Text-mining Services of the Swiss Variant Interpretation Platform for Oncology

Déborah CAUCHETEUR[a,b,*,1], Julien GOBEILL[a,b,*], Anaïs MOTTAZ[a,b,*], Emilie PASCHE[a,b,*], Pierre-André MICHEL[a,b], Luc MOTTIN[a,b], Daniel J. STEKHOVEN[b,c], Valérie BARBIÉ[b] and Patrick RUCH[a,b]
[a] *HES-SO / HEG Geneva, Information Sciences, Geneva, Switzerland*
[b] *SIB Swiss Institute of Bioinformatics, Geneva, Switzerland*
[c] *NEXUS Personalized Health Technologies, ETHZ, Zürich, Switzerland*

Abstract. The Swiss Variant Interpretation Platform for Oncology is a centralized, joint and curated database for clinical somatic variants piloted by a board of Swiss healthcare institutions and operated by the SIB Swiss Institute of Bioinformatics. To support this effort, SIB Text Mining designed a set of text analytics services. This report focuses on three of those services. First, the automatic annotations of the literature with a set of terminologies has been performed, resulting in a large annotated version of MEDLINE and PMC. Second, a generator of variant synonyms for single nucleotide variants has been developed using publicly available data resources, as well as patterns of non-standard formats, often found in the literature. Third, a literature ranking service enables to retrieve a ranked set of MEDLINE abstracts given a variant and optionally a diagnosis. The annotation of MEDLINE and PMC resulted in a total of respectively 785,181,199 and 1,156,060,212 annotations, which means an average of 26 and 425 annotations per abstract and full-text article. The generator of variant synonyms enables to retrieve up to 42 synonyms for a variant. The literature ranking service reaches a precision (P10) of 63%, which means that almost two thirds of the top-10 returned abstracts are judged relevant. Further services will be implemented to complete this set of services, such as a service to retrieve relevant clinical trials for a patient and a literature ranking service for full-text articles.

Keywords. Precision medicine, literature, variant, terminology, text-mining

1. Introduction

After several years of collaboration to improve and harmonize the NGS (next-generation sequencing) practices in somatic mutation calling, a number of Swiss hospitals, pathology institutes and the Swiss Institute of Bioinformatics members have pointed to a set of shortcomings, most prominently the lack of a central repository for clinically verified variant annotations in cancer. To support the harmonization of variant annotation in diagnosis, to provide a centralized curated database of somatic variants coming from

[1] Corresponding Author: Deborah Caucheteur, Campus Battelle, Bâtiment B, Rte de la Tambourine 17, 1227 Carouge, Switzerland; E-mail: deborah.caucheteur@hesge.ch.
* contributed equally

Swiss hospitals, and to enable a seamless interaction between participating institutes and global initiatives, the Swiss Variant Interpretation Platform for Oncology (SVIP-O) was developed [1].

The curation of variants is a labour-intensive process. First, NGS analyzes result in a huge amount of variants in the patient sample. Second, for each variant, many sources must be consulted, including the scientific literature. While for some already well-studied variants (e.g. BRAF V600E), identifying a set of relevant scientific literature is not an issue, for variants of uncertain significance (VUS), the task can be more difficult. Indeed, the curator must comprehensively gather the relevant literature to assign a standardized tier level (ASCO/AMP/CAP guidelines), in which the variant, the gene or the diagnosis could have been labelled in different ways [2]. Thus, the curator must author multiple queries to avoid missing out on an important paper. Moreover, when large sets of literature are available, triage of the literature (i.e. selection of relevant papers as well as rejection of irrelevant papers) can be a very time-consuming task [3].

Therefore, we developed a set of text analytics services intended to facilitate and improve the comprehensive collection of literature to support further processing steps, which will include capture of textual evidence by Swiss-Prot curators and clinical validation by medical specialists for final storage in the SVIP-O knowledgebase. First, we developed a service to annotate the clinically relevant information contained in the scientific literature. This service enables not only to accelerate searches (i.e. queries are performed using unique identifiers in a standardized field), but also to increase the recall (i.e. the annotation step conflates synonyms and string variants). However, the annotation of variants faces a serious limitation: no universal terminology is available. In order to offset this issue, variants can be searched in literature using synonym expansion. Thus, as a second task, we propose a service to expand a variant name with a set of expressions, including Human Genome Variation Society (HGVS) standard descriptions for the protein, transcript and genomic DNA levels, but also non-standard formats found in the literature (e.g. V-600-E or BRAFV600E) [4]. Finally, a service collecting and prioritizing literature is available: it combines heterogeneous information retrieval results for an optimal article triage, which altogether can reduce the triage effort by a factor of 3 [5].

