nucleic Acids Res*., **40(Web Server issue)**, W585–W591.

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