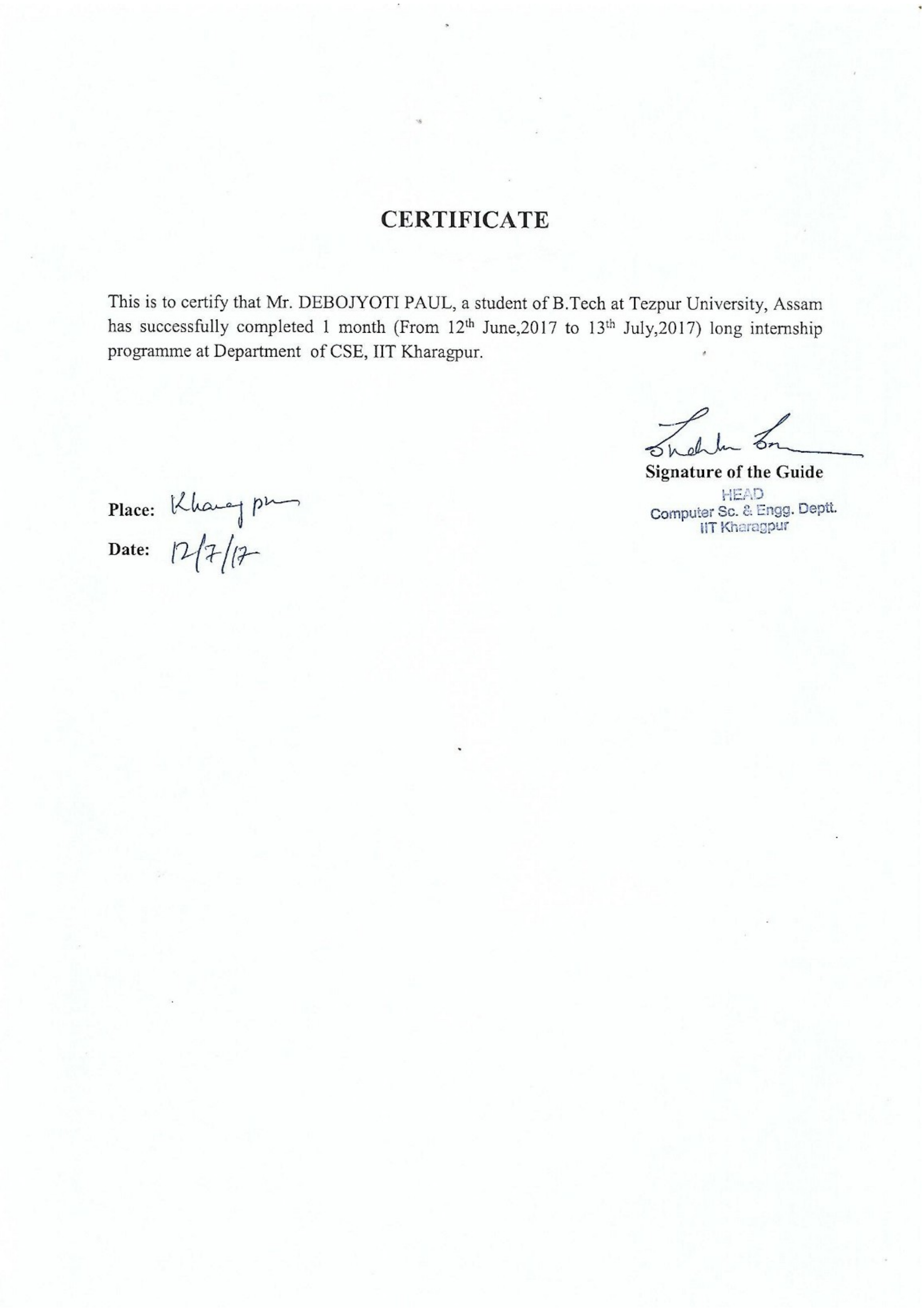
Indian Institute of Technology
Kharagpur

Summer Internship Project report

Information Extraction using co-occurrence and co-occurrence networks.

Submitted To:
Sudeshna Sarkar
Professor and Head
Computer Science
Department

Submitted By :
Debojyoti Paul
CSB14005
Computer Sceince
Department
7th semester



CERTIFICATE

This is to certify that Mr. DEBOJYOTI PAUL, a student of B.Tech at Tezpur University, Assam has successfully completed 1 month (From 12th June,2017 to 13th July,2017) long internship programme at Department of CSE, IIT Kharagpur.

