Advances in Informatics, Management and Technology in Healthcare
J. Mantas et al. (Eds.)

doi:10.3233/SHTI220735

Natural Language Processing Approaches for Retrieval of Clinically Relevant Genomic Information in Cancer

Taxiarchis BOTSIS[a,1], Joseph MURRAY[a], Alessandro LEAL[a], Doreen PALSGROVE[a], Wei WANG[a], James R. WHITE[a], Victor E. VELCULESCU[a], the Johns Hopkins Molecular Tumor Board Investigators[a] and Valsamo ANAGNOSTOU[a]
[a] *The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA*

Abstract. The accelerating impact of genomic data in clinical decision-making has generated a paradigm shift from treatment based on the anatomic origin of the tumor to the incorporation of key genomic features to guide therapy. Assessing the clinical validity and utility of the genomic background of a patient's cancer represents one of the emerging challenges in oncology practice, demanding the development of automated platforms for extracting clinically relevant genomic information from medical texts. We developed PubMiner, a natural language processing tool to extract and interpret cancer type, therapy, and genomic information from biomedical abstracts. Our initial focus has been the retrieval of gene names, variants, and negations, where PubMiner performed highly in terms of total recall (91.7%) with a precision of 79.7%. Our next steps include developing a web-based interface to promote personalized treatment based on each tumor's unique genomic fingerprints.

Keywords. Natural language processing, cancer, actionable genomic alterations

1. Introduction

Precision oncology is based on our increased knowledge of the cancer mutation repertoire and the availability of agents that target specific genes or pathways. The anticipated widespread adoption of large-scale genomic analyses in precision oncology mandates the development of a gene actionability framework to best evaluate the clinical utility of genomic technologies. While publicly available databases facilitate the association between mutations and response to therapy, these databases have been limited to analyses of individual genes and alterations using methods that are largely dependent on expert opinion and manual curation of the literature. Despite the potential of predictive genomic testing, a recent study reported that a quarter of physicians at an academic tertiary-care comprehensive cancer center reported low confidence in interpreting genomic sequencing data, highlighting the need to develop comprehensive, evidence-based strategies to enhance physician genomic education [1].

Furthermore, the massive production of biomedical information in the last decades has led to an overwhelming increase in scientific publications. Building automated

[1] Corresponding Author, Taxiarchis Botsis, The Sidney Kimmel Comprehensive Cancer Center Johns Hopkins University School of Medicine, USA; E-mail: tbotsis1@jhmi.edu.

methodologies for the processing of biomedical literature has been extensively pursued [2, 3]; however, no database currently captures the compendium of potentially actionable genes in the human genome. Here we report the development and prediction accuracy of a natural language processing (NLP) platform to extract and characterize cancer-specific genomic alterations in the published biomedical literature (PubMed). We envision that such approaches will facilitate incorporating scientific knowledge in precision oncology.

2. Methods

We developed PubMiner, an NLP platform to mine biomedical abstracts from PubMed and extract treatment, cancer type, gene (symbols. names and aliases), mutations, and negation statements. PubMiner is a rule-based system that performs the information extraction based on lexical resources containing all cancer types from the National Center for Biotechnology Information (NCBI); therapeutic agents from the National Cancer Institute drug dictionary; gene names, symbols, and aliases from the Entrez Global Query Cross-Database Search System at NCBI; and genomic alterations based on mutation nomenclature of the Human Genome Variation Society, the Sequence Ontology terms and a manually curated list of terms generated to address the vast variability in published mutation nomenclature. PubMiner was built in Python 3.x using the Natural Language Toolkit 3.x and Biopython 1.7.

We generated a user-defined automated PubMed query to identify studies reporting clinically relevant genomic alterations in cancer. The query was optimized to select studies with genomic alterations and their association with the outcome for cancer patients. Only studies annotated as clinical trials were selected for further review to derive evidence generated from uniformly treated patients under standardized conditions. We initially focused on the *KRAS* and *EGFR* genes that are commonly mutated oncogenes in human cancer. Our team of medical experts defined the fields to be extracted from each abstract and all abstracts were reviewed by three independent reviewers (AL, DP, JM).