2. Methods

Our MEDLINE and PMC collections consisted of respectively 30,415,832 and 2,632,396 documents (in January, 7th 2020), daily updated and loaded into a MongoDB document database and then into an ElasticSearch index.

2.1. Recognition of terminological entities in the literature

The collection was annotated with codes from various terminologies and ontologies, such as neXtProt [6] for genes, Drugbank [7] and WHO-ATC [8] for drugs, NCI Thesaurus [9] and ICD-O-3 [10] for diseases, and HPO (Human Phenotype Ontology) for phenotypes. Querying MEDLINE through annotations is not only faster because the indexes are pre-computed but also it results in a better recall as each occurrence of a concept receives one unique ID for all its synonyms. We also applied string pre-processing methods. For instance, if dashes are present, they are treated to form a “new”

word (e.g. “AB-C” becomes “AB”, “C” and “ABC”). Papers which contain only the word without dash will thus be retrieved.

2.2. Generation of synonyms for variants

While many databases of polymorphisms and variants exist, such as ClinVar, COSMIC or dbSNP, using those resources as terminologies is fairly challenging. They describe variants using a standard nomenclature recommended by the HGVS [11]. These standards require a precise syntax and a reference sequence on which the variation is described to avoid uncertainty about the position of the change. Both of them are rarely observed in publications. Depending on the database, variants entries are also centered on different levels: genomic, transcript or protein, which are not true synonyms.

Therefore, we developed our own synonym generation tool that enables us in addition to annotate variants not necessarily present in those databases. Our efforts focused until now on SNPs. Our tool generates synonyms given a gene and a variant. It includes a validation step, where we check whether the given base or amino acid exists at the given position. Then, we compute the description of the variant at the other levels, using the Mutalyzer tool [12]. We finally generate synonyms with many syntactic variations as encountered in the literature [4]. We also extend the search to all possible mutations in the case the replacing amino acid is not defined.

2.3. Prioritization of literature

The report focuses on the literature triage tasks, i.e. the ability to rank MEDLINE abstracts - as opposed to full-text articles - to support further curation steps. Triage is usually performed on abstracts, whose content is sufficient to help a domain expert to decide whether a particular report needs in-depth reading or not.

Our literature prioritization system is based on two steps: collecting a complete set of abstracts and reranking the MEDLINE set. A most complete set of abstracts related to a particular triplet (i.e. a variant in a gene for a specific diagnosis) is built by the intersection of several queries’ output: normalized entities (i.e. gene and diagnosis) are searched with unique identifiers within the MEDLINE annotations, while a keyword search expanded with synonyms is performed within free texts from MEDLINE for not normalized entities (i.e. variant). Moreover, a set of queries, with decreasing levels of specificity (e.g. abstracts not mentioning diagnosis) is also performed. Results are linearly combined with previous abstracts set. Then, we apply different strategies to re-rank the MEDLINE set: 1) based on the number of occurrences of some annotated entities (e.g. abstracts mentioning drugs); 2) based on demographic information; 3) based on a set of keywords that should (e.g. treat) or should not (e.g. marker) be present in the abstracts and 4) based on the breadth of treatments returned in the top articles. The last re-ranking strategy aims at avoiding that all top returned articles are related to the same treatment, but rather favoring abstracts that are discussing different treatment options.

The system is evaluated following standard TREC procedures using TREC PM 2019 benchmarks [13], comprising 40 synthetic patient cases, each consisting of a disease, a variant, a gene and some demographic information.

3. Results

3.1. Recognition of terminological entities in the literature

Thanks to our annotations, we are able to retrieve more papers than classical queries on PubMed. The query *BRAF* retrieves 14,099 papers on PubMed versus 14,952 papers on our index with his corresponding code *NX_P15056*. Another example, query *non-small cell lung cancer* retrieves 51,858 papers on PubMed versus 70,972 (code *C2926*) with our index. These results are explained by expansions provided with annotations: synonyms (*BRAF1* or *RAFB1*; *NSCLC*) and processing of hyphens (*B-RAF* becomes *BRAF*, thus a paper containing *B-RAF* will be retrieved). Table 1 represents an overview of annotations which are available for our collections for some of the terminologies.

Table 1. Extract of statistics about annotations on MEDLINE/PMC collections obtained in January 2020.

Type	Terminologies	Nb of entities annotated in	
		MEDLINE	**PMC**
Drugs	DrugBank	80,100,2684	86,769,820
Drugs	ATC	46,173,559	30,404,688
Diseases	NCIt	131,577,959	108,049,190
Diseases	ICD-O-3	4,837,713	2,955,200
Genes	neXtProt	36,320,459	91,930,787
Total annotations		785,181,199	1,156,060,212
Average per document		26	425

3.2. Generation of synonyms for variants

Our tool generates 42 synonyms for a given valid variant when successfully mapped to all levels, with 16 synonyms for protein variant, 13 for transcript variant including COSMIC id and 13 for genomic variant including dbSNP id. It increases the retrieval between a few percent and several times the number of publications depending on the variant frequency, with smaller effect for very popular variants. For instance, 708 additional abstracts, over 1715, are retrieved for the variant V617F in JAK2.

