ISSN 2319 – 1953

International Journal of Scientific Research in Computer Science Applications and Management Studies

Extracting Protein Names from Medline Abstract Using N-gram Techniques

Annalakshmi V[1], Bhuvaneswari V[2], Sheerin Rakshana A[3]

[1]*Asst. Professor, Dept. of Computer Science, Jayaraj Annapackiam College for Women (Autonomous), Periyakulam, Theni, Tamilnadu, India.*

[2]*AssociateProfessor, School of Computer Science & Engineering, Bharathiar University, Coimbatore, Tamil Nadu, India.*

[3]*PG Student, Dept. of Computer Science, Jayaraj Annapackiam College for Women (Autonomous), Periyakulam, Theni, Tamilnadu, India.*

[1]annalakshmivmca@gmail.com,[2]bhuvanes_v@gmail.com,[3]rakshanasheerin@gmail.com

***Abstract*— Extracting the protein name is an essential difficulty in the region of bioinformatics and biomedicine. This is used for measuring the several useful aspects of gene and protein groups. It is also exercised for ranking novel documents with importance to gene names and protein. Normally, protein names are referred in terms of protein names, gene symbol, synonyms, typographical variants and gene name. Classifying the gene / protein names are presented in biological literature is based on a variety of methods. Protein names are recognized from the dataset by using the Roman alphabets, capital letters, Roman numerals, Arabic numerals, and repeated words showing in protein names. In our effort we have planned a method to extracting protein name from Medline abstract using n-gram techniques.**

Keywords— Extracting Protein Names; Medline Abstract; n-Gram Techniques; Text Mining; Bioinformatics.

I. Introduction

Protein is a lengthy sequence molecule made up of amino acids connected by peptide bonds. Protein forms the structural material of bodily tissues. Proteins, the primary ingredients of the protoplasm of all cells, are of tall molecular weight and consist fundamentally of mixtures of amino acids in peptide relationships. Twenty dissimilar amino acids are normally established in proteins and each protein has a limited, hereditarily distinct amino acid sequence which decides its exacting function and shape.

Bioinformatics [1] obtain information from computer psychoanalysis of biological data. The biological data consists of genetic information, scientific literature and patient statistics. Research in bioinformatics contains technique growth for storage, analysis and retrieval of the data. Bioinformatics is a fast expanding division of biology and is really interdisciplinary, using concepts and techniques from statistics, informatics, mathematics, physics, chemistry, biochemistry, and linguistics.

The main purpose of this work is automatically extracting protein / gene name using n-gram approach from the biomedical literature. Text Mining illustrates an automated process of analysing natural language text with the goal of discovering information and knowledge.

II. REVIEW OF LITERATURE

Martin Krallinger [2] developed sub-tag set include protein variations which were produced from end to end, a rule supported pipeline of protein name processing.

Collier et al., [3] applied statistical methods for recognizing and categorizing gene and gene product names including proteins. The characteristics utilized in their techniques are the majority the similar as those applied in rule-based approaches, such as, surface clues and parts of speech.

Hong Yu [5] proposed proteins and genes are frequently characterized via names and symbols in literature. The names often are the lengthy forms of their symbols and illustrate the purposes of the proteins and genes.

Tanabe and Wilbur [6] have retrained Brill_s tagger on the biomedical area for protein / gene name-recognition. Statistical approaches have come together abstracts for keyword recognition [3].

Naive Bayes [7] described the Machine-learning approaches. He has applied Hidden Markov Models [8], and decision trees, to classify gene/protein terms. Other approaches comprise lookup in information sources such as SWISSPROT and GenBank [9].

Yoshida M et al.., developed, particularly for mapping protein symbols to filled names is PNAD-CSS (for ‘‘Protein full Name abbreviation Dictionary - Construction Support System’’). PNAD-CSS employed morphological characteristics to identify correct nouns as protein conditions in biological abstracts [11]. Identifying a phrase may hold a protein symbol and full name, PNAD-CSS texted dogged and parentheses whether the parenthetical expression was a short form of the outer phrase. To plan a protein symbol to its name, PNAD-CSS broke up words of the preceding phrase, and decided whether the parenthetical abbreviation candidate maps to the first letters of the broken-up phrase.

