BioKB - Text Mining and Semantic Technologies for Biomedical Content Discovery

Maria Biryukov, Valentin Grouès, Venkata Satagopam, and Reinhard Schneider

Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg.
maria.biryukov@uni.lu, valentin.groues@uni.lu, venkata.satagopam@uni.lu, reinhard.schneider@uni.lu,
WWW home page: https://wwwen.uni.lu/lcsb

Abstract. The ever-increasing number of publicly available biomedical articles calls for automatic information extraction from digitized publications. We have implemented a pipeline which, by exploiting text mining and semantic technologies, helps researchers easily access semantic content of thousands of abstracts and full text articles from PubMed and Elsevier. The text mining component analyzes the articles content and extracts relations between a wide variety of concepts, extending the scope from proteins, chemicals and pathologies to biological processes and molecular functions. Moreover, the relations are extracted along with the context which specifies localization of the detected events, preconditions, temporal and logic order, mutual dependency and/or exclusion. Extracted knowledge is stored in a knowledge base publicly available for both, human and machine access, via web interface and SPARQL endpoint. To address the data accessibility, reusability and interoperability, all the extracted relations are standardized using unique resource identifiers (URIs) and a custom ontology based on Genia ontology.

1 Introduction

Information extraction from biomedical literature is becoming a common practice due to the huge amounts of available textual data, and technological maturity which allows to gain insight into scientific content. Text analysis evolved from spotting relevant concepts in the text [18, 19] to co-occurrence statistics [20–22] and, finally, extraction of complex events which seek to reveal cause-effect relation between various entities involved in the biomedical processes [23–25]. Some approaches use textual data as the only source for the analysis [26, 27] while some other combine it with experimental data available from dedicated databases [28, 29]. Although there have been efforts to harmonize the output of several named entity recognition systems [30, 31], the wealth of the results obtained from heterogeneous sources has relatively limited outreach due to lack of a common language: each system typically comes up with its own nomenclature if any [32]. It is where semantic technologies come into play to become an integral part of the information extraction process. To increase data reusability

and interoperability several solutions have been proposed. PubAnnotation [33], micro [34] - and nanopublications [35] are important examples of how to represent extracted knowledge in a standardized format as to be accessible and shared between machines and human.

Knowledge discovery systems and platforms vary in scope. Many of them are focused on specific sub-domains. For example, EVEX [23] targets directed interactions between proteins; DisGeNET [36] explores genetic mechanisms of diseases, while LimTox [37] searches for toxicity associations of compounds, drugs and genes with the special interest in liver. Other systems adopt less centered strategies, trying to cover more aspects involved in biomedical processes. One such system is PolySearch [38], which searches associations between more than twenty entity types, exploiting data from medical literature, Wikipedia articles and 14 databases, among which are UniProt, DrugBank and HMDB. While leading in scope, Polysearch does not specify association types or directionality, leaving these important pieces of knowledge to be completed by the user. The BioKB platform[1] we introduce here aims to discover cause-effect relations between multiple entity types and deliver standardized representation of knowledge.

The paper is organized as follows. Section 2 gives an overview of the BioKB platform. In Section 3 we describe the text mining component. Section 4 focuses on semantic technologies employed by BioKB. Description of the web interface follows in Section 5. Section 6 offers a discussion, while conclusions are presented in Section 7.

2 System Overview

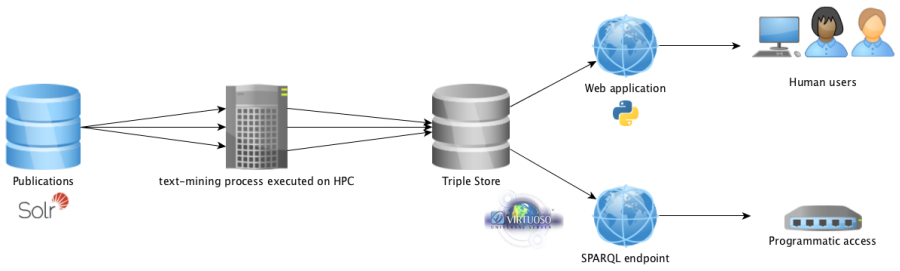

Fig. 1. BioKB platform. The publications, initially stored in Solr, are processed by the text-mining module. The results are then stored in Virtuoso for dissemination through a web interface and a SPARQL endpoint.

