MutDR – A Resource for Protein Mutation-Disease Relations Assembled from Biomedical Literature

Ravikumar Komandur Elayavilli, Majid Rastegar-Mojarad, Hongfang Liu
Department of Health Sciences Research
Mayo Clinic
Rochester, MN, 55901

***Abstract*— Text mining approaches can accelerate the process of assembling knowledge from literature. In this abstract, we present our effort in assembling a resource for protein mutation-disease relations assembled from literature.**

Keywords—literature mining; protein mutation-disease

I. INTRODUCTION

A large amount of information about the role of gene variants and mutations in diseases is available in curated databases such as OMIM [1], ClinVar [2], and UniprotKB [3]. However, much of this information remains 'locked' in the unstructured form in the scientific publications. Since manual curation involves significant human effort and time there is always a lag in the information between the curated databases and the literature. The recent findings published in the literature takes significant time to find its way into the curated knowledgebase. Text mining approaches can accelerate the process of assembling this knowledge from the published literature. However, developing a text-mining system with semantic understanding capability in the biomedical domain is very challenging. In an earlier work, we described MutD [4], a literature mining system that extracts relationship between protein point mutation and diseases from bio-medical abstracts. In this abstract, we present access to a PubMed scale resource through a web interface that allows users to retrieve protein point mutation-disease relations extracted through biomedical literature mining.

II. BACKGROUND

MutD is a literature mining system that uses an ensemble of state of the art named entity extraction and normalization tools and graph based dependency parse representation to extract relations between protein point mutations and diseases mentioned in biomedical abstracts. It also extends the scope of literature mining to across multiple sentences through discourse processing and heuristics. MutD achieved a precision of 71% and recall of 58% (F-Measure: 64%) when compared against the annotations of UniProtKB.

III. METHODS

In this work, we describe the extension of MutD to create a PubMed scale resource of literature-mined Protein Point mutation and disease (PMD) relations, MutDR. Figure 1 outlines the overall workflow in developing MutDR.

Using MutD, we performed a large-scale mining of PMD relations on the complete PubMed data set (till May 2016). The extracted PMD relations were indexed using Elastic Search and a simple web based search interface was developed to enable users to retrieve the literature-mined PMD relations. The user interface has three major functionalities: 1) Query by a gene/protein, a disease or both 2) Retrieve results ranked according to the relevance of the query and further by date. 3) Link the normalized entities genes and diseases to the external knowledge resources namely UniProtKB and Comparative Toxicogenomics database (CTD) [5] respectively.

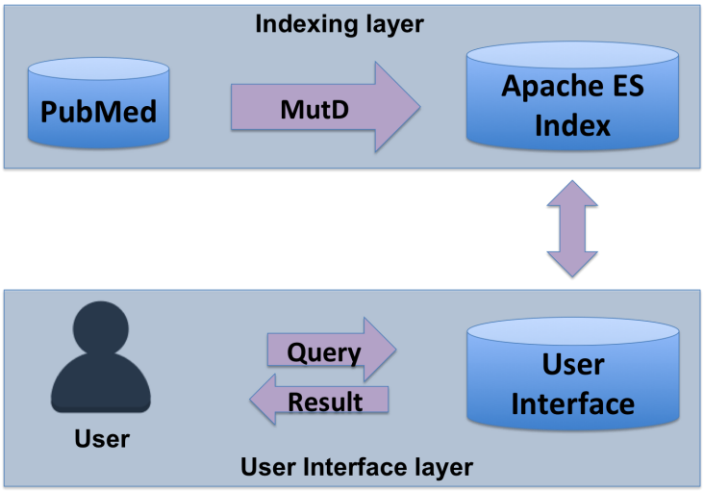

Fig.1 – Overall architecture workflow

IV. RESULTS

MutD extracted 27, 213 protein mutation disease relations from nearly 81, 048 PubMed abstracts (out of the total 21 million abstracts). Figure 2 shows some of the user interface features of MutD resource. The PMD relations extracted by MutD are indexed using Elastic Search [6].

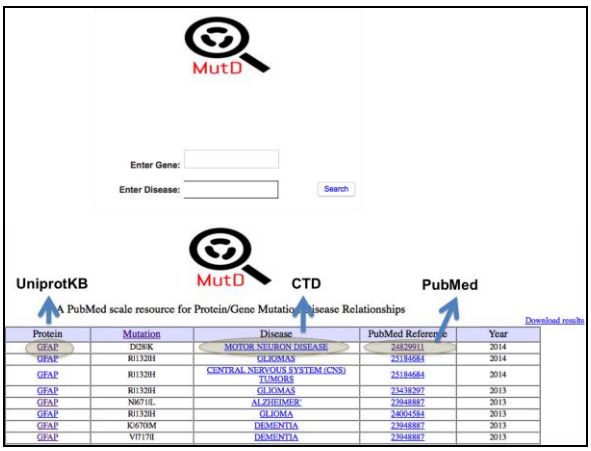

Fig. 2 – Overall architecture workflow

REFERENCE

[1] J. Amberger, C. A. Bocchini, A. F. Scott, A. HamoshA, " McKusick's Online Mendelian Inheritance in Man (OMIM)," Nucleic Acids Res 37: D793–6, 2009

[2] M. J. Landrum, J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church, et al., "ClinVar: public archive of relationships among sequence variation and human phenotype." Nucleic Acids Res 42: D980–5, 2014

[3] The UniProt Consortium, "Activities at the Universal Protein Resource (UniProt)," Nucleic Acids Res 42:D191–D198, 2014

[4] K. E. Ravikumar, K. B. Wagholikar, D Li, J. P. Kocher, and H. Liu, " Text mining facilitates database curation-extraction of mutation-disease associations from Bio-medical literature," BMC Bioinformatics, 16(1), 185, 2015.

[5] A. P. Davis, C. J. Grondin, K. Lennon-Hopkins, C. Saraceni-Richards, D. Sciaky, B. L. King, T. C. Wiegers, and C. J. Mattingly, "The Comparative Toxicogenomics Database's 10th year anniversary: update 2015," Nucleic Acids Res 43: D914-D920, 2015

[6] R. Kuc and M. Rogozinski. Elasticsearch Server. Packt Publishing Ltd, 2013.