Towards automatic large-scale curation of genomic variation: improving coverage based on supplementary material

Antonio Jimeno Yepes
National ICT Australia
Victoria Research Laboratory
Melbourne, Australia
antonio.jimeno@gmail.com

Karin Verspoor
National ICT Australia
Victoria Research Laboratory
Melbourne, Australia
karin.verspoor@nicta.com.au

Abstract

There are ongoing large-scale efforts to catalog genomic variation related to disease in structured databases. Much of the relevant information is available only from unstructured sources, including the scientific literature. The ability of text mining tools to recover the mutations catalogued in the COSMIC and InSiGHT databases based on the article text has been demonstrated to be far less than what would be expected based on the excellent performance on intrinsic evaluation of mutation extraction tools. We explore the impact of processing tables and supplementary material associated to relevant literature, and find that the coverage of variants improves dramatically, from 2% to over 50%. This result highlights the importance of processing all of the data associated with a publication.

1 Introduction

A major thrust of modern biological research is the understanding of how genomic variation relates to disease. There are large-scale efforts to catalog the results of this research in structured databases, including in the Online Mendelian Inheritance in Man (OMIM) database and the Human Gene Mutation Database (HGMD). Much of this information is available only from unstructured sources, including the scientific literature.

There have been several systems developed to target extraction of mutations and other genetic variation from the literature (Baker and Witte, 2006; Caporaso et al., 2007; Krallinger et al., 2009; Doughty et al., 2011; Naderi and Witte, 2012), *inter alia*. The performance of these tools has been claimed to achieve high precision and recall on *intrinsic* evaluation.

In previous work, we performed an *extrinsic* evaluation of a mutation extraction tool with respect to the task of curation of the literature for the purpose of populating a database of genetic variation information (Jimeno Yepes and Verspoor, 2013). We found that the ability of the text mining tool to recover the mutations catalogued in the databases is far less than what would be expected based on the excellent performance on intrinsic evaluation.

Here, we extend our mutation extraction analysis to include not only the mutations extracted from the abstract and full text of the articles, but also those in supplementary material. Our results show that the coverage obtained by using supplementary material reaches over 50% of the gene-mutation pairs for the analyzed articles, while the coverage with full text alone is approximately 2%. This indicates that the supplementary material is critically important for complete processing of the information available from publications.

2 Methods

We identified two mutation databases that have explicit, curated links to the source literature for individual genetic variants. We accessed that literature, collecting full text where possible, and applied a tool to identify genetic variants in the text.

2.1 The mutation databases

2.1.1 COSMIC database

COSMIC (Bamford et al., 2004) contains comprehensive, curated, information on somatic mutations in human cancer. We used version v62 available from COSMIC's FTP site[1],

[1] ftp://ftp.sanger.ac.uk/pub/CGP/cosmic/

including mutation information curated from 9,950 unique PubMed® citations (referenced via PubMed identifiers, or PMIDs). The database associates 7,868 publications to individual mutations in specific genes. The remaining articles contain non-coding mutations and COSMIC does not record specific mutations. Genes are referenced by name and by HGNC (HUGO Gene Nomenclature Committee) (Povey et al., 2001) identifier.

2.1.2 InSiGHT database

The International Society for Gastrointestinal Hereditary Tumours (InSiGHT) maintains a database of genetic variants for both Lynch Syndrome and Familial Adenomatous Polyposis. The database has curated mutations for four genes: MLH1, MSH2, MSH6 and PMS2.

We accessed the database on 02/Jan/2013. The data includes variants with curated associations linked to 809 PubMed citations. Amino acids in protein variants were normalized to single letter abbreviation form.

2.2 Article collection

We collected the PMIDs available from each of the databases and searched PubMed Central® Open Access (PMC-OA) subset for available articles. We were able to obtain PMC XML files for 13 articles in InSiGHT and 563 in COSMIC. We aim to extract the abstract and article text from these XML files. However, of the 13 InSiGHT PMC XML files, 4 files contained only the abstract with a link to the full text in PDF format. In the COSMIC collection, this issue occurred in 76 of the 563 PMC XML files. We downloaded the PDF articles for these articles to obtain the full text content; the PDF version of the articles contains the article text, references and images and tables that are not included as supplementary material. These PDF versions were downloaded from the European PMC[2], which offers a straightforward link to them. They were converted into plain text using Apache Tika 1.3[3]; no specific problems were noted. The conversion maintains the column formatting; for a two column layout, the second column is appended at the end of the first.