[1] Not to be confused with two other independent systems: `http://www.cs.cmu.edu/~biokb/` and `http://www.bioinf.mvm.ed.ac.uk/twiki/bin/view/TWiki/BioKbPlugin`.

Systems architecture is illustrated in Figure 1. Publications retrieved from PubMed and PubMed Central are indexed by a Solr instance; each publication is processed by the text mining component; results are converted to RDF (N-Quads) and stored in a triple store. To allow both human and machine access to the knowledge base, SPARQL endpoint provides machine readable access while a web application allows users to browse the content of the knowledge base. The web application is developed in Python 3 using the Flask framework and the SPARQLWrapper library to query the triple store. We use the vis.js [17] library to render the bio-medical events as a graph.

3 Text Mining Component

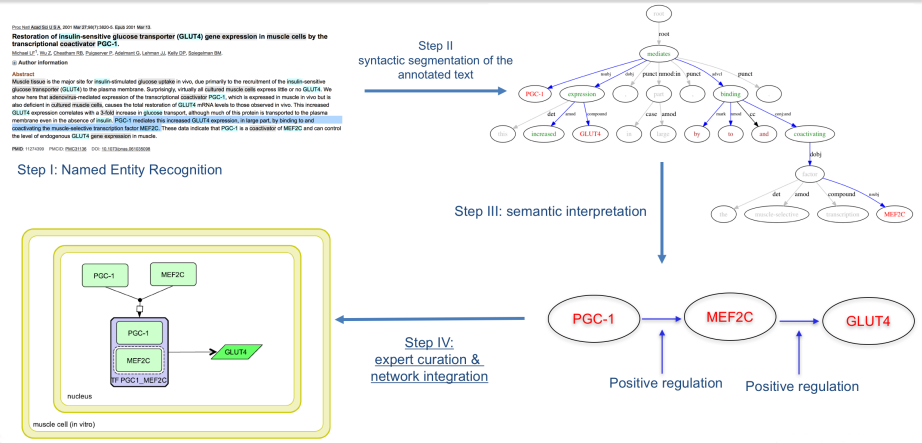

Fig. 2. Main stages of the text mining processing. The sentence detailed in Steps II and III is *PGC-1 mediates this increased GLUT4 expression, in large part, by binding to and coactivating the muscle-selective transcription factor MEF2C.* Triggers are marked in green, entities in red.

The main steps executed by the text mining component of BioKB are: i) named entity recognition; ii) syntactic parsing; iii) semantic interpretation (see Figure 2). They are briefly described in the following subsections.

3.1 Named Entity Recognition

During the Named Entity Recognition (NER) stage, biomedical concepts are identified in the text. Our choice of a NER engine was driven by two major requirements: a) capability to identify multiple concept types (bio-entities) to avoid using and synchronizing multiple NER tools within one pipeline; b) ability of the engine to map entity name to its unique identifier in a dedicated database.

The latter is known as *normalization* process and is indispensable in order to ensure database and semantic graph coherence. One of the systems which meets our criteria is *Reflect* [1]. Reflect recognizes proteins, chemicals, diseases, tissues, cell types, GO processes. In Step I of the Figure 2 entities identified by Reflect are marked in turquoise and grey.

3.2 Trigger Generation

Availability of the trigger dictionary is another prerequisite for semantic analysis. Triggers are words or expressions used to describe a biological process. For example, *mediates, increased, expression, binding* and *coactivating* are examples of triggers in the phrase in Figure 2. Our trigger dictionary is derived from Genia annotated corpus [2] which is a collection of PubMed abstracts with the detected biomolecular events of various types: gene expression, (positive/negative) regulation, binding, cell process etc. Genia corpus is used also to learn so-called 'knowledge cues' which express negative statements and author attitude toward facts being described, such as hypothesis, uncertainty, etc. Each entry in the trigger/knowledge cue dictionary is assigned a relative weight calculated based on positive and negative examples learned from the corpus. During the text analysis, triggers and knowledge cues are detected as dictionary match; those which satisfy a pre-set threshold are retained. Since Genia corpus is limited to 2000 abstracts, we try to increase potential coverage of the text mining component and expand triggers and knowledge cues with synonyms using WordNet [4], which we access via NLTK [3].