Signature of the Guide

HEAD
Computer Sc. & Engg. Deptt.
IIT Kharagpur

Place: Kharagpur

Date: 12/7/17

Acknowledgements

The internship opportunity I had with the Department of Computer Science & Engineering, Indian Institute of Technology, Kharagpur was a great chance for learning and professional development. Therefore, I consider myself as a very lucky individual as I was provided with an opportunity to be a part of it. I am also grateful for having a chance to meet so many wonderful people and professionals who led me through this internship period.

I am using this opportunity to express my deepest gratitude and special thanks to Dr. Sudeshna Sarkar, Professor, Department of CSE, IIT Kharagpur who in spite of being extraordinarily busy with her duties, took time out to hear, guide and keep me on the correct path and allowing me to carry out my project at their esteemed organization.

I perceive this opportunity as a big milestone in my career development. I shall strive to use the gained skills and knowledge in the best possible way, and continue to work on their improvement, in order to attain desired career objectives. Hope to continue cooperation with all of you in the future.

Sincerely,
Debojyoti Paul.

Abstract

Automatically extracting information from biomedical articles is important for consolidating large amounts of biological knowledge in computer accessible form. The projects aims to aid in information extraction by firstly analyzing the entities extracted from 10,000 PubMed abstracts. This is achieved by computing features like frequency distribution, co-occurrence between entities and thus forming a co-occurrence network that may be used for information extraction. Also, a web based portal is developed that returns similar entities from the a entity name entered. Entities with higher co-occurrence value are displayed.

Keywords: Named Entity recognition, Co-occurrence, Co-occurrence networks.

Contents

List of Figures

1 Introduction

An incredible wealth of biological information generated using biochemical and genetic approaches is stored in published articles in scientific journals. However, retrieving and processing this information is very difficult due to the lack of formal structure in the natural language in these documents. Thus, automatically extracting information from biomedical text is important for consolidating large amounts of biological knowledge in computer accessible form.

Information extraction (IE) systems could gather information on gene relationships, gene functions, gene-gene interactions, and other important information on biological processes.

Also,the text mining and information extraction strategies that are applied to the biomedical and molecular biology literature is possible due to the increasing number of available publications stored in databases such as PubMed.

However, the first phase of extracting information about genes or proteins consists of the initial challenge of recognizing the gene/protein names in the text. This process is complicated because of the lack of standardized gene naming conventions biology. Gene/protein name styles vary considerably from organism to organism, and also might differ for the same organism. Thus, Grammar and Dictionary based methods fail for the task of understanding of text giving importance to rise of text mining techniques.

Text mining applied to texts and literature of the biomedical and molecular biology domain is known as **BioMedical text mining** (also known as BioNLP)

1.1 Problem defination

As name entity recognition is such a important first step in the process of information extraction, I would like to analyze the entities found in this phase and develop a co-occurrence network network that show relation between pair of entities found from 10,000 PubMed abstracts. The network is based on co-occurrence value, which is defined in section 2.5. Also, I have developed a web portal for retrieving the most likely associated entities to a entity given by a user. The association is represented in terms of co-occurrence.

2 Chapter 2: Background and Overview

2.1 Information Retrival

Information retrieval is concerned with identifying, a subset of docuements within a large document collection,content is most relevant to a our need. Also, to be more precise, given a specification represented as query - the goal of IR is to find the documents in the database that satisfy the information need.

PubMed[3] is an example of Information retrival System in medical domain. It is service of the National Library of Medicine that includes over 15 million bibliographic citations from MEDLINE and other life science journals for biomedical articles.

2.2 Named Entity Recognition

Named entity Recognition is the classification of proper nouns, dates, time, measures and locations,etc. However in Medical Domain, Bio-entity recognition aims to identify and classify technical terms in the domain of molecular biology that correspond to instances of concepts. Entity recognition is a core part several higher level information access tasks such as information extraction, summarization, etc. It aim find structure in unstructured text data and which will aid in finding relevant factual information.

Meanwhile, the task of entity recognition consists of lot of Challenges for example due to ambiguity in determining the boundaries(the previous and next word are also considered, i.e sequence labeling problem).Additional problems arises in BioMed text Mining because of the irregular Naming conventions of entities, very long names of entities,complicated constructions,etc. This makes annotation of human training data very difficult.

Label is the entity names to which entities are mapped. Abner[1] classifies biomed entities into 5 types (or labels), i.e. DNA, RNA, cell line and cell type.