[2] http://europepmc.org/

[3] http://tika.apache.org/1.3

2.3 Mutation identification in text

We selected the EMU tool (Doughty et al., 2011) to perform mutation extraction from text. Compared to other existing mutation annotation tools, EMU is able to identify a broader range of mutations, including DNA insertions and deletions, dbSNP (Sherry et al., 2001) identifiers, and point mutations. In addition, it links the mutations to the proteins and genes that appear in text and, optionally, performs sequence verification against existing databases to increase the precision of the annotations.

We post-processed the output of EMU to be comparable to the information in the COSMIC and InSiGHT databases, i.e., normalizing the mutation mentions to the HGVS format (Den Dunnen et al., 2000). The dbSNP API[4] is queried to recover all available candidates for DNA and protein mutations for a given dbSNP identifier.

Protein missense mutations, mutations in the DNA that result in an amino acid change, identified by EMU are normalized to amino acid (wild type), position, amino acid (mutated). Single letter amino acid abbreviations are used. Thus, a mutation identified by EMU with wild type amino acid *Ala*, position *140* and mutated amino acid *Thr* is converted into *A140T*.

We normalize DNA mutations identified by EMU to the format "c.[position][wild type nucleotide]>[mutated nucleotide]". In the case of insertion and deletions, given position ranges, hyphens are replaced by the underscore character (e.g. *c.597-598delGA* to *c.597_598delGA*).

We identified some mentions in which the position of the DNA or protein mutation as the exon/intron number or codon position. The codon positions were converted to the three candidate nucleotide positions. Exon and intron mentions were removed since no precise position could be derived.

2.4 Gene normalization

EMU identifies gene mentions based on string matching of a dictionary of gene names from the Human Genome Organization (HUGO) and from NCBI's gene database. From this

[4] http://www.ncbi.nlm.nih.gov/projects/SNP/batchquery.html

dictionary, gene names identical to codon names were removed and the P53 gene name, absent in both gene dictionaries, was added. InSiGHT curated genes are easy to map since only 4 genes are included. The COSMIC database contains the gene name and in most cases a HGNC identifier. We normalized the gene mentions identified by EMU to the NCBI Gene database, and then mapped them to the corresponding HGNC identifier.

3 Article processing

Previous work (Jimeno Yepes and Verspoor, 2013) hypothesized the presence of mutations in tables and supplementary materials as one explanation for the low recall of mutations in full text. Here, we access the tables and supplementary materials to investigate the impact of processing those elements.

3.1 Table processing

It has been previously shown that genetic mutation information can appear in tables (Wong et al., 2009). We therefore extracted the tables and table captions associated to the XML articles in our data sets and processed them with EMU.

From the COSMIC database we found 394 articles with tables. After processing the articles with EMU, 197 articles were identified as having mutations in the tables. From the InSiGHT database there are only 8 articles with tables, of which 4 contain mutations. In these articles, no mutations were found in the abstract or full text content at all.

3.2 Supplementary material

Authors often include supplementary material with their publication that contains information supporting the claims in the paper. Supplementary material appears in a variety of file formats as shown in table 1. The InSiGHT set includes only one supplementary material file, while COSMIC has a larger set linked to the papers (505 files associated to 138 articles).

In order to process the supplementary material files with EMU, their content was converted to text using Apache Tika. It handles a large number of file types while preserving the original layout of the document, compared to other possible solutions like Open Office SDK.