3.3 Syntactic Analysis

With the entities and triggers in place, we can proceed toward syntactic analysis. In order to maximize the probability of identification of "subject-predicate-object" triples (e.g., "*RFLAT-1* **activates** *RATES*"), only the sentences with at least two entities and one trigger are processed. For syntactic analysis we use Stanford parser [5] with Stanford dependencies [6]. Step II in Figure 2 shows dependency graph into which the surface structure of the sentence has been transformed by the parser. A proven benefit of using dependency parsing in information extraction task is the ability to map syntactic dependencies onto semantic roles [7, 24, 8].

3.4 Semantic Interpretation

In order to ensure transfer between syntax to semantics, we opt for the rule-based approach. It consists of assigning semantic roles to entities which are syntactic arguments of a trigger. As a result, relations are typed (mostly, the type is inherited from the type of their trigger) and, whenever applicable, directed. For example, direction of a regulatory event is from semantic subject (cause) toward semantic object (theme). On the contrary, relations of type *binding* and *correlation* are naturally not directed. We collect syntactic arguments of the triggers

via the depth-first search (DFS) of the sentence graph. The rules are applied on the ensemble of trigger and its dependencies. For example, syntactic subject of *mediates*, *PGC-1*, is the semantic subject of the regulatory event whose predicate is *mediates*. Sometimes nodes are merged in favor of a more straightforward semantic interpretation. Thus, *increased* and *expression* are jointly interpreted as *Positive regulation*, loosing their individual correspondence to *Positive regulation* and *Gene expression* relation types.

Biomedical processes are subject to rich variety of conditions under which they could take place. We attempt to account for these by processing information conveyed by certain lexical and syntactic elements. For example, the main event in Figure 2, *PCG-1 mediates positive regulation of GLUT4*, is communicated along with the description of its mechanism introduced by the adverbial clause headed by trigger verb *binding*. By taking this bit of information into account we can logically order the events described in the sentence: (1) PCG-1 binds and coactivates MEF2C; (2) GLUT4's expression is increased (Step III of the Figure 2).

4 Semantic Web Technologies

The choice of using semantic web technologies for this project was dictated by two main reasons. First, using an ontology to represent the hierarchy of relationships offers different level of query granularity. For instance, one can ask if two entities are connected by a property *regulates* and be able to retrieve also results for the property *increases* because the two properties are linked by a sub-property relation. Additionally, the ontology and thus the hierarchy of properties can be updated without having to re-process the publications. Besides this reasoning capability, using semantic web technologies offers full machine readable access to the complete knowledge base. Not only can the knowledge base then be used by third parties directly but it becomes possible to combine BioKB data with external sources using federated queries.

4.1 BioKB Ontology

We created a simple ontology (Figures 3, 4), to represent the hierarchy of classes and properties that are used to categorize entities and relations identified by the text-mining component. This ontology is heavily inspired by the GENIA ontology. Our decision to allow inferences on sub-relationships resulted in the need to create a custom ontology. Indeed, in the GENIA ontology, relationships are represented by classes rather than properties. In the proposed ontology, a relationship between two bio-medical entities can be directly translated to a single triple, $s\ p\ o$ where s and o are the entities and p is a sub-property of *biokb:bioRelation*, the top level property in the BioKB model. We then use the named graphs feature of Virtuoso to add metadata about this relationship. This includes information such as creation date, provenance and confidence score.