Named entity recognition can be considered as sequence labeling problem. Several models for this NER considered such as Maximum Entropy Model, Hidden Markov Model, Conditional Random Fields. Abner[1] uses CRF framework to label the tokens in input sentences.

2.2.1 Conditional Random Fields

Biomedical named entity recognition can be considered of as a sequence segmentation problem: each word is a token in a sequence to be assigned a label (e.g. PROTEIN, DNA, RNA,CELL-LINE, CELL-TYPE, or OTHER).

Conditional Random Fields (CRFs) are undirected statistical graphical models, and is a special case of which is a linear chain that corresponds to a conditionally trained finite-state machine.[2]. This kind of models is well suited to sequence analysis, and CRFs in particular have shown to be very useful in part-of-speech tagging and named entity recognition .

$$f = \frac{1}{Z_0} exp\left(\sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_j f_j(s_{i-1}, s_i, o, i) \right) \tag{1}$$

where Z_o is a normalization factor of all state sequences, $f_j(s_{i-1}, s_i, o, i)$ is one of m functions that describes a feature, and λ_j is a learned weight for each such feature function.

2.3 Information Extraction

Information extraction is identifying information from text articles. Information extraction systems automatically identify entities and their relationships from free text, producing a structured representation of the relevant information stated in the input text. For example, if a protein is mentioned often in the same abstracts as a disease, it is reasonable to hypothesize that the protein is involved in some aspect of the disease.

2.4 Co-occurrence

Co-occurrence of a pair is the measurement of number of abstracts where both the entities are found. The relation formed between entities is based on the fact that terms that appear together tends to be related.

An example is relation extraction of Protein-protein interaction where text mining system with high coverage were based on the concept of co-occurrence: that is, two proteins co-occurring in a textual unit of some defined size.

Formally, Co-occurrence of two terms t1 and t2, is the simplest confidence metric is the count,c of texts that include both terms

$$c(t_1 \wedge t_2)$$

However, to remove the possibility of random co-occurrences due to more popularity of one term over another the measure is normalized. For example,

$$\frac{c(t_1 \wedge t_2)}{c(t_1).c(t_2)}$$

Pointwise mutual information (PMI) is similarly derived as

$$PMI(t_1, t_2) or I(t1, t2) = log_2(\frac{N}{d}.\frac{p(t_1 \wedge t_2)}{p(t_1).p(t_2)})$$

where, N is the total number of texts.
and,d is the size of context window.

Two terms are said to be contextually similar if they they occur together within the context window.

2.4.1 Application of the co-occurrence

Synonym detection method can use the concept of Similarity – for identifying synonyms based on contextual similarity.[6] The **contextual similarity** gives sets of terms that appear in similar contexts. The idea behind is that synonyms of a term t can be observed by finding terms that appear in the same contexts as t. contexts of $t1$ and $t2$ are similar, then $t1$ and $t2$ are considered synonyms.
Also, $I(t1, w)$ and $I(w, t)$ as not same as they all mentions of w is checked wrt t only after t has been found, i.e to the right side of t.

Using mutual information, we can now define the similarity Sim between two terms t1 and t2, based on their respective contexts as:

$$\frac{\sum_{w \in lexicons} min(I(w, t1), I(w, t2)) + min(I(t1, w), I(t2, w))}{\sum_{w \in lexicons} max(I(w, t1), I(w, t2)) + max(I(t1, w), I(t2, w))}$$

where w ranges over the set of all of the words that appear in the respective contexts of t1 and t2. The value of the similarity $Sim(t1, t2)$ indicates whether t1 and t2 are synonyms.

3 Resources used

3.1 PubMed/Medline Abstracts

PubMed[3] is a free resource and is developed and maintained by the National Center for Biotechnology Information (NCBI), at the U.S. National Library of Medicine (NLM), located at the National Institutes of Health (NIH)[3]. it consists of over 26 million citations for biomedical literature from MEDLINE, life science journals, and online books. The PMC Open Access Subset(some or all openaccess content) is a part of the total collection of articles in Pubmed cenral.

The ftp service provided by the PubMed central (PMC) is used to download the abstracts. Important to note here , that abstracts only contain the main medical literature, and not full text including author names, journal number, figures, etc.