We subsequently developed a prototype version of PubMiner based on requirements generated by the medical experts. This prototype version was optimized in multiple iterations by using a convenience set of abstracts for *KRAS* (N=467) and *EGFR* (N=652), resulting in recall >90% for cancer type and treatment. We then focused on refining the prototype version and developing a system that captures any given genomic alteration and accounts for variability in published mutation nomenclature. The same pool of *KRAS* and *EGFR* abstracts was used to: (1) develop a set of annotation guidelines that supported the generation of an annotated corpus in the next step; (2) familize the two annotators (AL, DP) with the guidelines and the annotation tool used for this process; and, (3) reach an inter-annotator agreement (IAA) of >90%. The latter was important to allow for solo annotations towards generating the annotated corpus, i.e., each abstract was reviewed and annotated by one expert only without any adjudication conducted [4].

IAA was evaluated in two corpus pre-generation phases over two small subsets of 20 and 10 abstracts randomly selected from the above corpus. In both phases, the two medical experts (AL, DP) annotated the same abstracts by identifying 3 named entities: GENEs, VARIANTs, and NEGATIONs. Inclusion of negations was considered important as there are instances where the absence of a genomic alteration determines clinical actionability (e.g., response to anti-*EGFR* therapy in colon cancer without *KRAS* mutations). We then compared their annotations to identify the full and partial matches.

For a full match (FM) both the label and text annotations should be exactly the same. For a partial match (PM) we allowed only one token-difference in the text annotation; labels could differ in the PM. Everything else was considered a mismatch (MM). IAA was equal to the sum of the full and the partial matches. To facilitate annotation, we developed an annotation tool in Python 3.x that supported text selection and tagging.

We subsequently ran the PubMed query for two independent oncogenes, *PIK3CA* and *AKT1* and retrieved all related abstracts. A reference standard was generated by manual annotation of these PubMed abstracts and the abstracts were randomly split into two subsets, one for each annotator. Annotators used the annotation tool to identify the same named entities as in the pre-generation phase. The generated corpus was randomly divided into a training and a testing set by applying the 20-80 rule to each of the two annotated subsets. The training set was used to refine the PubMiner prototype and build the final version that was evaluated over the testing set.

3. Results

The IAA in the first corpus pre-generation phase was equal to 86.2%, with FMs and PMs at 77.3% and 8.9%, respectively. IAA was equal to 91% in the second pre-generation phase, with FMs and PMs at 82.1% and 8.9%, respectively. After achieving an IAA of >90%, we proceeded with the generation of the reference standard. This was conducted in three steps. First, we retrieved all abstracts for *PIK3CA* and *AKT1* (N=219) as described above. Second, we split the corpus into two subsets (110 and 109 abstracts each). Third, each annotator worked on a single subset and identified the GENE, VARIANT, and NEGATION entities using an annotation tool developed for this task.

The reference standard was randomly divided into a training and a testing set by applying the 20-80 rule to each of the two annotated subsets. The training set was used to refine the PubMiner prototype by maximizing F-measure. The final version was evaluated over the testing set. In terms of full and partial matches, recall for annotator 1 and annotator 2 testing subsets was 90.2% and 93.1%, respectively; precision was 78.5% and 80.9%, respectively; and, F-measure was 0.839 and 0.866, respectively. Also, the type match (i.e., whether PubMiner assigned the right entity to extracted term) for Annotator 1 and Annotator 2 testing subsets was 94.1% and 97.8%, respectively. Table 1 includes the detailed breakdown of FMs, PMs, and MMs per annotator and entity type. Most MMs were observed in the GENE entity (both testing subsets) and the VARIANT entity (Annotator 2 testing subset). These findings and potential reasons for the GENE and VARIANT MMs were explored in the qualitative error analysis.