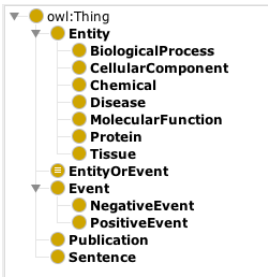

Fig. 3. Classes hierarchy of the BioKB ontology

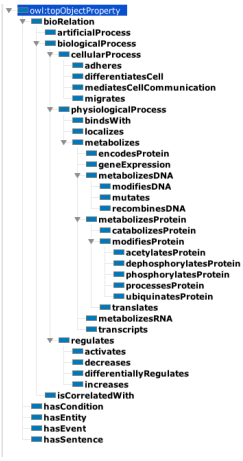

Fig. 4. Properties hierarchy of the BioKB ontology

4.2 Triple Store

In the current deployment of the platform, a single instance of the open source edition of Virtuoso 7 hosts the knowledge base and provides the SPARQL endpoint. The server hosting the Virtuoso instance has the following characteristics: 128GB Ram, 8 cores, Hard Drive 500GB 10000 RPM. At the date of this publication, the size of the database is 22GB for 215 million triples. On top of the content generated by the text-mining module, the different ontologies mentioned in Section 4.1 have also been loaded into the triple store. The actual number of triples constituting the BioKB specific content is about 156 million triples. Those triples are the result of the processing of more than 800 000 publications. About 10 million events were extracted from approximately 6.5 million sentences.

5 BioKB Web Interface

Besides the SPARQL endpoint, we created a web interface to access the BioKB content. This web application is publicly and freely available at `https://biokb.lcsb.uni.lu`. The home page displays a unique search field providing autocomplete functionality for all supported bio-medicial entities. Once the user clicks on an entity, the entity page will be displayed. This page shows a textual description of the entity, the list of most common co-occurrences for this entity (as a tag cloud) and two tables with the list of relationships involving this entity as extracted by the text-mining module. Those incoming and outgoing relationships are also represented visually as an interactive graph (Figure 5). On this graph, the central node is the entity corresponding to the current page and all other nodes and edges represent the most common relationships involving this entity. For each edge, on mouse over, the label and the number of occurrences

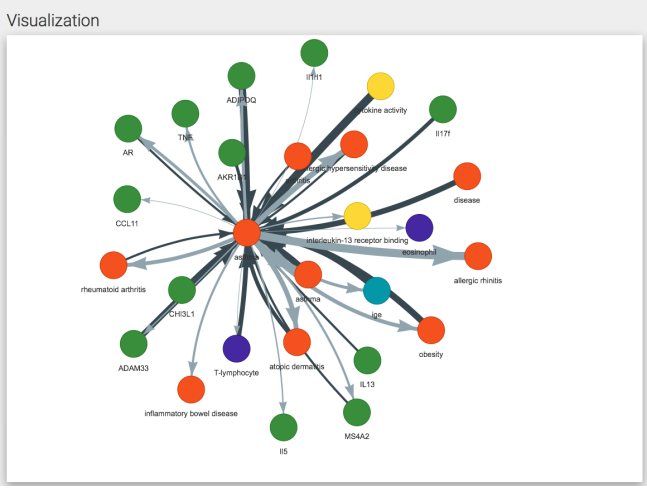

Fig. 5. Graph visualization of *Asthma*. Central node is Asthma. Other nodes with corresponding edges represent relationships identified by the text-mining component. Each color correspond to a different entity type (Disease, Genes, etc).

of this relationship are displayed. Each node is clickable and leads to the corresponding entity page. Each edge is also clickable and results in the display of the relationship details page (Figure 6). This page displays the list of publications where this relationship was found and the specific sentences. On the entity page, a download button proposes an export of the result of the SPARQL *DESCRIBE* command in RDF/XML and in CSV.

6 Discussion

6.1 Use Cases

The primary goal of our information extraction system and knowledge base is to help researchers focusing on various types of biomedical data analysis. We illustrate its functionality with two use cases related to disease network construction and enrichment.

Chronic obstructive pulmonary disease (COPD): the network verification challenge. Gathering disease-related factors into a large-scale network became a common practice. Such networks provide a comprehensive model which helps to elucidate mechanisms involved in pathological processes. For this network verification challenge, we used our system ability to provide typed, directed (if applicable) relations between various concepts. We have scanned the literature and extracted candidate relations which have been verified by a human expert and made part of the collaborative community curated network yielded by the

Fig. 6. Specific bio-medical event. The page shows a list of publications containing this specific event.

Challenge [15]. Specifically, we have identified gene/proteins related to the disease condition, characterizing every time the nature of the relations: up- or under-regulation; correlation; susceptibility or potential involvement (research hypothesis), as well as contradictory evidence brought down by various articles.

Parkinson's Disease map: integration and visualization of disease related data Similar in flavour, our system is used to extract supporting evidence and/or suggest new candidates for inclusion to disease maps which is another instance of disease modeling networks. Parkinson's Disease map is one such example [16]. Step IV in the Figure 2 shows how *GLUT4* was approved and appropriately integrated in the PD map.