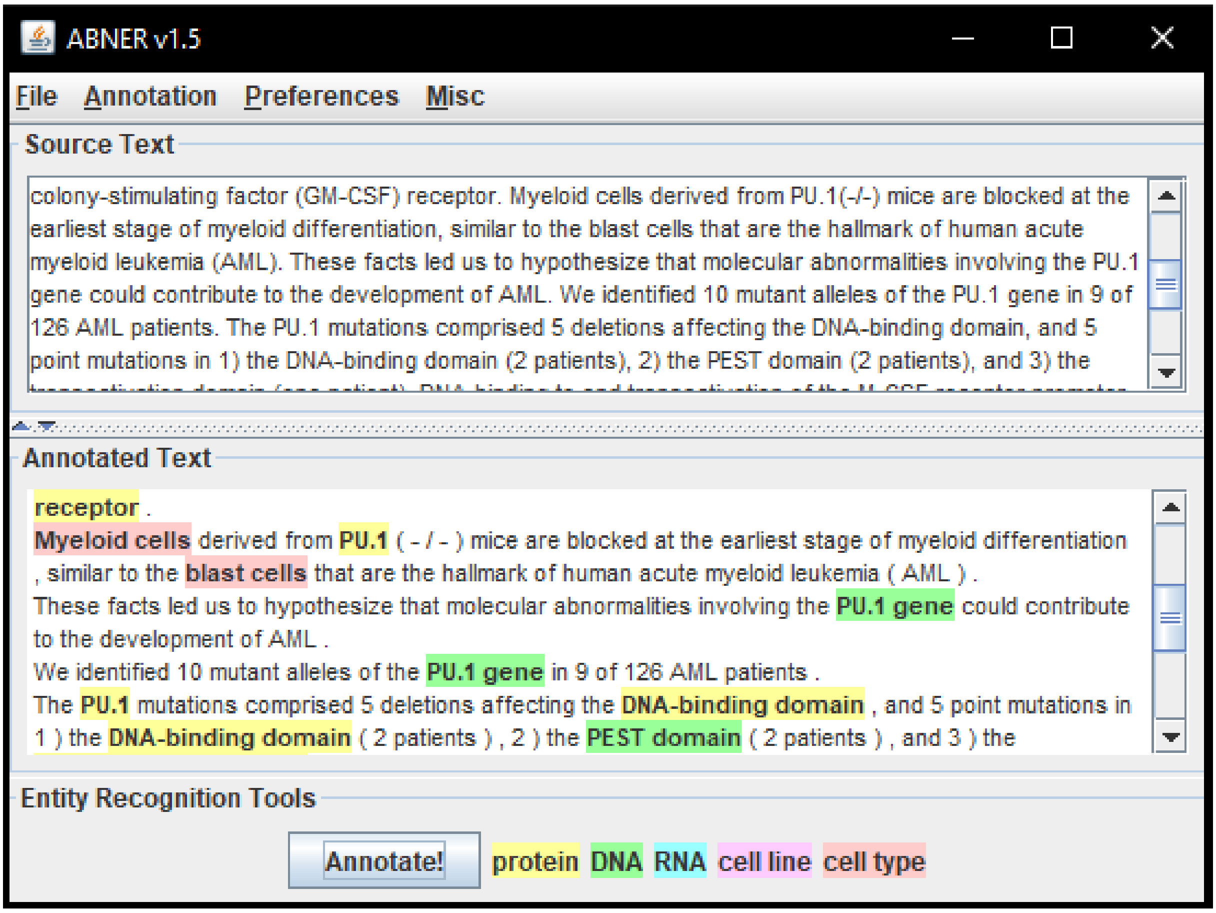
A subset of the bulk of articles named between C to H is used. The number of text files included in this subset is **10,000**. The Abstracts is passed through a ENTITY RECOGNIZER, ABNER [1]. It identifies all the entities which are then analyzed.

3.2 ABNER

ABNER (A Biomedical Named Entity Recognizer) is an open source software tool for molecular biology text mining. It is based on a machine learning system using conditional random fields with a variety of orthographic and contextual features[2]. It identifies words and phrases belonging to certain classes (e.g. protein and cell line).

The software includes two built-in entity tagging modules that are trained and evaluated on the standard NLPBA and BioCreative corpora[1].

