A Semantic Web-based System for Mining Genetic Mutations in Cancer Clinical Trials

Sambhawa Priya[1,2], Guoqian Jiang, MD, PhD[1], Surendra Dasari, PhD[1], Michael T. Zimmermann, PhD[1], Chen Wang, PhD[1], Jeff Heflin, PhD[2], Christopher G. Chute, MD. Dr. PH[1]
[1]Mayo Clinic, Rochester, MN; [2]Lehigh University, Bethlehem, PA

Abstract

Textual eligibility criteria in clinical trial protocols contain important information about potential clinically relevant pharmacogenomic events. Manual curation for harvesting this evidence is intractable as it is error prone and time consuming. In this paper, we develop and evaluate a Semantic Web-based system that captures and manages mutation evidences and related contextual information from cancer clinical trials. The system has 2 main components: an NLP-based annotator and a Semantic Web ontology-based annotation manager. We evaluated the performance of the annotator in terms of precision and recall. We demonstrated the usefulness of the system by conducting case studies in retrieving relevant clinical trials using a collection of mutations identified from TCGA Leukemia patients and Atlas of Genetics and Cytogenetics in Oncology and Haematology. In conclusion, our system using Semantic Web technologies provides an effective framework for extraction, annotation, standardization and management of genetic mutations in cancer clinical trials.

1 Introduction

The eligibility criteria in a clinical trial protocol may contain information about genetic mutations that determine the appropriateness of a cancer patient being administered a particular drug/therapy. This genetic mutation information and related context has been considered as an important source for compiling pharmacogenomics (PGx) evidences. Unfortunately, this information occurs in free-text format eligibility criteria, which is difficult to parse using standard text mining techniques. In previous studies, different approaches have been explored for capturing and structuring information in clinical trials, however most systems are not focused on capturing mutations mentioned in eligibility criteria. For example, Li et. al.[1] used a dictionary based approach to detect genes, drugs and diseases based on the condition, intervention and study description fields. Wu et. al.[2] used machine learning techniques to identify genes and their categorical status (mutated or not) from eligibility criteria, however they do not identify larger structural variations or specific variants. eTACTS[3] mines frequently occurring tags from the free-text eligibility criteria to provide efficient filtering of trials and facilitate search for trials. Since this system only preserves the frequently occurring tags for high-level concepts, it is likely to miss less recurrent mutation mentions. Other works by Kanagasabai et. al.[4], Naderi et. al[5] and Laurila et. al[6] focus on extracting mutation mentions from biomedical literature, but not clinical trials.

Thus, most of the systems lack the capability to provide sufficient mutation annotations for clinical trials, which makes it hard to search for relevant trials based on patients' mutation information. We consider that clinical research community would need such a system that can capture and store these annotations in a structured format so that it can be queried to retrieve relevant information from clinical trials. For this reason, we leveraged Semantic Web technology, that provides a scalable framework for standards-based data representation, integration and sharing.We store the extracted annotations in a structured representation using Resource Description Framework (RDF)*, a Semantic Web standard.

Such a system can be used to support personalized genomics medicine or individualizing medicine, a goal of which is to match the right patient to the right medicine based on the patient's genomic information. It is a common practice to compile PGx evidences from heterogeneous sources, such as PubMed, PharmGKB[16], COSMIC[17], etc. Even though clinical trials may not have concluded, they are designed and carried out based on knowledge from preliminary studies[1]. Hence, PGx evidences from clinical trials are reasonable candidates for potential cancer treatment when a standard guideline is not clear. However, they should be tagged specifically if there are no published results confirming the study.