6.2 System Strengths, Limitations and Future Work

Our system is constructed with the goal of detailed knowledge extraction from textual data, its availability to human and machine. Its strength is the ability to process abstracts as well as full texts; extract semantic relations between various concept types and contextualize them in terms of location, conditions, logic and temporary order. A web interface offers public and free access to the knowledge base while a SPARQL endpoint offers a machine readable access.

Some aspects of the system will be further developed and there remains room for change and improvement. First of all, the benchmarking of the system accuracy needs to be performed. From the text mining perspective, it operates on the sentence level which limits its recall. Although extracted knowledge is normalized with respect to concepts and relations, various nomenclatures are used. To increase knowledge interoperability we plan to adopt Unified Medical Language System (UMLS) which capitalizes on straightforward communication between various systems processing biomedical and health related data. Currently

triple store covers main attributes of the extracted relations, such as subject-predicate-object while contextual aspects need to be incorporated. Future work will include enriching the scope of entity types, extending the current web application by adding, among other developments, an advanced search feature, a personalized notification system, a REST web service and some bibliographic management system to easily cite the publications. BioKB will also have to be continuously extended by processing more publications.

7 Conclusions

In this paper we described an information extraction system along with the storage database and web interface in the field of biomedicine. The system employs text mining and semantic technologies to help discovery and accessibility of biomedical knowledge. As a proof of concept, we have shown its applicability to disease network construction and enrichment. Along with the strengths, we have pointed out the system's limitations and outlined future work directions.

8 Acknowledgements

This work was partially conducted in the scope of the eTRIKS project that received funding from the European Union and from the European Federation of Pharmaceutical Industries and Associations as an IMI JU funded project (no. 115446). The Reproducible Research Results (R3) team of the Luxembourg Centre for Systems Biomedicine is acknowledged for support of the project and for promoting reproducible research.

References

1. Pafilis E., et al. *Reflect: augmented browsing for the life scientist.* Nat. Biotechnol., 2009, vol. 27, pp. 508-510.
2. Kim Jin-Dong, Tomoko Ohta, Yuka Tateisi and Jun'ichi Tsujii. *GENIA corpus - a semantically annotated corpus for bio-textmining.* Proceedings of the Eleventh International Conference on Intelligent Systems for Molecular Biology, Brisbane, Australia, 2003, pp. 180-182.
3. *NLTK - Natural Language Toolkit.* `http://www.nltk.org`
4. Princeton University *"About WordNet."* Princeton University. 2010. `http://wordnet.princeton.edu`
5. *The Stanford Parser: A statistical parser.* `https://nlp.stanford.edu/software/lex-parser.shtml`
6. *Stanford Dependencies.* `https://nlp.stanford.edu/software/stanford-dependencies.html`.
7. David McClosky, Mihai Surdeanu, and Christopher D. Manning. 2011. *Event Extraction as Dependency Parsing.* In Proceedings of the Association for Computational Linguistics - Human Language Technologies 2011 Conference (ACL-HLT 2011), Main Conference.

