Ilkogretim Online - Elementary Education Online, 2019; Vol 18 (Issue 1): pp. 636-644
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2019.527672

A Case Study For Protein Localization-Mining Pubmed Data Using Loctext

Zahrah Ayub Department of Computer Science and Engineering, Islamic University of Science and Technology, Jammu and Kashmir, India.

Asif Ali Banka Department of Computer Science and Engineering, Islamic University of Science and Technology, Jammu and Kashmir, India.

Muneer Ahmad Department of Computer Science Technical University of Munich, Germany.

Roohie Naaz Department of Computer Science and Engineering, NIT Srinagar, Jammu and Kashmir, India.

Correspondence: *Zahrah Ayub

Abstract
Data sharing and web 2.0 have revolutionized the whole new idea and context of thinking about data. Generation and consumption of data has become integral part of human life. Data generated is in various formats and most widely accepted is text data. The biomedical literature is exploding and it is complexity increases to keep track of publications in relevant areas of interest. In this work we attempt to implement LocText on PubMed data to illustrate relationships between protein and sub-cellular locations. The aim of experimentation is to match the UniPortKB accession numbers, GO Cellular Component identifiers to NCBI taxonomy identifiers available in PubMed data. The experiments are performed on a commodity cluster that comprise of 16 machines installed with Ubuntu 16.04 Operating System, Java 8, Hadoop 2.7, Spark2.1, Elasticsearch, Nalaf, LocText and Kibana.

Keywords LocText, Protein Localization, PubMed, Text Mining.

I. INTRODUCTION

Data sharing and web 2.0 have revolutionized the whole new idea and context of thinking about data. The interconnected devices have practically resulted in data intensive computing where data is now essential part of human life [1]. Generation and consumption of data has become integral part of human life. Raw data available to users is diverse and complex in nature, consisting primarily of un- structured and unsupervised data. Pre-processing of data to extract ordered representation of data for downstream consumption is challenging. Data generated is in various

formats and most widely accepted is text data. Text data is simplest form of raw unprocessed information. This unstructured text data needs to be mined for processing and generating inferences. Text mining refers to identifying useful patterns in data [2] [3] [4]. Pattern of interaction between data elements reflect a lot of information about nature of underlying datasets.

A general text mining pipeline begins with normalizing names of entities mentioned in text. That is referred to as named-entity recognition (NER) and it is a step to match entities to the accepted vocabulary. It is followed by relation extraction (RE), which is explained in detail below. One of most widely terms used in bioinformatics is ontology, which actually is representation of relationship between entities and their properties. The concept of ontologies is extensively used in information ex- traction and relation extraction [2] [5]. Information extraction is generating structured inferences from unstructured text data using automation methods. These ontology based information extraction methods provide semantic identity to extracted entities that are further used to located and classify entities like gene names, protein names and localizations. This is called Named Entity Recognition (NER). The unstoppable growth of entities and their existence in literature and their existence with equivalent word is challenging for NER system to cope up [2] [5] [6] [7]. The Relation Extraction (RE) determines the relationship among various entities and properties. The relationship can be binary or m-ary but primarily in bioinformatics the relation-ship between genes and proteins, their location and its effect on functionality of gene is considered to be exciting area of research. Annotating and archiving these relations in databases is one hell of a task for database curators [7] [8] [9]. Thus proper mining and archiving of these medical informatics articles is of great interest to researchers [6] [10]. There exist a number of methods to address these issues but identifying relationships between protein and sub- cellular locations in same sentence is still challenges even with current state-of-art computing resources [11]. In this work we attempt to implement one such framework; LocText on PubMed data to illustrate relationships between protein and subcellular locations.

II. RELATED WORK AND FRAMEWORK

In this work we implement LocText on PubMed data to illustrate relationships between protein and subcellular locations. LocText is a state-of-art method that helps to identify and annotate relationships between protein and subcellular locations from text data [11]. Sub cellular location is an important factor that governs functioning of a protein. Annotating these scientific literatures with text mining techniques is expected to aid database curators to identify relationship between the protein and location that shall help understanding the functioning of proteins which remains to be unattended field for bioinformatics. Some methods like WoLF PSORT [12], SignalP [13] and LocTree3 [14] have been proposed that try to extract the ex- act location of a protein to assist Gene Ontology database but there is no generic text mining frame-work to facilitate the pipeline. As we know the text mining begins with normalizing names of entities like proteins mentioned in text. That is referred to as named-entity recognition (NER) and it is a step to match entities to the accepted vocabulary. As shown in Figure 1, the named-entity recognition (NER) identifies the protein to its UniportKB identifier and respective cell compartments to Gene Ontology. This is followed by relation extraction (RE) which identifies

the relations deduced from semantic context. LocText is a fully automated pipeline that tries to address the challenge using machine learnt semantics and syntax of scientific data.