Table 1. Matches over the Testing Subsets. FM: Full Match; PM: Partial Match; MM: MisMatch

		GENE	VARIANT	NEGATION
FM	Annotator 1	809	531	12
FM	Annotator 2	974	556	17
PM	Annotator 1	48	74	3
PM	Annotator 2	30	31	0
MM	Annotator 1	96	62	3
MM	Annotator 2	92	23	3

For the qualitative error analysis, we initially counted the unique annotations missed by PubMiner. Annotations occurring once were excluded from this analysis. Most of the MMs belonged to the GENE category and were largely attributed to either the absence of the corresponding entries in the gene lexicon or the rule that excluded “all gene

symbols and aliases with letters<=4 and only the leading letter capitalized" (e.g., "Akt"). PubMiner also missed VARIANT annotations when they were not reported according to standard nomenclature. We also missed abbreviated VARIANTS, such as "mt" and "mut", when these were found in the abstract without a corresponding gene.

4. Discussion

This study presents PubMiner, a novel NLP tool that extracts clinically-relevant genomic information from PubMed abstracts. PubMiner performed highly in total recall (91.7%) with an acceptable cost in terms of precision (79.7%). In this case, the type match for each testing subset was very high, ranging from 94.1% to 97.8%. Moreover, the differences over the testing subsets for the two annotators were generally marginal except for the FM recall (82.5% vs. 89.6%). Overall, PubMiner reliably retrieved the GENE, VARIANT, and NEGATION annotations.

Our study has certain limitations. The current version of our algorithm does not extract outcome information, such as the prognostic or predictive value of the retrieved genomic alteration. Although our PubMed query was generated such that only studies that report clinically actionable genomic alterations are retrieved, a dedicated NLP approach to retrieve outcome is warranted. We acknowledge both limitations and plan to address them in the next release. Furthermore, one might criticize the absence of any adjudication in generating the reference standard. As we have previously shown, single annotations may work very efficiently when IAA exceeds a certain threshold (90% or more) in the pre-production phase, and we therefore followed this scheme [4].

In summary, we have developed a novel NLP algorithm to extract and characterize cancer-specific genomic alterations in the published biomedical literature. Our approach is not based on expert opinion determined genomic alterations as it was generated on the premise of capturing genome-wide actionable cancer-specific mutations. Future steps include the interpretation of clinical actionability based on levels of evidence spanning mutations in all genes in the human genome. This task will be supported through the detailed patient review conducted in our Molecular Tumor Board that will further validate the collected evidence. Ultimately, we envision that our efforts would be translated into web-based portals accommodating both healthcare providers and patients to promote personalized treatment based on each tumor's unique genomic fingerprints.

References

[1] Gray SW, Gollust SE, Carere DA, Chen CA, Cronin A, Kalia SS, Rana HQ, Ruffin IV MT, Wang C, Roberts JS, Green RC. Personal genomic testing for cancer risk: results from the impact of personal genomics study. Journal of Clinical Oncology. 2017 Feb 20;35(6):636.

[2] Kulick S, Bies A, Liberman M, Mandel M, McDonald R, Palmer M, Schein A, Ungar L, Winters S, White P. Integrated annotation for biomedical information extraction. InHLT-NAACL 2004 Workshop: Linking Biological Literature, Ontologies and Databases 2004. p. 61-68.

[3] Collier N, Park HS, Ogata N, Tateisi Y, Nobata C, Ohta T, Sekimizu T, Imai H, Ibushi K, Tsujii JI. The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers. InNinth Conference of the European Chapter of the Association for Computational Linguistics 1999 Jun. p. 271-272.

[4] Foster M, Pandey A, Kreimeyer K, Botsis T. Generation of an annotated reference standard for vaccine adverse event reports. Vaccine. 2018 Jul 5;36(29):4325-30.