Krauthammer et al [12] presented a Basic Local Alignment Search Tool (BLAST)-trapped system approach. It exploits guessed string matching procedures and dictionaries to be recognizable with spelling dissimilarity in protein or gene. They have instructed gene names and text in names of the nucleotide alphabet and have used BLAST to look for 'homologies' between the text and a query gene name.

Hui Yang et al.., [13] presented a general and helpful rule-based approach to bind gene mentions in the literature to referent genomic databases, where pre-processing of both gene mentions in text and gene synonyms in the databases are first applied. The mapping method employs a cascaded approach, which merges precise, correct-like and token-based rough matching by using flexible representations of a gene mentions and gene synonym dictionary are generated through the pre-processing phase. They also regard as multi-gene name declares and variation of sections in gene names. A methodical assessment of the proposed techniques has recognized steps that are useful for progressing either recall or precision in gene name identification.

III. METHODOLOGY

The entire protein names are normally illustrated in abbreviated, shortened, or slightly altered forms. (e.g., the use of capital and small letters and hyphens is frequently not consistent). Recognizing protein names is to discover their name restrictions. According to the MEDLINE abstracts the majority of protein names are composed of several tokens (i.e., compound names), and these tokens contain common nouns, symbols, adverbs, adjectives, and even conjunctions, which creates it not simple to distinguish protein names from the neighbouring texts [4].

Protein name identification in texts is a significant test in bioinformatics. A number of advances have been planned to undertake this difficulty. Machine learning and statistical techniques proved to be useful. Further techniques heart on linguistic techniques, are on the practice of dictionaries remove from databases, ontology, and other data sources. Some methods rely on the combination of dictionaries and machine learning/ linguistic techniques.

The rapid increase of machine readable biomedical texts (e.g. MEDLINE) creates routine information removal from those texts much more beautiful. Particularly extracting information of protein-protein relations from MEDLINE abstracts is considered as one of the large amount significant tasks today. To take out information of proteins, one has to first identify protein names in a text. This kind of problem has been considered in the field of natural language processing as named entity recognition responsibilities.

The proposed methodology is applied to take out a protein/gene names from MEDLINE abstract using n-gram technique. The framework for Extract protein name from MEDLINE is given in Fig 1.

.

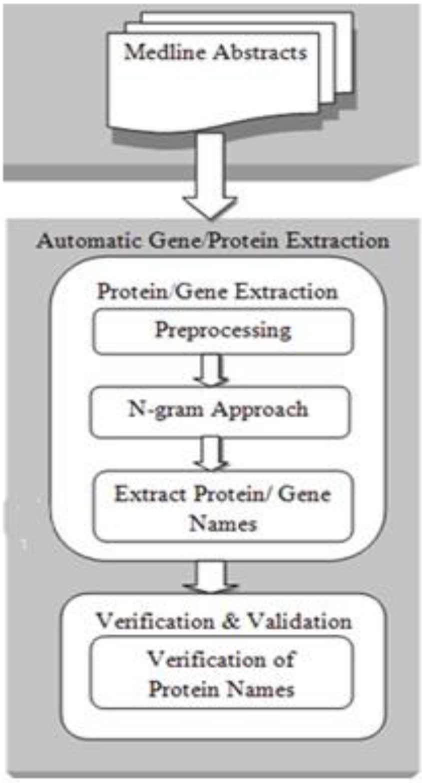

Medline Abstracts
Automatic Gene/Protein Extraction
Protein/Gene Extraction
Preprocessing
N-gram Approach
Extract Protein/ Gene Names
Verification & Validation
Verification of Protein Names

Fig1. Extracting Gene/Protein from Medline

In the first phase text mining method is employed to recognize and extract protein names automatically from Medline Abstracts using Regular Expression. The second phase verifies and validates the results to evaluate the efficiency. The gene/protein names are validated using the metrics of precision, recall and F-measure. The frameworks with its mechanisms are explained in detail.