A. PubMed and PubMed Central

One of major applications of text mining in bioinformatics is to provide preprocessed data to its stakeholders which include splitting sentences, identifying parts of speech, tokenization and named entity recognition. PubMed is an effort to provide preprocessed, computationally expensive data to scientific com- munity [7] [8] [15]. Data is fuel to technology and we look to move beyond data archival. PubMed, is a publicly available prime source of biomedical datawith around 26 M abstracts, 1.4 M full articles, 388 M analyzed sentences [7] [16]. Manual curation of such huge database is infeasible so we realize need for automated curation of these massive datasets. The database resource is a reliable, fully updated and automated distributed data and provides data in readily and suitable *.xml and *.txt formats [7] [16] [17]. The automated curation of data is expected to allow easier and swifter information and relation extraction.

B. Nalaf

There is a fast growing gap between theusers of specialized repositories and technical people working on top of these repositories trying to help with their natural language processing expertise. NALAF (NA)tural (LA)nguage (F)ramework is a NLP framework written in python that aimsto generate domain specific entity tagging from free text currently maintained by JuanMiguel Cejuela [18]. The framework was proposed as thesis at Rostlab at Technische Universität München to cover the task of named entity recognition and relationship extraction with automated training and annotation. The framework is open source andwidely adopted by many biomedical and bioinformatics natural language processing to integrate relationships between entities due to its high precision and low recall [10]. The framework is bridging the gap between semantic resource annotations and efficient discovery of entities in literature.

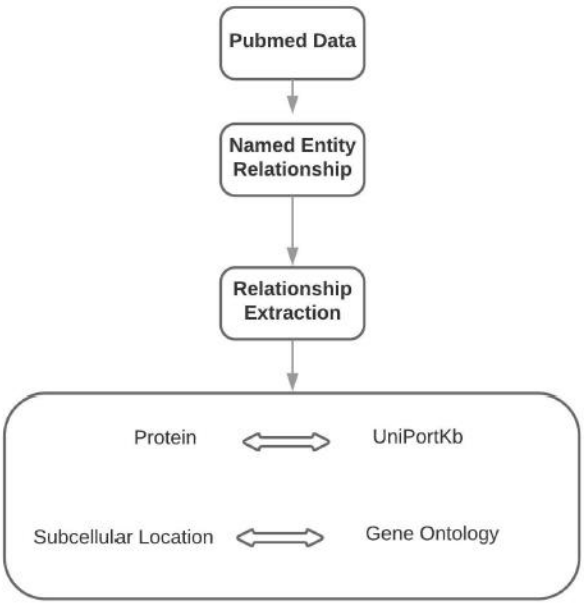


Fig 1: Process Outline

C. LocText

It is a method to identify relation between location of a protein in subcellular compartment and its functionality [11]. The information is extracted from abstracts and texts in PubMed data and its relation with named entity is observed.LocText learns the patterns from parse tree and is trained and evaluated over it. STRING Tagger (NER) is used to extract proteins, subcellular localizations and organism information to improve performance [11]. It improves Gene Ontologyannotations [19] were enhanced in Swiss-Port and UniPortKB [20] with efficient computational cost [21] [22] [23] [24]. LocText works in combination with nalaf for enhancing relationship extraction. Various techniques are used for feature selection, however, LocText claims Lasso L1 regularization to be most efficient [11][25] [26][27]. The aim of experimentation is to match the UniPortKB accession numbers, GO Cellular Component identifiers to NCBI taxonomy identifiers available in PubMed data [28]. The tagging was done using nalaf and results are sorted using GO identifiers. The protein location relationship is further plotted using Kibana.

III. IMPLEMENTATION DETAILS

The experiments are performed on a commodity cluster that comprise of 16 machines installed with Ubuntu 16.04 Operating System, Java 8, Hadoop 2.7 7 [29], Spark 2.1 [30], Elasticsearch [31], Nalaf [18], STRING Tagger, LocText [11] and Kibana [31]. Stepwise implementation walk through detailsare presented in Figure 2.