*http://www.w3.org/RDF/

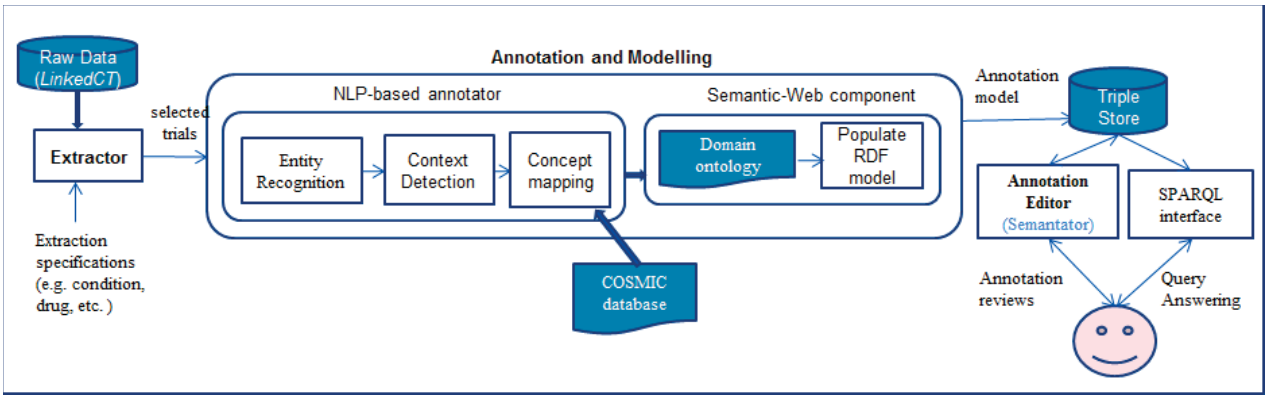

Figure1: System Architecture

2 Materials and Methods

2.1 System Architecture Overview

The system is comprised of 4 main modules: an extractor, an NLP-based annotator, a Semantic Web ontology-based annotation manager and a user interaction module with editing tools and a SPARQL[14] query interface (Figure 1). The extractor selects the clinical trials based on some extration specifications (e.g. condition.). The NLP component annotates the eligibility criteria of the selected trials and maps it to a dictionary database. These annotations are serialized into structured format by the Semantic Web component. The user can review and edit the generated annotations using the annotation editing tool and explore the generated model using the SPARQL query interface.

2.2 Prototype Implementation

The Extractor - We use LinkedCT[7], an RDF version of ClinicalTrials.gov, as our raw dataset. In this implementation, the extractor selects some clinical trials associated with the condition of leukemia. We focus on leukemia trials because such trials are enriched with mutation evidences.

The NLP-based annotator - There are three modules of the NLP subcomponent which achieve entity recognition, concept mapping, and dependency detection. For the entity recognition task, we created a rule set for capturing different types of mutation evidences. Based on manual curation of 50 leukemia trials, we created rules (regular expressions) to capture structural variations like translocations, deletions and inversions. We borrowed the regular expressions from the MutationFinder[9] tool for capturing the point-mutations. We also include phrase-based rules to capture frequently occurring mutation phrases like 'MLL gene rearrangements', 'hypodipoidy', etc.

After identifying the mutation entities, we mapped them to the Catalog of Somatic Mutations in Cancer (COSMIC)[17] which serves as our dictionary database. We could map the extracted point-mutations and structural variations to those in COSMIC, however mapping broad low-resolution categories like 'MLL rearrangement', 'hypodipoidy', was more difficult. In these cases, we used bedtools[12] to convert positions with variants in COSMIC (coordinates from GRCh37) to their corresponding cytoband, the resolution most commonly appearing in eligibility criteria. This process also enriched our annotation model with additional evidences from the COSMIC database, like PMID reference and associated gene(s) for the matching mutations.

Sentences in eligibility criteria often contain negation or other context which may modify meaning (e.g. "The absence of good risk molecular features: t(8;21) ..." or "any cytogenetic abnormality except inv16..."). It is important to capture this information and associate it with the corresponding mutation evidence for accurate information retrieval. The free-form narrative format of the eligibility criteria renders standard parsing tools (e.g. the Stanford parser[15]) ineffective. We manually reviewed 50 clinical trials with negative status for mutations in criteria and identified a list of common negative signals for mutation evidences (e.g. 'absence of', 'except', etc.). If any negative word from the identified list appeared within a sentence or at the beginning of an enumerated list of sentences containing mutation evidence, we assigned a negative context to that mutation evidence.