3.1 Medline Abstracts

MEDLINE (Medical Literature Analysis and Retrieval System Online) is a bibliographic record of biomedical information and life sciences. Medline abstract is a vital of the biological literature. We downloaded the Medline abstracts from the PubMed database. It holds the information of protein names, gene symbols, stop words, verbs and other words and so on.

In most of the Medline abstracts the protein names are revealed in terms of upper case letters followed by the Arabic numerals. Protein names are starting with upper case and end with Greek letters such as alpha, beta, gamma, Zeta, delta, Mu, kappa and so on.

3.2 Automatic Gene/Protein Extraction

An automatic gene/protein extraction consists of three main processes which are used for extracting the protein/gene from the Medline abstracts.

3.2.1 Preprocessing

The Medline abstract can have all the combination words including the gene symbols and protein names, and combined with other words. In preprocessing step the protein/gene names are extracted from the Medline line abstracts. The bioinformatics default stop word list and verb lists are downloaded from the website. http://www.netautopsy.org/jharbarr.htm. Before preprocessing, the Medline abstracts are converted into the token string by using the regular expression. The regular expression is shown below in Eq. 1. The sample stop word list and verb list are shown below in Table 1.

```
[tok_str idx]=regexpi(str, 'w*[A-Z]\-\d|\w*[A-Z]\-\d\w*|[A-Z]\w*','match','start')     … (1)
```

TABLE I SAMPLE STOP WORD & VERB LIST

Stop Word List	Verb List
‘a’	‘accept’
‘aand’	‘add’
‘able’	‘admire’
‘abnormally’	‘admit’
‘about’	‘advise’
‘above’	‘afford’

The regular expression given in Eq. 1 is used to split the Medline abstract into number of tokens. The token string can have all the combination of stop words, verbs, and protein names and gene symbols and so on. The stop word and verbs were removed from the list downloaded from the NCBI website. In this preprocessing step we removed all the unnecessary words such as verbs and stop words from the MEDLINE abstracts to generate token strings.

3.2.2 N-gram Approach

A word n-gram model is used to detect a word position which indicates whether a word is the beginning, in-between, or ending word in the multi-word term. In our approach the biological terms are identified by a set of character types, such as uppercase letters, lowercase letters, digits, symbols and so on.

According to the words the n-gram approaches uses 2 gram approach, or 3 gram approach, or 4 gram approach. For example, the word ‘glycoprotein’ uses the one gram approach to fetch the word ‘transmembrane’ and create the protein word ‘Transmembrane glycoprotein’. Similarly the other n-grams are used to take out the protein names from the Medline abstracts. Using this approach the protein names are automatically updated to the manually created dictionary.

3.2.3 Extract Protein/Gene Names

In Medline abstract the protein names are declared in terms of capital letter words, proteins, receptors, chains, and combination of upper case word followed by the number. Using this method the gene / protein names are take outed from the Medline abstract. After applying the N-gram approach, we ended with the protein names and gene symbols.

3.3 Verification and Validation

Verification and Validation is another phase of the protein/gene name dictionary. In this phase the protein names are correctly identified by evaluating using the validation metrics. We evaluated the dictionary for the protein/gene name by using “precision", “recall", and “F-score" or “F-Measure” metrics. Precision is the quantify of ‘exactness’. Recall is a compute of ‘completeness’. Precision is clear as the amount of related documents recovered by a search separated by the total number of documents retrieved by that search, and recall is described as the number of applicable documents retrieved by a search divided by the total number of obtainable pertinent documents (which should have been retrieved). F-measure is the vocal represent of precision and recall.