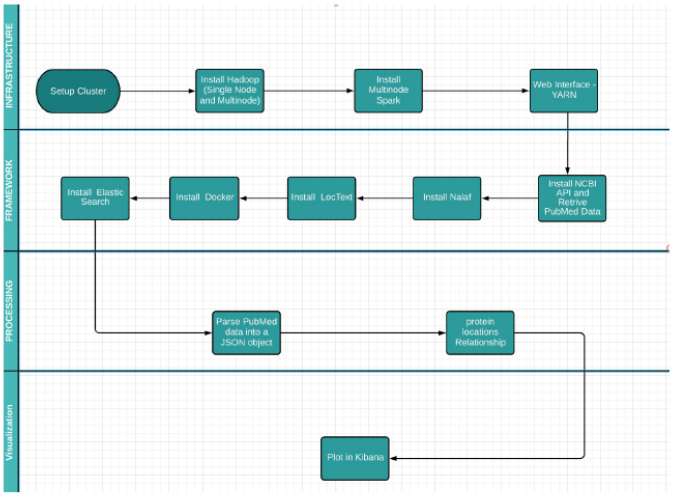


Fig 2: Step-wise Implementation Walk-through

IV. RESULTS

Each protein is associated to some location or some-times to more than one location in a cell. It is a m*m association. In the plot 3 below we have 12 UniPortKB association numbers plotted against location. This show UniPortKB association numbers (uac) vs Gene Ontology (GO) association. The inner slices represent 12 uacs and outer ring represents the presence at different locations.

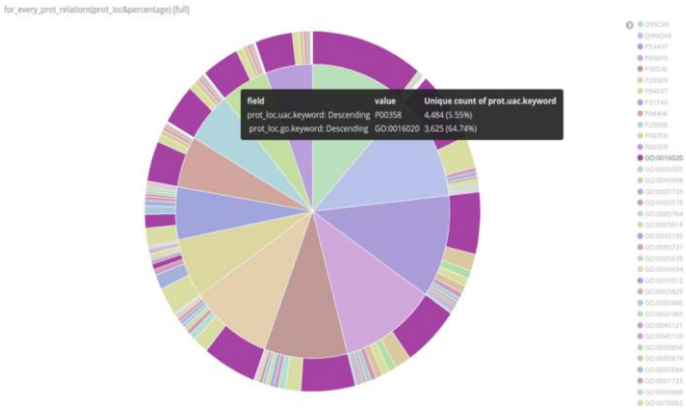


Fig 3: Association of uac and protein location

The plot below 4 gives number of occurrences ofprotein for each protein reference. We have plotted occurrences for 15 protein references. These are plotted in order of occurrences. It represents the number of references of a protein at a particular location. The plot below 5 depicts the distribution of UniPortKB association numbers in various Gene Ontologies (GO). Here we plot 5 protein for

15 different location and observe their presence. In every text there is presence of UniPortKB association numbers (protein). In the plot below we look for presence of uac in text. The plot below 6shows the distribution of uacs taken randomly.

