CancerMine

Knowledge base construction for personalised cancer treatment

Jake Lever
Genome Sciences Centre
BC Cancer Agency
Vancouver, Canada
jlever@bcgsc.ca

Martin Jones
Genome Sciences Centre
BC Cancer Agency
Vancouver, Canada
mjones@bcgsc.ca

Steven JM Jones
Genome Sciences Centre
BC Cancer Agency
Vancouver, Canada
sjones@bcgsc.ca

***Abstract*— Knowledge of the relevant genomic aberrations that drive a particular cancer type is necessary to accelerate efficient interpretation of genomic data and enable large-scale endeavours in precision medicine. Currently, this field is limited by the lack of focused and scalable literature curation tools that can reliably capture the required information. Here we present a knowledge-base of genes that have been described in the literature as drivers, oncogenes or tumour suppressors with respect to a specific type of cancer. We have annotated a large body of literature which reports oncogenic aberrations using a custom designed annotation tool. We then applied VERSE, an in-house relation extraction tool, to catalogue driver mutations and illustrate the ability to build a useful resource for clinical interpretation of genomic data for personalised treatment approaches.**

Keywords—relation extraction, oncogenomics, driver mutations

I. Introduction

Improvements in sequencing technology now allow for investigation of individual cancers in a clinically actionable time frame. These technologies reveal a set of mutations in the genome of an individual patient's cancer. These mutations may disable molecular pathways, up-regulate them or dramatically change their function in the quest for increased tumour growth and drug resistance. A bioinformatician examining these sets of mutations must identify the important changes and highlight those relevant for clinical decisions.

Distinguishing between driver mutations, that are important in the tumour development, and passenger mutations, that are coincidental mutations, remains a huge challenge in cancer research. Large scale projects, including The Cancer Genome Atlas (TCGA) [1], have shone a light on the mutational landscapes of a variety of cancer types. However, TCGA by necessity focuses on only the most common or accessible types of cancer and only on primary tumours. Metastatic tumours are a hugely important area, causing 90% of cancer-related mortality [2], and are not as well studied. Existing resources (such as IntOGen [3]) listing known or statistically derived driver genes rely on these large-scale projects but miss variants which may be exquisitely characterised in smaller scale studies or are associated with incidental findings discussed in the literature. Smaller studies on specific cancer types are an important resource for cancer researchers in understanding driver mutations. However, the information from these studies is commonly locked in the text of associated publications and has not been curated into a usable database. Cancer types also play an extremely important contextual role in understanding the function of a particular gene. The NOTCH gene can have oncogenic effects in blood cancers and be tumour suppressive in head & neck cancers [4]. Therefore it is very important to link specific genes with a specific form of cancer.

Previous work has linked gene mutations with diseases based on simple distance metrics [5] and used crowdsourcing to annotate gene mutation relations [6]. Our approach uses syntactic and semantic information to predict relations between cancer types and genes to generate a usable knowledge base based on a smaller set of expert annotated data.

II. Methods

In order to identify sentences that discussed both a human gene and a cancer type, word-lists were generated from popular bioinformatics ontologies. Due to existing named entity recognition tools missing some specific cancer types, a custom word list was created from the UMLS Metathesaurus [7]. All terms and their synonyms of the type Neoplasm (T191) were selected. This list was then manually trimmed to remove very general cancer terms so that only cancer types remained. The NCBI Gene list [8] with all alternative names was used to create a list of human genes with their synonyms and was manually trimmed for several gene names that are common words in biomedical literature (e.g. MICE). The cancer type list contained 12,522 terms and the gene list contained 59,860 terms. Both word lists were filtered by a list of common English words. This word list was built from the stop words from the NLTK toolkit [9], the most frequent 5,000 words based on the Corpus of Contemporary American English [10] and a stop word list associated with the NCBI gene data.

Table 1. Examples of annotated sentences used as training data for (a) driving, (b) oncogenic and (c) tumour suppressive associations with PubMed IDs. Gene names are underlined and cancers are bolded.

(a)	Recent studies reported S100A2 protein is a molecular driver in TGF-β induced cell invasion and migration in **hepatic carcinoma**.(PMID:25591983)
(b)	In summary, our work suggests a new direction for understanding the oncogenic function of TRAF4 in **breast cancer**. (PMID:25738361)
(c)	In present report, the tumor suppressive role of DMTF1 was studied and confirmed in **bladder cancer**. (PMID:25965824)

This research was supported by a Vanier Canada Graduate Scholarship and funding from Compute Canada.

Medical literature was downloaded in XML format from the MEDLINE database of PubMed citations and the Pubmed Central Open Access subset. The raw text was extracted from the files and processed using the Stanford CoreNLP tools [11]. Text was split into sentences and tokenized. A sentence that contained a term from the cancer types word list and a term from the human gene names wordlist was flagged and stored in a MySQL database.