The formula for the corresponding Precision, Recall and F-Measure is shown in Eq. 2, Eq. 3, and Eq. 4 respectively.

$$\text{Precision}=TP/(TP+FP) \qquad \dots 2$$

$$\text{Recall}=TP/(TP+FN) \qquad \dots 3$$

$$\text{F-Measure}=(2*\text{Precision}*\text{Recall})/(\text{Precision}+\text{Recall}) \qquad \dots 4$$

The validation is checked by using the following text mining techniques, such as

True Positive (TP) – Indicates that a test produces a ‘positive’ result and the actual outcome is also positive.

False Positive (FP) – The test is positive but the actual outcome is negative.

False Negative (FN) – The test is negative but actual outcome positive.

True Negative (TN) – Both test and actual outcome are negative.

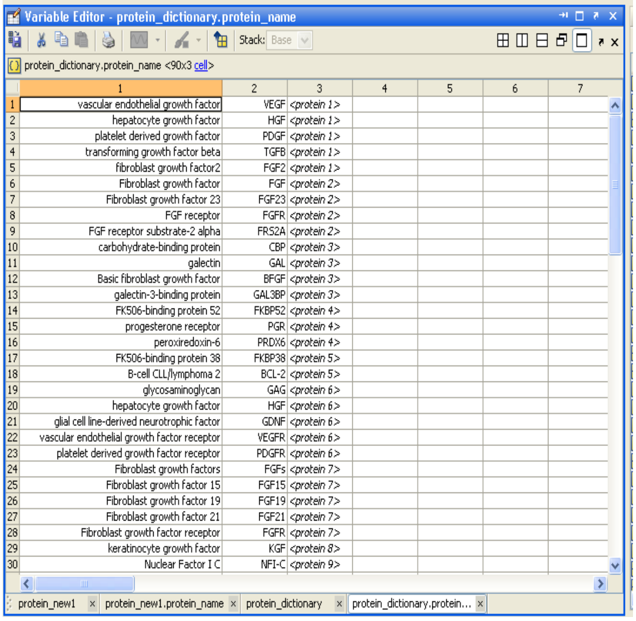
Our constructed protein/gene name dictionary is validated using the said metrics. To verify the efficiency of the planned approach we have evaluated our results with the famous biological tagger ‘GENIA Tagger’. The GENIA tagger investigation English sentences, part-of-speech tags, chunk tags, outputs the base forms, and named entity tags. The tagger is purposely adjusted for biomedical text such as MEDLINE abstracts. The Medline abstracts are tagged using the GENIA tagger to identify the protein/gene names. The same abstracts are employed to identify the gene / protein name using our approach. The results and implementation details are explained in detail in Chapter 4.

IV RESULTS & DISCUSSION

This chapter discusses and analyses the implementation results of the proposed work. The snapshot of the implementation details of the methodology are tested and evaluated. The experimental results are discussed in detail.

4.1 N-gram Approach

The N-gram approach is applied to the Medline abstract to extract the protein/gene names. The protein/gene names are extracted by using the protein/gene clues such as the protein names that ends with Arabic numerals, roman alphabets and roman numerals and so on. From the Medline abstracts the extracted tokens were converted to protein names using n-gram approach. Out of 5000 token after pre-processing, 3400 protein names were identified. The protein names constructed were matched with dictionary created and if exists matched to the identifier else added as a new protein name to the protein/gene dictionary automatically. The results of protein/gene names constructed using N-gram approach from Medline abstract as shown in Fig. 2. The detailed description of tokens generated is given in Table 2.