Abner is modeled based on Conditional random fields (CRFs) which are undirected statistical graph- ical models, a special case of which corresponds to conditionally trained finite-state machines well suited for labeling and segmenting sequence data. Words in a sentence are tokens to be assigned labels by states in the CRF framework(sequence labeling problem).



colony-stimulating factor (GM-CSF) receptor. Myeloid cells derived from PU.1(-/-) mice are blocked at the earliest stage of myeloid differentiation, similar to the blast cells that are the hallmark of human acute myeloid leukemia (AML). These facts led us to hypothesize that molecular abnormalities involving the PU.1 gene could contribute to the development of AML. We identified 10 mutant alleles of the PU.1 gene in 9 of 126 AML patients. The PU.1 mutations comprised 5 deletions affecting the DNA-binding domain, and 5 point mutations in 1) the DNA-binding domain (2 patients), 2) the PEST domain (2 patients), and 3) the

receptor .
Myeloid cells derived from PU.1 (- / -) mice are blocked at the earliest stage of myeloid differentiation , similar to the blast cells that are the hallmark of human acute myeloid leukemia (AML) .
These facts led us to hypothesize that molecular abnormalities involving the PU.1 gene could contribute to the development of AML .
We identified 10 mutant alleles of the PU.1 gene in 9 of 126 AML patients .
The PU.1 mutations comprised 5 deletions affecting the DNA-binding domain , and 5 point mutations in 1) the DNA-binding domain (2 patients) , 2) the PEST domain (2 patients) , and 3) the

Figure 1: Abner, named entity recognizing tool.

Here, ABNER identifies entities and marks them with different colors (like Red color for protein entities). Also, Batch Annotation can be done, by selecting a folder and a file type.

4 Observation

4.1 Frequency distribution in recognized named entities

As 10,000 abstracts where ran on ABNER, certain entities were discovered which are more common relative to others. Higher, frequency indicates more relations and more information can be extracted from this terms.

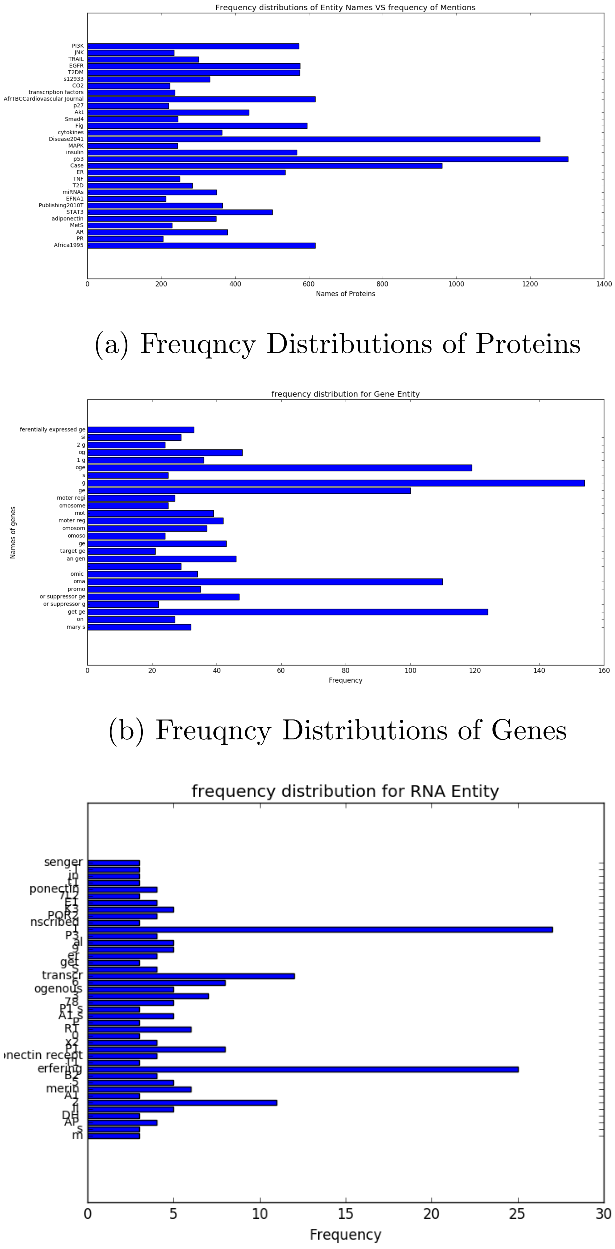
Figure 2.a. Proteins frequency distribution. Only protein with frequency more then 300 are choosen.
Figure 2.b. Gene frequency distribution. Only GENE mentions with frequency more then 20 are choosen.
Figure 2.c. RNA frequency distribution. Only RNA mentions with frequency more then 2 are choosen.

As we can see that certain non-biomed entities are falsely recognized as entities. This this known as **False Positives**.

