International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 02 | Feb -2017 www.irjet.net p-ISSN: 2395-0072

Dictionary Based Approaches in Protein Name Recognition

Annalakshmi V[1], Bhuvaneswari V[2], Aruna L[3]

[1]Assistant Professor,
Dept. of Computer Science,
Jayaraj Annapckiam College for Women (Autonomous),
Periyakulam-625 601, Tamil Nadu, India

[2]Assistant Professor
School of Computer Science and Engineering, Bharathiar University
Coimbatore-641 046, Tamil Nadu, India

[3]Assistant Professor,
Dept. of Computer Science,
Jayaraj Annapckiam College for Women (Autonomous),
Periyakulam-625 601, Tamil Nadu, India

[1]annalakshmivmca@gmail.com
[2]bhuvanesh_v@yahoo.com
[3]arunaswarni@gmail.com

Abstract—*Bioinformatics is the science of organizing and analyzing biological data. Identifying protein/gene name from Medline abstracts is an important task in the biomedical literature. Constructing the protein/gene name dictionary is a major task of the biological literature. Protein names are mentioned in terms of gene symbol, protein names, synonyms, gene name and typographical variants. Dictionary based approaches normalize gene and protein names, reducing many synonyms and phrases representing the same concept to a single identifier for that protein/gene. Protein names are identified from the dataset by using the capital letters, Arabic numerals, Roman alphabets, Roman numerals and frequent words appearing in protein names. In our work we have proposed a method to identify protein/gene name using regular expression to construct dictionary.*

Keywords— Bioinformatics, Text Mining, Gene, Protein, MEDLINE Abstract.

1.INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses [1]. Data mining techniques are the result of a long process of research and product development. Data mining is a component of a wider process called Knowledge discovery from databases.

Bioinformatics is the science of organizing and analyzing biological data that involves collecting, manipulating, analyzing, and transmitting huge quantities of data. Bioinformatics and data mining provide exciting and challenging researches in several application areas especially in computer science. Bioinformatics is the science of managing, mining and interpreting information from biological sequences and structures [2].

Text mining is the process of searching, collecting and deriving high-quality useful material from text sources. It involves setting up patterns in text files, deriving rule patterns, applying them to the text, and producing the output as meaningful information. Most information on the Web is not numeric data but text. So, text mining is a very useful technique to discover customer information from the unstructured text. Text Mining describes the automated process of analyzing natural language text with the goal of discovering information and knowledge. A number of terms describe specific aspects of automatic text analysis:

Bioinformatics is part of the larger science of computational biology. Computational biology is the application of quantitative analytical techniques in modeling and solving problems in the biological systems bioinformatics is a broad term covering the use of computer algorithms to analyze biological data. A gene is a basic unit of heredity in a living organism. It is normally a stretch of DNA (Deoxyribo Nucleic Acid) that codes for a type of protein or for an RNA (Ribo Nucleic Acid) chain that has a function in the organism. All proteins and functional RNA chains are specified by genes. Protein is a long chain molecule made up of amino acids joined by peptide bonds. Protein forms the structural material of

bodily tissues. Proteins, the principal constituents of the protoplasm of all cells, are of high molecular weight and consist essentially of combinations of amino acids in peptide linkages.

Dictionary-based approaches can also normalize gene and protein names, reducing many synonyms and phrases representing the same concept to a single identifier for that gene or protein. In addition, dictionary-based approaches can make use of the huge amount of information in curate genomics databases. Currently, there is an enormous amount of manual curation activity related to gene and protein function. Several genomics databases contain large amounts of curate gene and protein name symbols as well as full names.

Genes and proteins are usually represented by symbols and names in literature. The names usually are the long forms of their symbols and describe the functions of the genes or proteins. Therefore literature redundancy (e.g., the same genes or proteins are represented by different authors in different articles) makes it possible that may obtain automatically a relatively exhaustive gene/ protein symbol and full name table from all of MEDLINE. Protein names are mentioned in terms of gene symbol, protein names, synonyms, gene name and typographical variants. Dictionary based approaches normalize gene and protein names, reducing many synonyms and phrases representing the same concept to a single identifier for that protein/gene. Protein names are identified from the dataset by using the capital letters, Arabic numerals, Roman alphabets, Roman numerals and frequent words appearing in protein names.

In this paper we have done a detailed study on the dictionary based approach for protein/gene name identification using N-gram approach. The paper is organized as follows. Section 2 provides the literature study of the dictionary based approaches. Sections 3 describe the methodology of the study. In section 4 provides modeling of dataset for NCBI (National Center for Biotechnology Information) and Medline Abstracts. The implemented results for the dictionary based approach and N-gram approach are analyzed and validated using the text mining precision, recall and F-Measure metrics. The final section draws the conclusion of the paper.