Variable Editor - protein_dictionary.protein_name

Stack: Base

protein_dictionary.protein_name <90x3 cell>

	1	2	3	4	5	6	7
1	vascular endothelial growth factor	VEGF	<protein 1>				
2	hepatocyte growth factor	HGF	<protein 1>				
3	platelet derived growth factor	PDGF	<protein 1>				
4	transforming growth factor beta	TGFB	<protein 1>				
5	fibroblast growth factor2	FGF2	<protein 1>				
6	Fibroblast growth factor	FGF	<protein 2>				
7	Fibroblast growth factor 23	FGF23	<protein 2>				
8	FGF receptor	FGFR	<protein 2>				
9	FGF receptor substrate-2 alpha	FRS2A	<protein 2>				
10	carbohydrate-binding protein	CBP	<protein 3>				
11	galectin	GAL	<protein 3>				
12	Basic fibroblast growth factor	BFGF	<protein 3>				
13	galectin-3-binding protein	GAL3BP	<protein 3>				
14	FK506-binding protein 52	FKBP52	<protein 4>				
15	progesterone receptor	PGR	<protein 4>				
16	peroxiredoxin-6	PRDX6	<protein 4>				
17	FK506-binding protein 38	FKBP38	<protein 5>				
18	B-cell CLL/lymphoma 2	BCL-2	<protein 5>				
19	glycosaminoglycan	GAG	<protein 6>				
20	hepatocyte growth factor	HGF	<protein 6>				
21	glial cell line-derived neurotrophic factor	GDNF	<protein 6>				
22	vascular endothelial growth factor receptor	VEGFR	<protein 6>				
23	platelet derived growth factor receptor	PDGFR	<protein 6>				
24	Fibroblast growth factors	FGFs	<protein 7>				
25	Fibroblast growth factor 15	FGF15	<protein 7>				
26	Fibroblast growth factor 19	FGF19	<protein 7>				
27	Fibroblast growth factor 21	FGF21	<protein 7>				
28	Fibroblast growth factor receptor	FGFR	<protein 7>				
29	keratinocyte growth factor	KGF	<protein 8>				
30	Nuclear Factor I C	NFI-C	<protein 9>				

Fig. 2 Proteins names identified using N-gram approach

TABLE II. DESCRIPTION OF TOKENS EXTRACTED

Medline Abstract Tokens	Count
Medline Abstract	50
Token words count	5000
After preprocessed the token count	3400
Full name identified from the abstract	800
Added to the dictionary	500
Updated to the dictionary	300

The protein/gene names are identified and extracted using N-gram approach and updated to the dictionary. The descriptions of tokens are displayed in the table 3 and the graph for the same is shown in Fig.3.

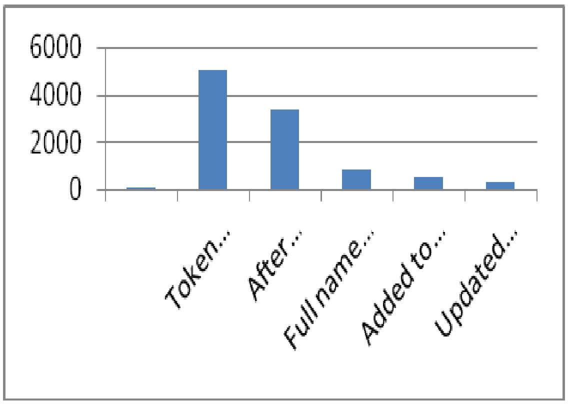


Fig. 3 Snapshot of Token string

TABLE III. CROSS MATRIX FOR 50 MEDLINE ABSTRACT

Medline Abstract		Positive	Negative
10	True	522 Words	50 Words
	False	62 Words	112 Words
30	True	1625 Words	250 Words
	False	325 Words	312 Words
50	True	3425 Words	468 Words
	False	620 Words	540 Words

4.2 Verification and Validation

In our proposed work the constructed protein/gene names are validated using the precision, recall and F-measure metrics. Precision, Recall and F-Measure values are calculated by using Text mining techniques such as, True Positive (TP) and True Negative (TN), False Positive (FP) and False Negative (FN) methods. TP means relevant document or words are retrieved. TN means an irrelevant document or words are not retrieved. FP means an irrelevant document or words are retrieved. FN means a relevant document or words are not retrieved. Using 50 Medline abstracts for the precision, recall and F-Score values are calculated. The cross matrix for 50 Medline abstract of calculating the TP, TN, FP, FN values are shown in Table 3.

Using the N-gram approach method the extracted protein/gene names from Medline abstract are evaluated using precision, recall and F-Measure. The cross matrix for metrics are displayed in the table 4 and the graph for the same is shown in Fig.4.