The Semantic Web ontology-based annotation management – In order to store the extracted annotations in the RDF format, we need an ontology that represents the knowledge of the domain. We explored some existing domain ontologies that include the concepts and properties relevant to mutation annotations. One of the ontologies we explored was the mSTRAP ontology[4], specifically designed for structuring annotations for point-mutations in

biomedical literature. However, since our focus was mainly on the eligibility criteria for clinical trials related to leukemia where structural abnormalities are more frequently occurring type of mutation evidence than the point-mutations, this ontology wasn't sufficient to cover our requirements. We also looked at the OMM impact ontology[5] that covers insertions, deletion and point-mutations but includes many other features that are irrelevant to our modelling requirements. Thus, by combining relevant features from the above two ontologies with some new classes and properties of our own, we designed our domain ontology for capturing mutation annotations in eligbility criteria of clinical trials. The domain ontology file can be accessed from the Github repository here: https://github.com/sambhawa/Mutations_In_LinkedCT_Ontology/blob/master/Mutations_In_LinkedCT_tbox.owl. The captured mutation annotations are used for instantiating this domain ontology to create the annotation model.

User-Interaction module - Apart from outputting the annotation model for eligibility criteria text, our system also creates output compatible with Semantator[8] , an OWL annotation editing tool. This output contains highlighting information about the annotated entities and the classes they are instances of. A domain expert can export the Semantator-compatible output files to Semantator to review and edit the annotations generated by our system.

We can store the annotation model in a triple store and query it using SPARQL[14]. We can use SPARQL to issue queries over distributed datasets, as shown in Figure 2.

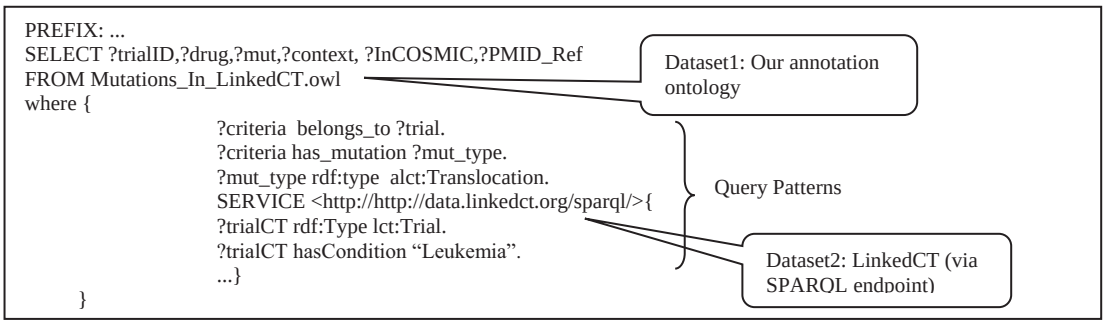

Figure 2:A SPARQL query template to retrieve information for leukemia trials that contain mutations of the type translocation. This query is executed over the datasets including our annotation ontology and LinkedCT.

2.3 System Evaluation Design

We use LinkedCT, the RDF version of ClinicalTrials.gov, as our input dataset. In this evaluation, the data collection consisted of 172,363 trials available as of August 2014. In total, we retrieved 3605 leukemia associated trials. Our annotation system detected a total of 2044 mutation mentions (247 unique mutations) in 472 trials which contained mutation evidences out of the complete set of retrieved trials. We evaluated the performance of the NLP component in terms of precision and recall. Standard ground-truth for eligibility criteria annotations is missing. Hence, with help from 3 co-authors who have extensive expertise in molecular biology and bioinformatics, we created ground truth for a set of 125 leukemia trials that contained some mutation evidences. We manually annotated mutation mentions and their status as inclusion or exclusion criteria, incorporating any negative context for mutations in the inclusion criteria. These trials serve as our test-set for evaluation.

We demonstrate the usefulness of our system by conducting case studies in retrieving relevant clinical trials using a collection of mutations identified from TCGA Leukemia patients and Atlas of Genetics and Cytogenetics in Oncology and Haematology (AGCOH)[11]. Using cBioPortal[13], we obtained a list of frequently mutated genes associated with Acute Myeloid Leukemia(AML). We randomly selected a few TCGA patient cases for some of these genes and, queried our annotation model to retrieve clinical trials associated with mutations matching with those in the selected patients. In order to demonstrate the effectiveness of our system in retrieving clinical trials with structural abnormalities, we performed an experiment to search for mutations from AGCOH in our annotation model. We extracted all the structural variations for leukemia from AGCOH and looked for partial or complete match in our annotation model and returned the corresponding trials.