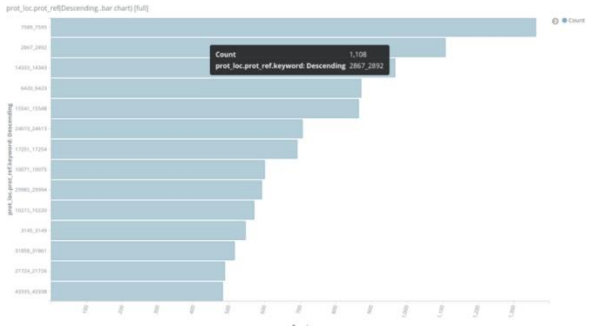


Fig 4: Number of occurrences of protein reference on protein location

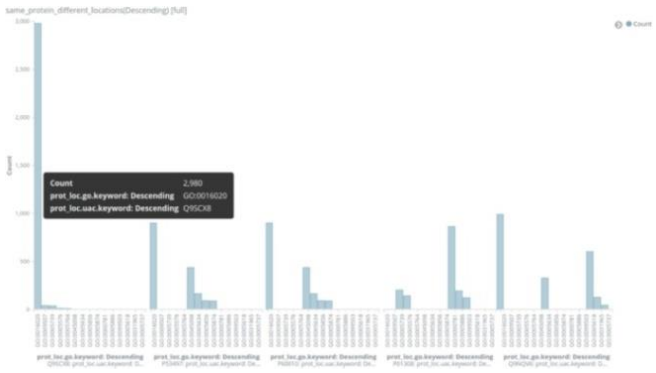
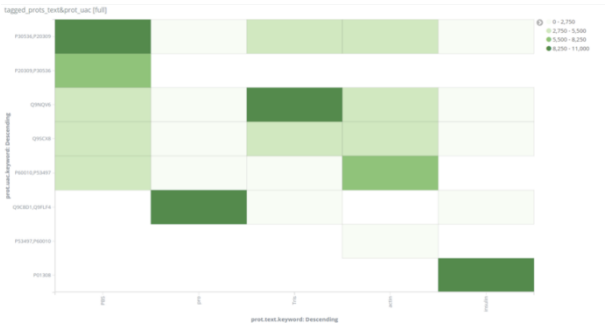


Fig 5: UAC distribution in various locations



s

Fig 6: uac presence in t

The plot below 7 is top 10 proteins and there uac-GO association according to their presence in PubMed data considered for experimentation.

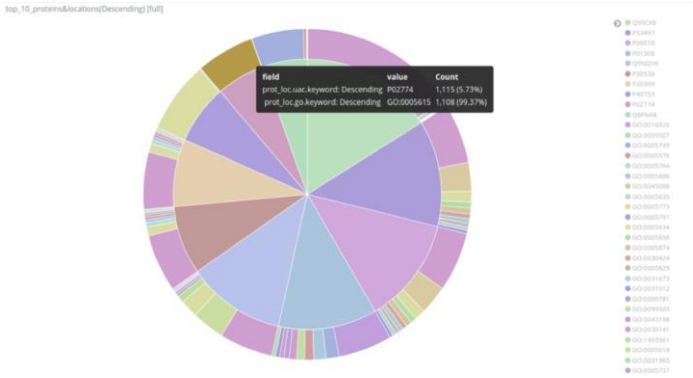


Fig 7: Top 10 proteins according to their uac and locations

The image below 8 is word cloud plotted to see the occurrences of words in the dataset. Sometimes sentences are put in double quotes so they also are present in the plot.

Fig 8: Word Cloud

V. CONCLUSION

In this paper we outlined the need for database curation and text mining in bioinformatics and biomedical domains. Increased growth of repositories and literature in these fields with varied nomenclature for their entities across globe makes it even more challenging to address such issues. Various platforms have been proposed to tackle the ever increasing scientific information. In this paper we attempted to implement once such framework (Loc- Text) as case study and infer its impact on database curation. A state-of-art method was adopted that runs on top of current age technologies to

identify relationship between protein and subcellular compartments. The experimentation was performed on publicly available on PubMed data. The aim of experimentation is to match the UniPortKB accession numbers, GO Cellular Component identifiers to NCBI taxonomy identifiers available in PubMed data.

REFERENCES

[1] Gantz, J., & Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. IDC IView: IDC Analyze the Future, 2007, 1–16.

[2] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. Retrieved from http://arxiv.org/abs/1707.02919

[3] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1–47. doi:10.1145/505282.505283C.

[4] Aggarwal, C., & Zhai, C. (2012). Mining text data. Springer Science & Business Media.

[5] Cohen, K. B., & Hunter, L. (2008). Getting started in text mining. PLoS Computational Biology, 4(1), e20. doi:10.1371/journal.pcbi.0040020

[6] Yu, L. (2011). A developer's guide to the semantic Web. Springer Science & Business Media.

[7] Hakala, K., Kaewphan, S., Salakoski, T., & Ginter, F. (2016). Syntactic analyses and named entity recognition for pubmed and pubmed central-up-to-the-minute. In Proceedings of the 15th Workshop on Biomedical Natural Language Processing (pp. 102–107).

[8] Wikipedia contributors. (2022, May 16). PubMed. Retrieved from Wikipedia, The Free Encyclopedia website: https://en.wikipedia.org/w/index.php?title=PubMed&oldid=1088240327

[9] Leaman, R., & Gonzalez, G. (2008). BANNER: an executable survey of advances in biomedical named entity recognition. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 652–663.

[10] Sfakianaki, P., Koumakis, L., Sfakianakis, S., Iatraki, G., Zacharioudakis, G., Graf, N., Tsiknakis, M. (2015). Semantic biomedical resource discovery: a natural lan- guage processing framework. BMC Medi- Cal Informatics and Decision Making, 15(1).