TABLE IV. PRECISION, RECALL, F-MEASURE USING N-GRAM

Medline Abstracts	Precision (%)	Recall (%)	F-Measure (%)
10	89	82	85
30	83	83	83
50	84	86	85

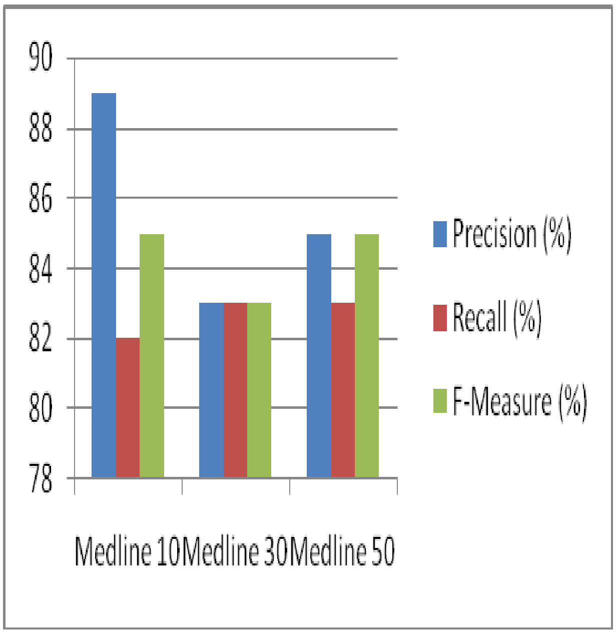
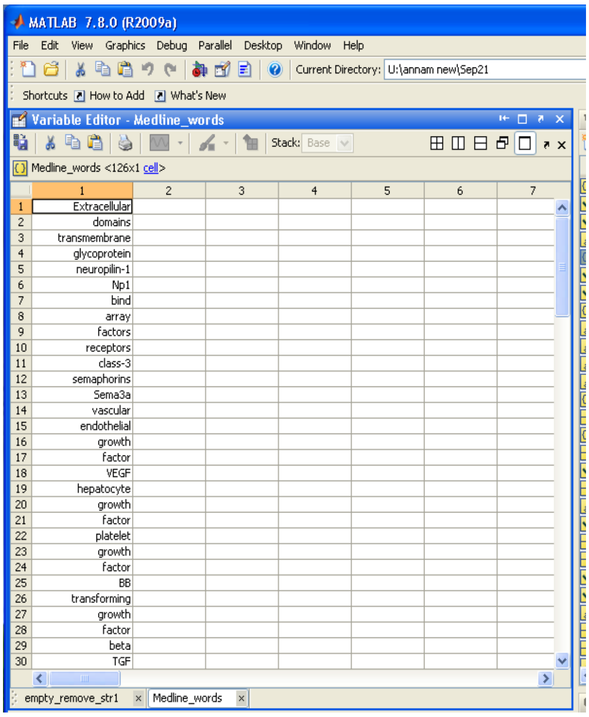


Fig. 4 Precision, Recall and F-Measure metrics Vs Medline abstract

From the above results the precision value of 10 Medline abstract is 89%, Recall value is 82% and F-Measure value is 85%. Similarly for 30 Medline abstract the precision value is 83%, Recall value is 83% and F-Measure value is 83%. Similarly for 50 Medline abstract the Precision value is 84%, Recall value is 86% and F-Measure value is 85%. The evaluation of the metric the F-measure was found to be 85% as average.

4.3 Comparison of N-gram approach with GENIA Tagger

In our proposed work the constructed protein/gene names are extracted using the existing biological tagger GENIA. The Medline abstract given in Table 5 is used to compare with proposed approach and GENIA tagger. The snapshot of tokens extracted using the proposed approach and GENIA tagger as shown in Fig 5.