ABNER was evaluated on with BioCreative and NLPBA corpus [1] and got a F-Score of 70.5[2]. F-Sore is defined as $f = \frac{2*R*P}{R+P}$

R is the Recall and is defines as fraction of relevant instances retrieved to the total instances in the corpara.
P is Precision and is defined as fraction of relevant instances among the retrieved instances.

Precision of ABNER was 69.1 and Recall 72.0 when trained on the mentioned corpus.



(a) Freuqncy Distributions of Proteins

(b) Freuqncy Distributions of Genes

(c) Freuqncy Distributions of RNA

Figure 2: Frequency distribution in recognized named entities

4.2 Network of the Genes and proteins

Here, I found co-occurrences of genes and proteins in 10,000 abstracts and used such co-occurrence to create a of network of 10000 genes and proteins. Edges between genes are weighted by the number of normalized articles that contained the pair. This network formed is a **co-occurrence network**.

By, definition, **co-occurrence** networks are the collective interconnection of terms based on their paired presence within a specified unit of text. Networks are generated by connecting pairs of terms using a set of criteria defining co-occurrence.[4]

The networks formed were :

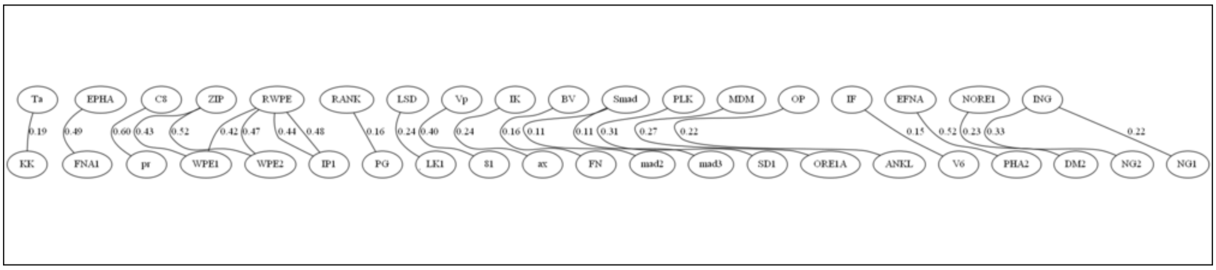


Figure 3: Co-occurrence network of proteins entities.

In the figure above, a network is formed among the most interacting proteins which appear most frequently in the Medline abstracts used. The coverage of this network can be measured by checking each pair of terms in the Database of Interacting Proteins[5].

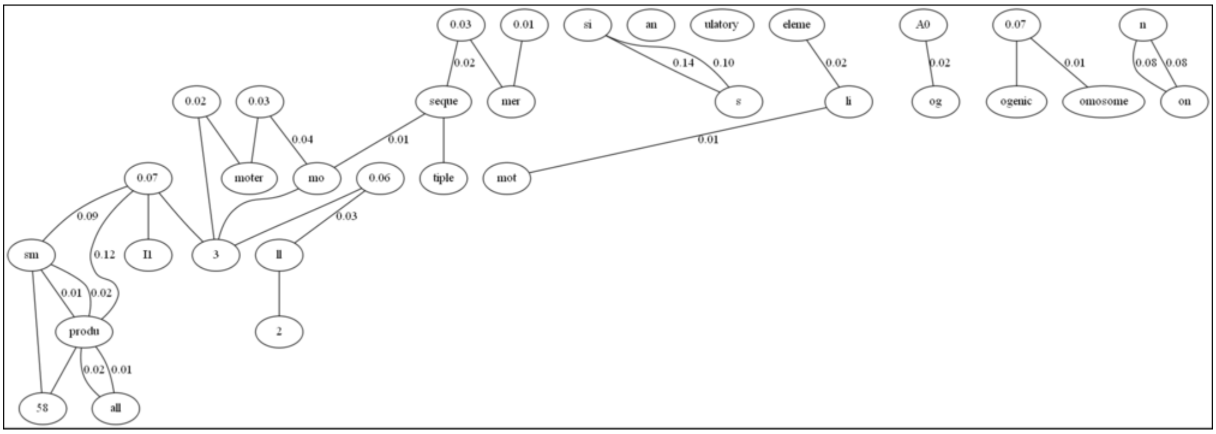


Figure 4: Co-occurrence network of Gene entities.

In the figure above, co-occurence network is formed based on Gene named entities. Edges indicates normalized co-occurrence scores.