Set	COSMIC		InSiGHT
	Files	PMIDs	
MS Word	176	87	1
MS Excel	111	57	0
PDF	82	70	0
MS Powerpoint	34	17	0
CSV	1	1	0
Images	101	36	0
Total	505	138	1

Table 1: Count of supplementary file types

Manual inspection shows that MS Word documents and MS Excel files are converted to text without any problem for our coverage analysis purposes, even though further processing might be required to properly identify different sections in these documents. MS Powerpoint documents seem to contain many images, which are not processed. On the other hand, no mutations seem to be reported in them. A small number of PDF files could not be converted properly since the PDF contained scanned images, but no mutations were reported in these files.

In this work, we did not attempt to process images in the supplementary material. Manual inspection of a random selection of images show that no mutation information can be found in them.

4 Results

We assessed coverage of mutation extraction by evaluating matching of each {PMID, gene, mutation} triple extracted by EMU to the curated mutations present in the databases. From the set of 13 articles for the InSiGHT database, we find 252 mutation triples. For COSMIC, 33,814 mutation triples were identified for the 563 articles.

Table 2 shows the mutations matched (M) and the recall (R) per database and mutation source. Recall measures how many of the database mutations are also identified through the text processing. As we reported in previous work, MEDLINE® abstracts and full text articles provide very limited mutation coverage (only 3% of COSMIC mutations and ˜8% of InSiGHT mutations in the complete collection we evaluated; the numbers for the PMC-OA subset considered here are comparable).

	InSiGHT			COSMIC		
Set	Art	M	R (%)	Art	M	R (%)
Abstracts	13	1	0.40	563	140	0.41
XML Full Text (FT)	9	20	7.94	487	694	2.05
PDF Full Text (PDFFT)	4	7	2.78	76	23	0.07
Tables	8	18	7.14	394	466	1.38
FT+PDFFT+Tables	13	44	17.46	563	929	2.75
Supp. Mat.	1	88	34.92	138	17015	50.59
All	13	115	45.63	563	17896	52.92

Table 2: Variant extraction results. Art=articles in set, M=Mutations matched, R=Recall (%)

The recall shows scant improvement using the few additional PDF full text files that we have added here. Relaxing the gene requirement from the evaluation triple, to allow for the possibility of gene normalization errors, does not provide a significant recall boost.

Our data shows that tables contribute another ˜1% of the mutations. Combining the information from both the full text, including the articles available as PDF, and the tables (FT+PDFFT+Tables) shows that these sources are complementary. Finally, supplementary material has the largest coverage, exceeding by far any other mutation source considered (35% recall for InSiGHT, and 50% for COSMIC). Combining all the sources results in close to 50% recall in the case of the InSiGHT database, and over 52% recall in the case of the COSMIC database. In the supplementary material, most of the mutations are found either in MS Word documents or MS Excel files.

5 Discussion

The results confirm the hypothesis that most of the mutations being curated in the considered databases are not in the main article text but appear in the supplementary material. Article tables also contribute to the extracted mutations but with more limited coverage.

Even with the boost in coverage provided by supplementary material, the coverage is still only around 50%. The main reason is that mutations are represented in tables and supplementary material differently to how they are expressed in unstructured text. These representations do not correspond directly to the patterns that tools such as EMU use to recognize mutations and they are therefore missed. In particular, elements of the mutation, such as a specific base change and the location of that change, can appear in different columns of a table or external data source. Figure 1 of (Wong et al., 2009) exemplifies this sort of representation. In addition, information can be distributed across discontiguous spans of text, such as across several sentences, document sections, or even multiple supplementary files. The current tools do not consider this.

Our work has implications for the curation of mutation databases. Biocuration workflows rely mainly on using textual data (Hirschman et al., 2012; Verspoor et al., 2013) but our results indicate that all the material linked to the articles is required to fully cover all the mutations. Our results could explain the limited coverage of not only mutations but, as well, residue extraction methods that rely on text data (Nagel et al., 2009) and possibly contribute to existing tools like LEAP-FS (Verspoor et al., 2012).

The current study is based on articles available from PMC-OA, which is a reduced set compared to the articles available from PubMed. Applicability of our work is limited to the access of licensed content and diversity of formats available from different journals.