3 Evaluation Results

3.1 Performance of the NLP subcomponent

For this evaluation, we extracted annotations using our system for the trials in the test-set described above. Table 1 shows precision, recall and f-score for different categories of mutation mentions (structural variations such as

'del(12q)', 'inv(16)', 't(16;16)(p13;q22)'; point-mutation such as 'Q252H', 'T315I'; and other mutation mentions such as 'MLL gene rearrangement', 'hypodiploidy').

Table 1: Precision, Recall and F-Score for different categories of mutation mentions appearing in leukemia trials.

Category	Precision	Recall	F-Score
Structural Variants (SV)	90.2%	83%	86.5%
Point-Mutation (PM)	100%	100%	100%
Others	98.4%	85.33%	91.4%

3.2 Utility of the system

We identified 2 TCGA patients, one with FLT3 D835Y mutation, the most frequently seen point-mutation in FLT3 and another one with KIT D816V mutation, the most frequently occurring point-mutation in KIT. For both the patients, our system found the clinical trials which specifically mentioned these mutations (Table 2). This use case demonstrates the capability of our system in retrieving trials given a patient's mutation information. ClinicalTrials.gov website only retrieves search results for mutations appearing in the title or the keyword list for the trial, but not when the mutation occurs only in the eligibility criteria.

Out of 455 chromosomal abnormalities found in AGCOH for leukemia, we found 98 matching unique variations in our annotation model (which has a total of 217 unique structural variations). A total of 273 unique matching trials were returned corresponding to the mutations in AGCOH. Most frequent matches for mutations from AGCOH were 11q23 abnormalities, t(9;22) and inv(16). Searching for occurrence of structural variations in eligibility criteria is not supported by basic or advanced search facility on ClinicalTrials.gov website. Hence, our system is useful for searching for trials with such broader categories of mutations.

Table 2: Information retrieved from our annotation model for clinical trials matching to the mutations in sample TCGA leukemia case-studies. The last column indicates if the search on ClinicalTrials.gov (CT.gov) website returned the matching trial or not.

CaseID	Matching-TrialID	Matching Mutation	Drug	Context	Trial-Status	COSMIC-ID	Search-CT.gov
TCGA-AB-2811	nct00045942	D835Y	PKC412	Inclusion	Completed	783	No
TCGA-AB-2945	nct00171912	D816V	Imatinib Mesylate	Exclusion	Completed	1314	Yes
	nct00233454	D816V	PKC412	Inclusion	Active	1314	No

4 Discussion

Our rule-based text-mining approach was effective in identifying mutation evidences and their associated contextual information from eligibility criteria of clinical trials associated with leukemia. Standard mutation extraction tools such as MutationMiner and tmVAR reviewed by Yepes et. al.[9] are not effective in capturing most of the mutation evidences occurring in the eligibility criteria because these tools do not handle non-standard formats of structural variations for translocations (e.g. t(1;2)(p12;q13)), deletions (e.g. del (5q), -5q), or complex mutation mentions (e.g. "karyotype t4", +4) that frequently appear in the eligibility criteria for leukemia trials. Hence, we decided to design our own set of rules. We currently do not handle double negatives (e.g. "for patients whose AML does not have good risk cytogenetic features, i.e. t (8;21) and t(15;17) without c-kit mutations) or conditional negatives (e.g. No MLL rearrangements if 12 to 24 months old). We plan to make the NLP component more robust by including richer rules and capturing annotations for other features in the eligibility criteria, including trials for other cancer types.

Our Semantic Web-based annotation model was useful in searching and retrieving relevant information for clinical trials associated with a diversity of variants in leukemia. The advantage of ontology-based annotation management is that it not only helps structure the extracted information, but also facilitates data integration. Most of the existing annotations for eligibility criteria, generally stored in relational databases or as indices[1,3], are data silos that cannot be integrated easily. Our ontology can be easily integrated with LinkedCT and this combined dataset can serve as an enriched version for the later. Our RDF model can also be integrated with other biomedical datasets for capturing additional evidences for interactions between biomedical entities such as drugs, diseases and genetic-mutations. We

can align our ontology with other mutation-specific ontologies (e.g. Sequence Ontology[10] or OMM impact ontology[5]). By converting the annotations to an instance of a domain ontology, an investigator can explore the annotation model using semantic queries in SPARQL or leverage other Semantic Web tools, such as reasoners that can check the consistency of the populated ontology.