In order to enrich the dataset for sentences likely discussing important cancer aberrations, the sentences were filtered for those containing "driv", "oncogen" or "tumo(u)r suppress". In literature from 2015, 13,765 sentences were extracted and examples are shown in Table 1. Equal numbers of sentences for each filter were then prepared for annotation.

The CancerMine annotation system displays each pair of cancer type term and gene name term that appear in the same sentence. The user can then tag the term pair as having a driver, oncogenic, tumour suppressive or no relation. Driver relations require the sentence to specifically discuss a genomic aberration driving cancer development. Oncogenic relations require the text to state that an aberration is involved in oncogenesis while a tumour suppressive relation requires the text to state that the aberration has a tumour suppressive role. In total, 1203 sentences were annotated by a single annotator providing 504 driver, 521 oncogenic and 215 tumour suppressive relations. Note that 352 sentences had no relations and 412 sentences had more than one relation.

Annotated sentences were then transformed into the input format data appropriate for use with the Vancouver Event and Relation System for Extraction (VERSE) [12]. It was used to predict triggerless events between gene and disease entities. VERSE utilises bag-of-words features based on the entire sentence, dependency paths and individual entities. A logistic regression classifier was used in order to generate a set of probabilities for each annotation type. Only annotations with a probability above a certain threshold were output.

III. Results

A two-fold cross validation approach was used during a parameters search on a 6000 core cluster. A stochastic search strategy was used. The F-score metric with beta=0.1 was used to evaluate the success of each run. This allowed a greater focus on precision to improve the quality of the resulting knowledge base. ~75,000 different runs were executed and the optimal parameters were selected based on an average F-score (beta=0.1) of 0.8845. These parameters provided an average precision of 0.941 and recall of 0.128.

The optimal classifier was then applied to the larger set of unannotated sentences. These sentences were from all accessible literature from 2010 to 2016. Table 2 shows an overview of the data included in the CancerMine knowledge base. The difference in proportion of relations in the training set and the final knowledge base is due to the selection of equal numbers of filtered sentences for each possible relation type for annotation. Importantly all annotations are associated with a PubMed or PubMedCentral ID to allow easy access to the original text of the article or abstract.

Table 2. Overview of data in CancerMine knowledge base

# of analysed sentences	60,464
# of gene terms	155,646
# of cancer terms	79,290
# of driver annotations	1,967
# of oncogenic annotations	6,877
# of tumour suppressive annotations	3,075

IV. Conclusion

In conclusion, we presented a full pipeline for identifying sentences that discuss a gene and cancer type, annotating a large number of sentences and training a high-quality relation classifier on them. This data is an important resource for improved personalised cancer treatment and can be expanded to address other specific questions relevant to genome interpretation, such as clinical outcome.

References

[1] Weinstein, John N., et al. "The cancer genome atlas pan-cancer analysis project." Nature genetics 45.10 (2013): 1113-1120.

[2] Mehlen, Patrick, and Alain Puisieux. "Metastasis: a question of life or death." Nature Reviews Cancer 6.6 (2006): 449-458.

[3] Gonzalez-Perez, Abel, et al. "IntOGen-mutations identifies cancer drivers across tumor types." Nature methods 10.11 (2013): 1081-1082.

[4] Radtke, Freddy, and Kenneth Raj. "The role of Notch in tumorigenesis: oncogene or tumour suppressor?." Nature Reviews Cancer 3.10 (2003): 756-767.

[5] Singhal, Ayush, Michael Simmons, and Zhiyong Lu. "Text mining for precision medicine: automating disease-mutation relationship extraction from biomedical literature." Journal of the American Medical Informatics Association (2016): ocw041.

[6] Burger, John D., Emily Doughty, Ritu Khare, Chih-Hsuan Wei, Rajashree Mishra, John Aberdeen et al. "Hybrid curation of gene–mutation relations combining automated extraction and crowdsourcing." Database 2014 (2014): bau094.

[7] Bodenreider, Olivier. "The unified medical language system (UMLS): integrating biomedical terminology." Nucleic acids research 32.suppl 1 (2004): D267-D270.

[8] Maglott, Donna, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. "Entrez Gene: gene-centered information at NCBI." Nucleic acids research 33.suppl 1 (2005): D54-D58.

[9] Bird, Steven. "NLTK: the natural language toolkit." Proceedings of the COLING/ACL on Interactive presentation sessions. Association for Computational Linguistics (2006).

[10] Davies, Mark. "The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights." International journal of corpus linguistics 14.2 (2009): 159-190.

[11] Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. "The Stanford CoreNLP Natural Language Processing Toolkit." ACL (System Demonstrations). (2014)

[12] Lever, Jake and Steven JM Jones. "VERSE: Event and relation extraction in the BioNLP 2016 Shared Task." Proceedings of the BioNLP Shared Task 2016 Workshop (2016). in press