[11] Cejuela, J. M., Vinchurkar, S., Goldberg, T., Shankar, M. S. P., Baghudana, A., Bo- Jchevski, A., Jensen, L. J. (2018). Loctext: relation extraction of protein localizations to assist database curation. BMC Bioinformatics, 19(1).

[12] Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., & Nakai, K. (2007). WoLF PSORT: protein localization predictor. Nucleic Acids Research, 35(Web Server issue), W585-7. doi:10.1093/nar/gkm259

[13] Petersen, T. N., Brunak, S., von Heijne, G., & Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods, 8(10), 785–786. doi:10.1038/nmeth.1701

[14] Goldberg, T., Hecht, M., Hamp, T., Karl, T., Yachdav, G., Ahmed, N., Balasz, K. (2014). Loc- tree3 prediction of localization. Nucleic Acids Research, 42(W1), W350–W355.

[15] Miwa, M., Thompson, P., Mcnaught, J., Kell, D. B., & Ananiadou, S. (2012). Extracting semantically

enriched events from biomedical literature. BMC Bioinformatics, 13(1).
[16] Moen, S., & Ananiadou, T. S. S. (2013). Distributional semantics resources for biomedical text processing. In Proceedings of LBM (pp. 39–44).
[17] Tanabe, L., Xie, N., Thom, L. H., Mat- Ten, W., & Wilbur, W. J. (2005). Genetag: a tagged corpus for gene/protein named entity recognition. BMC Bioinformatics, 6(1).
[18] Nalaf: NLP framework in python for entity recognition and relationship extraction. (n.d.).
[19] Ashburner, M., Ball, C. A., Blake, J. A., Bot- Stein, D., Butler, H., Cherry, J. M., Eppig, J. T. (2000). Gene ontology: tool for the unification of bi- ology. Nature Genetics, 25(1).
[20] Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., & Bairoch, A. (2007). UniProtKB/Swiss-Prot. Methods in Molecular Biology (Clifton, N.J.), 406, 89–112. doi:10.1007/978-1-59745-535-0_4
[21] Van Auken, K., Jaffery, J., Chan, J., Muller, H.-M., & Sternberg, P. W. (2009). Semi- automated curation of protein subcellular localization: a text mining-based approach to gene ontology (go) cellular component cura- tion. BMC Bioinformatics, 10(1).
[22] Zheng, W., & Blake, C. (2015). Using distant super- vised learning to identify protein subcellular localizations from full-text scientific articles. Journal of Biomedical Informatics, 57, 134–144.
[23] Binder, J. X., Pletscher-Frankild, S., Tsafou, K., Stolte, C., O'Donoghue, S. I., Schneider, R., & Jensen, L. J. (2014). COMPARTMENTS: unification and visualization of protein subcellular localization evidence. Database: The Journal of Biological Databases and Curation, 2014(0), bau012. doi:10.1093/database/bau012
[24] Fyshe, A., Liu, Y., Szafron, D., Greiner, R., & Lu, P. (2008). Improving subcellular localization prediction using text classification and the gene ontology. Bioinformatics, 24(21), 2512–2517.
[25] Tibshirani, R. (1996). Regression shrinkage and se- lection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288.
[26] Yu, C.-S., Chen, Y.-C., Lu, C.-H., & Hwang, J.-K. (2006). Prediction of protein subcellular localization," Proteins: Structure. Proteins: Structure, Function, and Bioinformatics, 64(3), 643–651.
[27] Briesemeister, S., Rahnenfi¿ ½hrer, J. ½., & Kohlbacher, O. (2010). YLoc—an interpretable web server for predicting subcellular localization. Nucleic acids research, 38(suppl_2), W497-W502.
[28] Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research, 30(1), 207-210.
[29] Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010, May). The hadoop distributed file system. In 2010 IEEE 26th symposium on mass storage systems and technologies (MSST) (pp. 1-10). IEEE.
[30] Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., & Stoica, I. (2016). Apache spark: a unified engine for big data processing. Communications of the ACM, 59(11), 56-65.
[31] Kononenko, O., Baysal, O., Holmes, R., & Godfrey, M. W. (2014, May). Mining modern repositories with elasticsearch. In Proceedings of the 11th working conference on mining software repositories (pp. 328-331).