PubGene is an example of an application that presents networks based on the co-occurrence of genetics related terms as these appear in MEDLINE records.

4.3 Web based View of co-occurrence network

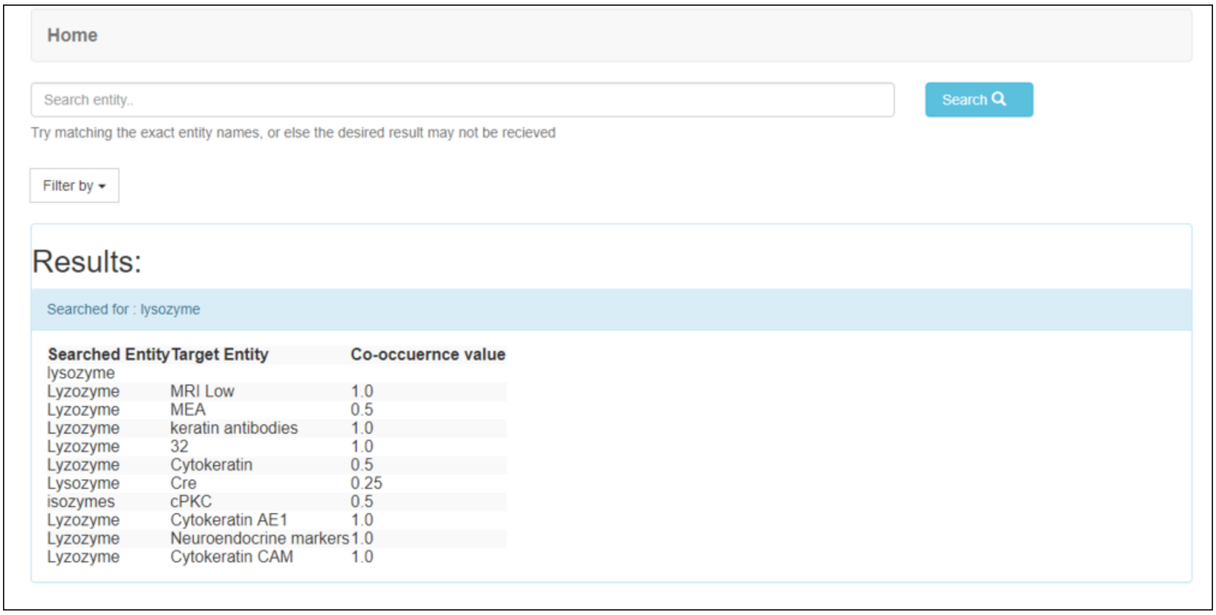
4.3.1 Description

The interface is developed based on the concept of similarity between entities. Two terms are considered more similar when co-occurrence value is high. It has two views :

- http://localhost:5000/search/genes. for genes names [5.3.2]
- http://localhost:5000/search/protein for protein names [5.3.2]

When made a Post request is made to to the API endpoint, list of entities is returned based on the level of similarity the input.
Similarly views for Proteins and RNA's are also created.

4.3.2 Screenshots



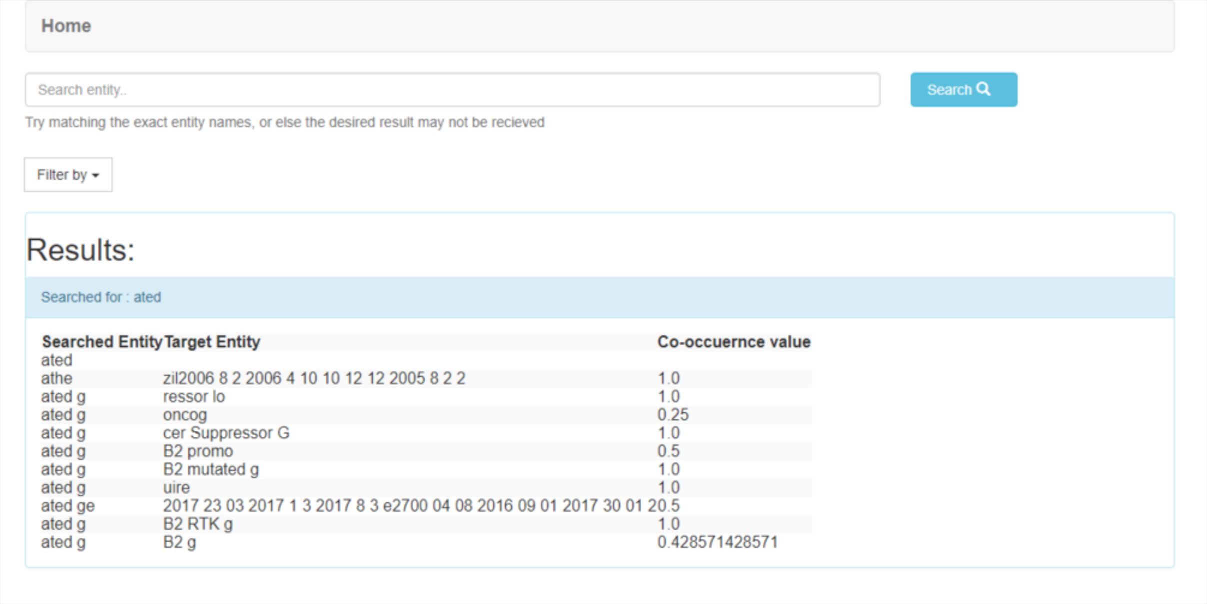
Try matching the exact entity names, or else the desired result may not be recieved