2. METHODOLOGY

A) Objective of the Study

Protein name extraction is an important problem in the area of biomedicine and bioinformatics which is used for assessing various functional aspects of protein and gene groups. It is also used for ranking literature documents with relevance to protein and gene names. Identifying the protein gene names are available in literature based on various approaches. The idea proposed in the work is to identify protein names by constructing a dictionary manually initially and update semi automatically update using n-gram approach. The main objective of the work is:

i) Construct protein/gene dictionary manually based existing information from literature using regular expression.

a) Rules for Constructing Protein/Gene Name Dictionary

Our approach is to construct a protein/gene dictionary using dictionary based approach based on text mining technique. We manually examined and generated the regular expression for the genes and proteins, and developed a set of pattern-matching rules that map gene and protein symbols to names. The pattern-matching rules include some special abbreviations that represent Keratocan for KTN and common conventions for apply an abbreviation matches the first letter of each word in the full form (e.g. the protein symbol of Glia activating factor is GAF).

The following rules are used to construct the dictionary for the protein/gene names. The rules are obtained from the protein/gene names to create abbreviations. The following regular expression rules are:

- The abbreviation matches the first letter of each word in the full name.
- The abbreviation letter matches the first capital letter of a word in the full name.
- The abbreviation letter matches the last capital letter of a word in the full name.
- The abbreviation matches the first letter of each word in the full name with alpha or beta or gamma combinations.
- Protein names starting with upper case word followed by the Arabic numerals followed by the normal word of protein names removed the Arabic numeral from the protein/gene name. (E.g.) 'FK506 binding protein 1' name is 'FKBP1'.
- Some of the abbreviation letter uses the special abbreviation of the full name.
- Protein names starting with upper case word followed by the normal words. For example the protein/gene symbol of 'GTP binding protein' is 'GTPBP'.
- Some of the protein/gene name uses the special abbreviation of protein/gene symbol combine with the normal word of the protein/gene names. (E.g.) 'G antigen 1' name is 'GAGE1'.

B) Methodology

The proposed methodology is used to construct a protein/gene dictionary using text mining technique a

dictionary based approach. The framework for protein/gene name construction is given in Fig 1. The framework consists of the following phases. In the phase the protein names are retrieved from the protein dataset from NCBI (National Center for Biotechnology Information) to construct the protein/gene dictionary manually using regular expression (REG EXP). The framework consists of the following components which are described in detail.

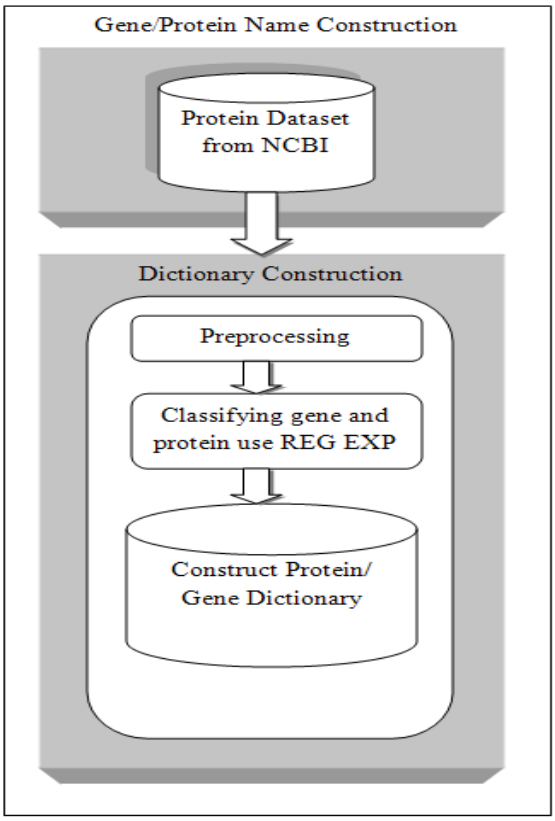

Fig. 1 Protein/Gene Name Construction

Phase I

a) Dictionary Construction

The dictionary construction consists of three main steps and processes the preprocessing step, classifying protein/gene names using regular expression process and constructing protein/gene dictionary.

i) Preprocessing

In the preprocessing phase the protein names are identified from the dataset by using the following clues extracted from literature and study.