8. Gunes Erkan, Arzucan Ozgur and Dragomir R. Radev. *Extracting Interacting Protein Pairs and Evidence Sentences by usnig Dependency parsing and Machine Learning Techniques*. In Proceedings of the Second BioCreAtIvE Challenge Workshop - Critical Assessment of Information Extraction in Molecular Biology, April 23-25, 2007.
9. The Gene Ontology Consortium. *Gene ontology: tool for the unification of biology*. Nature Genetics, 25(1), 25–29, 2000.
10. Gremse, M., Chang, A., Schomburg, I., Grote, A., Scheer, M., Ebeling, C., and Schomburg, D. *The BRENDA Tissue Ontology (BTO): The first all-integrating ontology of all organisms for enzyme sources*. Nucleic Acids Research 39, 2011.
11. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, B. S. *The PubChem Project*. Nucleic Acids Research, 2016.
12. Schriml, L. M., Arze, C., Nadendla, S., Chang, Y. W. W., Mazaitis, M., Felix, V., ... Kibbe, W. A. *Disease ontology: A backbone for disease semantic integration*. Nucleic Acids Research, 40(D1), 2012.
13. Aken, B. L., Achuthan, P., Akanni, W., Amode, M. R., Bernsdorff, F., Bhai, J., ... Flicek, P. *Ensembl 2017*. Nucleic Acids Research, 45(D1), D635–D642, 2017.
14. *Genia Tagger* `http://www.nactem.ac.uk/GENIA/tagger/`
15. Aishwarya Alex Namasivayam et. al. *Community-Reviewed Biological Network Models for Toxicology and Drug Discovery Applications*. Gene Regulation and System Biology, vol 10, pp.51 - 66, 2016.
16. Satagopam Venkata et. al. *Integration and Visualization of Translational Medicine Data for Better Understanding of Human Diseases*. Big Data. June 2016, 4(2): 97-108.
17. *Vis.js* `http://visjs.org`
18. Leaman, R. and Gonzalez G. (2008) *BANNER: An executable survey of advances in biomedical named entity recognition*. Pacific Symposium on Biocomputing, pp. 652-663, 2008.
19. *ABNER: a Biomedical Named Entity Recognized*. `http://pages.cs.wisc.edu/~bsettles/abner/`
20. Li J, Zhu X, Chen JY. *Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts.* PLoS Computational Biology 5(7), 2009.
21. Rosario B and M.A. Hearst. *Classifying semantic relations in bioscience texts*. In Proceeding of the 42nd Annual Meeting on Association for Computational Linguistics, 2004.
22. Hoffmann R., Krallinger M., Andres E., Tamames J., Blaschke C., and Valencia A. *Text mining for metabolic pathways, signaling cascades, and protein networks.* Sci STKE, 2005.
23. Landeghem S., et. al. *Exploring Biomolecular Literature with EVEX: Connecting Genes through Events, Homology, and Indirect Associations.* Advances in Bioinformatics, 2012.
24. Kilicoglu, H., S. Bergler. *Syntactic Dependency Based Heuristics for Biological Event Extraction*. In Proceedings of the Workshop on BioNLP: Shared Task, pp. 119-127, 2009.
25. Björne, L. et. al. *Extracting Complex Biological Events with Rich Graph-Based Feature Sets*. In Proceedings of the Workshop on BioNLP: Shared Task, pp. 10-18, 2009.

26. Gawronska, B. Erlendsson and B. Olsson. *Tracking biological relations in texts: a Referent Grammar based approach.* Biomedical Ontologies and Text Processing, ECCB 2005.
27. Peng Y. et.al. *An extended dependency graph for relation extraction in biomedical text.* In Proceedings of the Workshop on Biomedical Natural Language Processing, pp. 21-30, 2015.
28. Liekens A. et.al. *BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation.* Genome Biology, 12(6), 2011.
29. *STRING: Protein-Protein interaction networks.* `https://string-db.org`
30. Li C., Liakata M. and Rebholz-Schumann D. *Biological network extraction from scientific literature: state of the art and challenges.* Briefings in Bioinformatics, 15(5), pp. 856-877. 2014.
31. Rebholz-Schumann, D. et.al. *Assessment of NER solutions against the first and second CALBC Silver Standard Corpus.* Journal of Biomedical Semantics, 2 (Suppl 5), 2011.
32. Johnson H., et. al. *Corpus Refactoring: a Feasibility Study.* Journal of Biomedical Discovery and Collaboration, 2007.
33. Kim J. and Y. Wang. *PubAnnotation - a persistent and sharable corpus and annotation reposit.* In Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012), pp. 202–205.
34. Clark, T., et. al. (2014). Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications.
35. Mons, B. and Velterop, J. (2009). *Nano-Publication in the e-Science era.* In Proceedings of the International Semantic Web Conference, 2009.
36. Piñero J. et. al (2016). *DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants.* Nucleic Acids Research, vol. 45(D1), pp.D833-D839, 2017.
37. Cañada A. et. al. *LimTox: a web tool for applied text mining of adverse event and toxicity associations of compounds, drugs and genes.* Nucleic Acics Reseaech, 45(W 1), 2017.
38. Cheng D. et.al. *PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites.* Nucleic Acics Reseaech, 36, 2008.