The mutation evidences captured by our system in the clinical trials may be disease-causing or disease-associated. Hence, the interactions between drugs and mutations that we extracted from the clinical trials may not specifically represent drug-target associations but rather reflect more loosely defined relationships between treatments and genetic-mutations. In the future, we plan to associate the disease-drug-mutation relationships extracted from clinical trials to relevant evidences found in sources like PubMed and PharmGKB[16] in order to authenticate the extracted relationships.

5 Conclusion

We successfully developed a Semantic Web-based system that provides an effective framework for extraction, annotation, standardization and management of genetic mutations in cancer clinical trials. In the future, we plan to generalize our approach for clinical trials associated with other cancer types. As we annotate different types of cancer clinical trials, we will leverage scalable infrastructures for managing large-scale RDF datasets. We also plan to design a user-friendly interface that would allow a non-SPARQL expert to input different criteria (e.g. genomic feature, primary disease context, age, etc.) that will be translated into appropriate SPARQL queries.

6 Acknowledgements

The study is supported in part by a NCI U01 Project – caCDE-QA (1U01CA180940-01A1). The authors would like to thank Dr. Ravikumar Komandur Elayavilli and Dr. Hongfang Liu for their helpful inputs and support.

7 References

1. Li J, Lu Z. Systematic identification of pharmacogenomics information from clinical trials. J. Biomed. Inform.. 2012; 45:870–878.
2. Wu Y, Levy MA, Micheel CM, Yeh P, Tang B, Cantrell MJ, Cooreman SM, Xu H. Identifying the status of genetic lesion in cancer clinical trial documents using machine learning. BMC Genomics. 2012; 13 S8:S21.
3. Miotto R, Jiang S, Weng C. eTACTS: A method for dynamically filtering clinical trial search results. J Biomed Inform. 2013; 46(6):1060-7.
4. Kanagasabai R, Choo KH, Ranganathan S, Baker CJ. A workflow for mutation extraction and structure annotation. J Bioinform Comput Biol. 2007;5(6):1319-37.
5. Naderi N and Witte R. Automated extraction and semantic analysis of mutation impacts from the biomedical literature. BMC Genomics. 2012;13 Suppl 4:S10.
6. Laurila JB, Naderi N, Witte R, Riazanov A, Kouznetsov A, Baker CJ. Algorithms and semantic infrastructure for mutation impact extraction and grounding, BMC Genomics. BMC Genomics. 2010;11 Suppl 4:S24.
7. Hassanzadeh O, Kementsietsidis A, Lim L, Miller RJ, Wang M. LinkedCT: A Linked Data Space for Clinical Trials. In: Proc. WWW 2009 Workshop on Linked Data on the Web, LDOW. 2009
8. Song D, Chute CG.,Tao C. Semantator: Annotating Clinical Narratives with Semantic Web Ontologies. 2012 AMIA Summit on Clinical Research Informatics (AMIA CRI), 2012; 20-29.
9. Yepes AJ and Verspoor K. Mutation extraction tools can be combined for robust recognition of genetic variants in the literature. F1000Research. 2014; 3:18
10. Eilbeck K, Lewis SE. Sequence Ontology Annotation Guide. Comp Funct Genomics. 2004; 5(8): 642–647.
11. Huret JL, Dessen P., Bernheim A. Atlas of Genetics and Cytogenetics in Oncology and Haematology, updated. Nucleic Acids Res. 2001; 29(1): 303–304.
12. http://bedtools.readthedocs.org/en/latest/
13. http://www.cbioportal.org/public-portal/index.do
14. Prud'hommeaux E., Seaborne A. SPARQL query language for RDF. http://www.w3.org/TR/rdf-sparql-query/
15. Klein D, Manning CD. Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics. 2003; pp. 423-430.
16. Klein TE, Chang JT, Cho MK, Easton KL, Fergerson R, Hewett M et. al. Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. Pharmacogenomics J. 2001;1(3):167-70.
17. Forbes SA, Bhamra G, Bamford S,Dawson E,Kok C et. al. The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet.2008;10.11.