- Capital letters (e.g., FGF, HBGF)
- Arabic numerals (e.g., GBP-2, HCF85)
- Roman alphabets (e.g., Folate receptor gamma, Hepatocyte nuclear factor 3-alpha)
- Roman numerals (e.g., dipeptidylpeptidase IV, factor XIII)
- Frequent words appearing in protein names (e.g., myelin basic protein, PI 3-kinase, nerve growth factor)

ii) Classifying gene and protein use REG EXP

The rules identified in preprocessing phase for gene and proteins names are used to form the Regular Expression (REG EXP) to create dictionary. The following regular expressions were framed to generalize the gene symbol naming conventions to construct the dictionary from the full description of protein names extracted from the dataset. The snapshot of rules for creating gene/protein using REG EXP is shown in Table 1.

Table 1 Rules for creating gene/protein use REG EXP

RULES	REG EXP
The abbreviation letter matches the first letter of each word in the full name.	'[A-Z]\w*|\d*|\[a-z]\w*|\d*$'
The abbreviation letter matches the last capital letter of a word in the full name.	'[A-Z]*'
The abbreviation matches the first letter of each word with Roman alphabets.	'[A-Z]\w*|\d*'
The abbreviation matches the first word of upper case followed by the normal word of protein name.	'[A-Z]\w*|[a-z]\w*|\d*$'
The abbreviation matches the first word of upper case followed by the Arabic numerals followed by the normal word of protein name.	'[A-Z]*|d*'
The abbreviation matches the special abbreviation of protein/gene symbol combine with the normal word of the protein name.	'[A-Z]\w*'
The abbreviation letter matches the first capital letter of a word in the full name.	'\w*[A-Z]\-\d'
Using some special abbreviation of the full name.	' [A-Z]\w*|\d*|\[a-z]|\d*$'

The above patterns are used to imparting the gene/protein dictionary using Regular Expression (REG EXP) method.

iii) Construct protein/gene dictionary

The downloaded information from NCBI gene/protein details is used to identify the gene/protein from the dataset. The framework nomenclature is used to construct the protein/gene dictionary using regular expression method. Gene names are automatically constructed by using the regular expression method. The sample snapshot for the protein name and protein/gene symbol are shown in Table 2.

Table 2 Sample protein names and symbols

Protein Names	Protein/Gene Symbol
'Fibroblast growth factor'	'FGF'
'FK506-binding protein'	'FKBP'
'Heparin-binding growth factor'	'HBGF'
'Keratinocyte growth factor'	'KGF'
'Glia-activating factor'	'GAF'
'Proto-oncogene tyrosine-protein kinase FGR'	'FGR'
'INT-2 proto-oncogene protein'	'INT-2'
'IGF receptor'	'IGF-R'
'IGF binding protein'	'IGFBP'
'Follicle-stimulating hormone receptor'	'FSHR'
'Guanylate-binding protein'	'GBP'
'Growth/differentiation factor'	'GDF'
'Growth hormone receptor'	'GHR'
'Growth hormone-releasing hormone receptor'	'GHRHR'
'Glutamate receptor, ionotropic kainate'	'KRIK'
'Glia-derived nexin'	'GDN'
'RAS guanyl-releasing protein'	RASGRP'
'Gastrin-releasing peptide receptor'	'GRPR'
'Glutathione S-transferase kappa'	'GSTK'
'Glutathione S-transferase Mu'	'GSTM'

3. RESULTS AND DISCUSSION

GO annotation site contains various data sets. Yeast data set is one of the types of data set. We can download the yeast data set form the Go

A) Implementation Results

This chapter discusses and analyzes 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. The methodology consist of 3 phases, the results of each phase are discussed in the following sections.

Phase 1 Results

i) Extract protein/gene using regular expression

The protein/gene names are constructed in the dictionary using regular expression. The regular expressions were used to generalize the protein name patterns to construct the protein/gene name dictionary.

The regular expression given in Eq. 5 is used to extract the normal form of a protein name abbreviation which will be in the form of starting with Capital letter and ending with Arabic numerals.

```
[split_str idx]=regexpi(str, '[A-Z]\w*|\d*|\[a-z]\w*|\d*$','match','start')
```
Eq.(5)

The results of constructing the protein/gene symbol from the protein/gene names using the regular expression rule 1 is shown below in Fig. 2.