6 Conclusions and Future Work

We have presented an analysis of text mining for genetic variant extraction, extending previous work by considering supplementary material. The achieved recall of approximately 50% of curated mutations dramatically exceeds previously reported results. We plan to build on these results by developing targeted methods for mutation extraction in tables and supplementary material, possibly also includ-

ing mutation extraction from images. The results we have provided here indicate that such methods are critical for achieving effective information extraction of genetic variant data from the literature.

Our work has focused on recall of mutations. We would also like to evaluate more carefully the precision of the mutation extraction. In particular, we hope to refine the current tools to address more specific requirements of database curators (e.g., a given database may be limited to either germ line or somatic mutations, or the database may be restricted to genetic variants related to a specific disease) or to provide further information about the mutations like the reference sequence considered in the reported mutations. We therefore plan to continue this work, linking extracted mutations to more contextual details.

7 Acknowledgements

We thank the InSiGHT database curator, John-Paul Plazzer of the Royal Melbourne Hospital, for sharing the InSiGHT data and helping us to interpret the database fields. We also thank the COSMIC team for helpful details about their database.

Funding: National ICT Australia (NICTA) is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

C.J.O. Baker and R. Witte. 2006. Mutation Mining: A Prospector's Tale. *Journal of Information Systems Frontiers*, 8(1):47–57, February.

S. Bamford, E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dogan, A. Flanagan, J. Teague, PA Futreal, MR Stratton, et al. 2004. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British journal of cancer*, 91(2):355–358.

J.G. Caporaso, W.A. Baumgartner, D.A. Randolph, K.B. Cohen, and L. Hunter. 2007. MutationFinder: A high-performance system for extracting point mutation mentions from text. *Bioinformatics*, 23(14):1862–1865.

J.T. Den Dunnen, S.E. Antonarakis, et al. 2000. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Human mutation*, 15(1):7–12.

E. Doughty, A. Kertesz-Farkas, O. Bodenreider, G. Thompson, A. Adadey, T. Peterson, and M.G. Kann. 2011. Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature. *Bioinformatics*, 27(3):408–415.

L. Hirschman, G.A.P.C. Burns, M. Krallinger, C. Arighi, K.B. Cohen, A. Valencia, C.H. Wu, A. Chatr-Aryamontri, K.G. Dowell, E. Huala, et al. 2012. Text mining for the biocuration workflow. *Database: The Journal of Biological Databases and Curation*, 2012.

A. Jimeno Yepes and K. Verspoor. 2013. A sobering analysis of the recovery of curated genetic variants through text mining, and some lessons. *(under review)*.

M. Krallinger, J.M.G. Izarzugaza, C. Rodriguez-Penagos, and A. Valencia. 2009. Extraction of human kinase mutations from literature, databases and genotyping studies. *BMC bioinformatics*, 10(Suppl 8):S1.

N. Naderi and R. Witte. 2012. Automated extraction and semantic analysis of mutation impacts from the biomedical literature. *BMC Genomics*, 13(Suppl 4):S20–.

K. Nagel, A. Jimeno-Yepes, and D. Rebholz-Schuhmann. 2009. Annotation of protein residues based on a literature analysis: Cross-validation against UniProtKb. *BMC bioinformatics*, 10(Suppl 8):S4.

S. Povey, R. Lovering, E. Bruford, M. Wright, M. Lush, and H. Wain. 2001. The HUGO gene nomenclature committee (HGNC). *Human genetics*, 109(6):678–680.

ST Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, EM Smigielski, and K. Sirotkin. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1):308–311.

K. Verspoor, J.D. Cohn, K.E. Ravikumar, and M.E. Wall. 2012. Text mining improves prediction of protein functional sites. *PloS one*, 7(2):e32171.

K. Verspoor, A. Jimeno Yepes, L. Cavedon, T. McIntosh, A. Herten-Crabb, Z. Thomas, and J.P. Plazzer. 2013. Annotating the biomedical literature for the human variome. *Database: The Journal of Biological Databases and Curation*, bat019.

W. Wong, D. Martinez, and L. Cavedon. 2009. Extraction of named entities from tables in gene mutation literature. *BioNLP 2009*, page 46.