3.3. Prioritization of literature

Precision at rank 10 (P10) has been used to evaluate our system. This metric reflects the proportion of relevant documents retrieved in the top ten results. Our system resulted in a P10 of 63%, which means that almost two thirds of the top-10 returned abstracts are judged relevant. This service is publicly available: http://candy.hesge.ch/Variomes/.

4. Discussion

We have thus developed a set of services that can be used to facilitate the process of variant curation and validation by physicians, and in particular molecular pathologists, oncologists and hematologists, as well as rare disease experts. The literature triage service, boosted by the variant expansion service and the MEDLINE annotations service, is able to reduce the search burden by simplifying the paper triage. Indeed, the system's evaluation demonstrated that in the top-10 abstracts proposed by our system, more than

six are relevant for the clinical decision-support task. In addition, such services can increase the possibility of finding a relevant paper. Although this aspect might be marginal for common variants, it is key for rare (or poorly studied) variants for which each publication matters. As an example, the search for the TP53 V143A variant in PubMed results in only seven abstracts, while our system is able to return 34 abstracts. Indeed, while PubMed strictly searches for V143A, our system is also able to search for Val143Ala or 428T>C.

Additional services will be implemented to complete this set of services, such as a literature ranking service for full-text articles. Indeed, scientific abstracts reporting on treatments do not always mention all the information regarding diagnosis, gene and variant, so that full-text articles are needed, as well as supplementary data when available. A service to retrieve relevant clinical trials for a given patient is also under development, enabling to connect patients with experimental treatments.

Acknowledgements. This project has been supported by Swiss Personalized Health Network (SPHN) and BioMedIT fundings, see https://svip.ch/.

References

[1] D.J. Stekhoven, P. Ruch, V. Barbié, Swiss Variant Interpretation Platform for Oncology (SVIP-O), Swiss Med Informatics **34** (2018), 00411.
[2] Z. Shi, B. Gu, F. Popowich, et al., Synonym-based Query Expansion and Boosting-based Re-ranking: A Two-phase Approach for Genomic Information Retrieval, In Proceedings of the Fourteenth Text REtrieval Conference TREC (2005).
[3] T.C. Wiegers, A.P. Davis, K.B. Cohen, et al., Text mining and manual curation of chemical-gene-disease networks for the Comparative Toxicogenomics Database (CTD), BMC Bioinformatics **10** (2009), 326.
[4] Y.L. Yip, N. Lachenal, V. Pillet, et al., Retrieving mutation-specific information for human proteins in UniProt/Swiss-Prot Knowledgebase, J Bioinform Comput Biol **5(6)** (2007), 1215-31.
[5] L. Mottin, J. Gobeill, E. Pasche, et al., neXtA5: accelerating annotation of articles via automated approaches in neXtProt, *Database (Oxford)* (2016).
[6] P. Gaudet, P.A. Michel, M. Zahn-Zabal, et al., The neXtProt knowledgebase on human proteins: 2017 update, *Nucleic Acids Res* **45(D1)** (2017), D177-D182.
[7] D.S. Wishart, Y.D. Feunang, A.C. Guo, et al., DrugBank 5.0: a major update to the DrugBank database for 2018, *Nucleic Acids Res* **46(D1)** (2018), D1074-D1082.
[8] WHO Collaborating Centre for Drug Statistics Methodology, ATC classification index with DDDs, 2019. Oslo, Norway (2018).
[9] N. Sioutos, S. de Coronado, H.W. Haber, et al., NCI Thesaurus: a semantic model integrating cancer related clinical and molecular information, *J Biomed Inform* **40(1)** (2007), 30-43.
[10] A. Fritzl, C. Percy, A. Jack, et al., International classification of diseases for oncology / editors, 3rd ed. *World Health Organization* (2000).
[11] J.T. den Dunnen, R. Dalgleish, D.R. Maglott, et al., HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Hum Mutat* **37(6) (**2016), 564-9.
[12] M. Wildeman, E. van Ophuizen, J.T. den Dunnen, et al., Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker, *Hum Mutat* **29(1)** (2008), 6-13.
[13] K. Roberts, D. Demner-Fushman, E.M. Vorrhees, et al., Overview of the TREC 2018 Precision Medicine Track, *In Proceedings of the Twenty-Seventh Text REtrieval Conference* (2018).