Protein Names	Protein/Gene Symbol
'Fibroblast growth factor'	'FGF'
'Heparin-binding growth factor'	'HBGF'
'Keratinocyte growth factor'	'KGF'
'Glia-activating factor'	'GAF'
'Follicle-stimulating hormone receptor'	'FSHR'
'Guanylate-binding protein'	'GBP'
'Growth/differentiation factor'	'GDF'
'Growth hormone receptor'	'GHR'
'Growth hormone-releasing hormone receptor'	'GHRHR'
'Glutamate receptor, ionotropic kainate'	'KRIK'
'Glia-derived nexin'	'GDN'
'Gastrin-releasing peptide receptor'	'GRPR'
'Glutathione S-transferase kappa'	'GSTK'

Fig. 2 Results of Using Regular Expression Rule 1

The regular expression is given in Eq. 6 is used to extract the protein name abbreviation which will be in the form of ending with Capital letter word.

```
[split_str idx]=regexpi(str, '[A-Z]*','match','start')
```
Eq.(6)

The results of constructing the protein/gene symbol from the protein/gene names using the regular expression rule 2 is shown below in Fig. 3.

Protein Name	Protein/Gene Symbol
'Proto-oncogene tyrosine-protein kinase FGR'	'FGR'
'Pre-mRNA 3'''-end-processing factor FIP1'	'FIP1'
'Peptidyl-prolyl cis-trans isomerase KFBP1A'	'KFBP1A'
'Peptidyl-prolyl cis-trans isomerase KFBP1B'	'KFBP1B'
'Pre-mRNA-splicing regulator WTAP'	'WTAP'
'Zinc finger protein ZFPM1'	'ZFPM1'
'Zinc finger protein ZFPM2'	'ZFPM2'
'ARF GTPase-activating protein GIT1'	'GIT1'
'ARF GTPase-activating protein GIT2'	'GIT2'
'Poly(A) RNA polymerase GLD2'	'GLD2'
'Nucleoporin GLE1'	'GLE1'

Fig. 3 Results of Using Regular Expression Rule 2

The sample snapshot of table 7 shows the number of protein/gene names use various regular expression rules.

Table 3 Number of count for various REG EXP rules

Regular Expression Rules	Number of Count
Regular Expression Rule 1	3500
Regular Expression Rule 2	500
Regular Expression Rule 3	750
Regular Expression Rule 4	1250
Regular Expression Rule 5	1300
Regular Expression Rule 6	300
Regular Expression Rule 7	200
Regular Expression Rule 8	220

The pictorial representaion of the Table 7 as shown in Fig. 4.

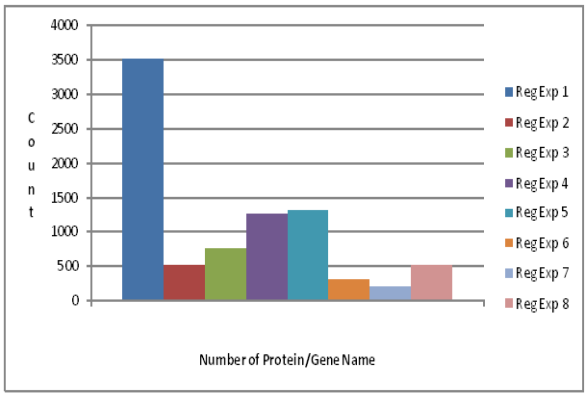

Fig. 4 Protein/gene name count use REG EXP

ii) Construct Protein/Gene name Dictionary Using Regular Expression

Using regular expression method to manually construct the protein/gene name dictionry. In our dicitionary the protein/gene symbol are calculated based on the protein/gene names and as well as the alternate names of the protein names of protein/gene symbols are added to the dictionary. The results are stored into the Matlab structure. The Fig. 11 shows the sample snapshot for the overall protein/gene name dictionry using regular expression as shown in below.

Protein Names	Alternate Names	Protein/Gene Symbol
'Fibroblast growth factor'	<6x2 cell>	'FGF'
'Heparin-binding growth factor'	<1x2 cell>	'HBGF'
'Keratinocyte growth factor'	<2x3 cell>	'KGF'
'Glia-activating factor'	<2x3 cell>	'GAF'
'Follicle-stimulating hormone receptor'	'Follitropin receptor'	'FSHR'
'Guanylate-binding protein'	<1x2 cell>	'GBP'
'Growth/differentiation factor'	<1x2 cell>	'GDF'
'Growth hormone receptor'	<2x2 cell>	'GHR'
'Growth hormone-releasing hormone receptor'	<1x2 cell>	'GHRHR'
'Glutamate receptor, ionotropic kainate'	<1x2 cell>	'KRIK'
'Glia-derived nexin'	<3x2 cell>	'GDN'
'Gastrin-releasing peptide receptor'	'GRP-preferring bombesin receptor'	'GRPR'
'Glutathione S-transferase kappa'	<1x2 cell>	'GSTK'
'Glutathione S-transferase Mu'	<1x2 cell>	'GSTM'
'Glycerol kinase'	<1x2 cell>	'GK'
'Host cell factor'	<1x2 cell>	'HCF'
'Histone deacetylase'	<1x2 cell>	'HD'
'Hepatoma-derived growth factor'	<1x3 cell>	'HDGF'
'Hepatoma-derived growth factor-related protein'	<1x2 cell>	'HDGFRP'
'Hepatocyte growth factor'	<2x2 cell>	'HGF'
'Hypoxia-inducible factor'	<4x2 cell>	'HIF'
'Hydroxy indole O-methyl transferase'	<1x2 cell>	'HIOMT'
'Hydroxy-acid oxidase'	<1x2 cell>	'HAO'
'Huntingtin-associated protein'	<1x2 cell>	'HAP'
'Homeodomain-interacting protein kinase'	<1x2 cell>	'HIPK'
'Hepatocyte nuclear factor'	<4x2 cell>	'HNF'