Results:

Searched for : lysozyme

Searched Entity	Target Entity	Co-occuernce value
lysozyme		
Lyzozyme	MRI Low	1.0
Lyzozyme	MEA	0.5
Lyzozyme	keratin antibodies	1.0
Lyzozyme	32	1.0
Lyzozyme	Cytokeratin	0.5
Lysozyme	Cre	0.25
isozymes	cPKC	0.5
Lyzozyme	Cytokeratin AE1	1.0
Lyzozyme	Neuroendocrine markers	1.0
Lyzozyme	Cytokeratin CAM	1.0

Figure 5: Search Results in the protein interface .



Try matching the exact entity names, or else the desired result may not be recieved

Results:

Searched for : ated

Searched Entity	Target Entity	Co-occuernce value
ated		
athe	zil2006 8 2 2006 4 10 10 12 12 2005 8 2 2	1.0
ated g	ressor lo	1.0
ated g	oncog	0.25
ated g	cer Suppressor G	1.0
ated g	B2 promo	0.5
ated g	B2 mutated g	1.0
ated g	uire	1.0
ated ge	2017 23 03 2017 1 3 2017 8 3 e2700 04 08 2016 09 01 2017 30 01 2	0.5
ated g	B2 RTK g	1.0
ated g	B2 g	0.428571428571

Figure 6: Search Results in the Gene interface.

4.3.3 Tools for creating the web interface

Flask : It is a microframework for Python based on Werkzeug, Jinja2. It is great for creating API as created above.
Jinja2: It is the templating engine that Flask uses.
Boostrap: For front-end design and javascript.

5 Conclusion and future work

Entity recognition is the primary step in the process of information extraction. Most common entities in Clinical articles being Genes and Gene Product mentions. Thus, high accuracy in named entity models is desirable. ABNER has a relatively high F-score and is a standard NER tool.

Co-occurrence networks drawn between similar entity terms gives rise new relationship being formed between pair of entities due to rule of transitivity.

Next step will be extracting entities from 10,000,00 abstracts. This is computationally very expensive and very resource extensive. However, this would increase the quality of the network formed such that richer information may be extracted. Another, particular improvement would be to be user tagger modules with higher F-score.

6 References

[1]. B. Settles (2005). ABNER: an open source tool for automatically tagging genes, proteins, and other entity names in text. Bioinformatics, 21(14):3191-3192., 2005.

[2]. Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi and Jun'ichi Tsujii. 2004.Introduction to the Bio-Entity Recognition Task at JNLPBA. *In Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications* (JNLPBA-2004)

[3]. *Pubmed.* US national Library of Medicine, national Institute of health.U pdated on 01 March 2016. Available online. Available form https://www.ncbi.nlm.nih.gov/books/NBK3827/#pub
Internet.

[4]. *Wikipedia: The Free Encyclopedia.* Wikimedia Foundation Inc. Updated 22 May 2017, at 15:09. Encyclopedia on-line. Available from https://en.wikipedia.org/wiki/Co-occurrence_networks, Internet.

[5]. Rodriguez-Esteban R (2009) Biomedical Text Mining and Its Applications. PLoS Comput Biol 5(12): e1000597. https://doi.org/10.1371/journal.pcbi.1000597

[6].Yu Hong and Agichten Eugene. Extracting Synonymous Gene and Protein Terms from Biological Literature, Vol. 1 no. 1 (2003):1-10.