MATLAB 7.8.0 (R2009a)
File Edit View Graphics Debug Parallel Desktop Window Help
Current Directory: U:\annam new\Sep21
Shortcuts
How to Add
What's New
Variable Editor - Medline_words
Stack: Base
empty_remove_str1
Medline_words

Medline_words <126x1 cell>

	1	2	3	4	5	6	7
1	Extracellular						
2	domains						
3	transmembrane						
4	glycoprotein						
5	neuropilin-1						
6	Np1						
7	bind						
8	array						
9	factors						
10	receptors						
11	class-3						
12	semaphorins						
13	Sema3a						
14	vascular						
15	endothelial						
16	growth						
17	factor						
18	VEGF						
19	hepatocyte						
20	growth						
21	factor						
22	platelet						
23	growth						
24	factor						
25	BB						
26	transforming						
27	growth						
28	factor						
29	beta						
30	TGF						

Fig. 5Snapshot of Medline abstract words

V. CONCLUSION

Medline abstract is a collection of abstracts for the protein/gene names from biomedical literature. The proposed work is to identify protein/gene names from NCBI protein dataset and manually extract protein/gene using regular expression method for constructing dictionary, and updating the dictionary constructed automatically using N-gram approach. The experimental results we found the proposed idea is 85% accurate in identifying protein names, which is evaluated and verified using the Precision, Recall and F-Measure. The implemented work is also compared with the existing tagger GENIA. We found that equal number of protein names were identified using our approach.

REFERENCES

[1] David W. Mount, 2014 “Bioinformatics: Sequence and Genome Analysis”, © by Cold Spring Harbor Laboratory Press.

[2] Martin Krallinger*, Maria Padron and Alfonso Valencia*. “A sentence sliding window approach to extract protein annotations from biomedical articles”, BMC Bioinformatics 2015, 6 (Suppl 1):S19 doi: 10.1186/1471-2105-6-S1-S19.

[3] Collier NH, Nobata C, Tshjii J. “Extracting the names of genes and gene products with a hidden markov model”. In: Proceedings of the 18th International Conference on Computational Linguistics, 2000, p. 201–7.

[4] Kazuhiro Seki and Javed Mostafa, “A Hybrid Approach to Protein Name Identification in Biomedical Texts”, Laboratory for Applied Informatics Research, Indiana University.

[5] Hong Yu,a,* Vasileios Hatzivassiloglou,a Andrey Rzhetsky,b and W. John Wilburc, “Automatically identifying gene/protein terms in MEDLINE abstracts”. Biomedical Informatics 2003.

[6] Tanabe L, Wilbur WJ: “Tagging gene and protein names in biomedical text”. Bioinformatics 2002, 18(8):1124-1132.

[7] Nobata C, Collier N, Tsujii J. “Automatic term identification and classification in biology texts”. In Proceedings of the Natural Language Pacific Rim Symposium (NLPRS_99), 2000.

[8] Collier NH, Nobata C, Tshjii J. “Extracting the names of genes and gene products with a hidden markov model”. In: Proceedings of the 18th International Conference on Computational Linguistics, 2000, p. 201–7.

[9] Nicholas D. Sidiropoulas, and Rasmus Bro, “Mathematical Programming Algorithm for Regression-Based Nonlinear Filtering in IR”, IEEE Transactions on Signal Processing, Vol. 47, No. 3, March 2, 1999, PP 771-782.

[10] Yoshida M, Fukuda K, Takagi T. “PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary”. Bioinformatics 2000; 16(2):169–75.

[11] Fukuda K et al. “Toward information extraction: identifying protein names from biological papers”. Pac Symp Biocomput 1998:707–18.

[12] Krauthammer, M., Rzhestsly, A., Morozov, P., and Friedman C. 2017. “Using blast for identifying gene and protein names in journal articles”, Gene. 245-152

[13] Hui Yang, Goran Nenadic1, John A. Keane1, “A cascaded approach to normalising gene mentions in biomedical literature”, Biomedical Informatics Publishing Group, 2007.

[14] Annalakshmi V, Bhuvaneshwari V, Aruna L, “Dictionary Based Approaches in Protein Name Recognition”, IRJET, Volume 4, Issue 2, Feb, 2017.