Fig. 5 Overall protein/gene name dictionary

4. CONCLUSION

In this paper we have done a detailed study of the dictionary based approach for the identifying protein/gene name using regular expression method. The modeling of the NCBI protein dataset and the Medline Abstracts for the extraction of protein/gene are provided in detail. The results on implementation we have found that gene/protein names are mentioned in terms of capital letters, Arabic numerals, roman alphabets, roman numerals etc. The extracted protein/gene names are updated into the dictionary. We conclude from our study that the regular expression method is more efficient for extracting gene/protein names from Medline abstracts.

REFERENCES

[1] Arun. K. Pujari, *Data Mining Techniques*, Universities press (India) Limited 2001, ISBN-81-7371-380-4.
[2] Agarwal .R, Imielinski .T, Swami .A, "Mining Association rules between set of items in large databases", Proc. 1993 ACM SIGMOD Int. Conf. On Management of Data, Washington, DC: ACM Press, pp 207-216.
[3] Koning D, Sarkar IN, Moritz T: TaxonGrab: "Extracting taxonomic names from text ", Biodiversity Informatics 2006, 2:79-82.
[4] Fukuda K et al. "Toward information extraction: identifying protein names from biological papers". Pac Symp Biocomput 1998:707–18.
[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] Jaiwei Han, Michelinne Kamber, "Data Mining: Concepts and Techniques", 2001, II Edition.
[7] Tanabe L, Wilbur WJ: "Tagging gene and protein names in biomedical text". *Bioinformatics* 2002, 18(8):1124-1132.
[8] Katrin Fundel*, Daniel Güttler, Ralf Zimmer and Joannis: Apostolakis, "A simple approach for protein name identification: prospects and limits". BMC bioinformatics 2005.
[9] Hanisch D, Fluck J, Mevissen H, Zimmer R: "Playing Biology's Name Game: Identifying Protein Names in Scientific Text". *Pacific Symposium on Biocomputing* 2003, 8:403-414.
[10] Hanisch D, Fundel K, Mevissen H, Zimmer R, Fluck J: ProMiner: "Rule-based protein and gene entity recognition". *BMC Bionformatics* 2005, 6(Suppl 1):S14.
[11] Hirschman L, Morgan AA, Yeh AS: "Rutabaga by any other name: extracting biological names". *Journal of Biomedical Informatics* 2002, 35(4):247-259.
[12] Hong Yu,a,* Vasileios Hatzivassiloglou,a Andrey Rzhetsky,b and W. John Wilburc, "Automatically identifying gene/protein terms in MEDLINE abstracts". Biomedical Informatics 2003.
[13] Kazuhiro Seki and Javed Mostafa, "A Hybrid Approach to Protein Name Identification in Biomedical Texts", Laboratory for Applied Informatics Research, Indiana University.
[14] Tanabe L, Wilbur WJ: "Tagging gene and protein names in biomedical text". *Bioinformatics* 2002, 18(8):1124-1132.
[15] Tsuruoka Y, Tsujii J: "Boosting Precision and Recall of Dictionary-Based Protein Name Recognition". *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine* 2003:41-48.
[16] Ono,. T., Hishigaki, H., Tanigami, A., and Takagi, T.2001. "Automated extraction of information on protein-protein interactions from the biological literature", Bioinformatics 17:155-161.
[17] Yoshida M, Fukuda K, Takagi T. "PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary". Bioinformatics 2000; 16(2):169–75.
[18] Yu H, Hatzivassiloglou V, Friedman C, Rzhetsky A, Wilbur WJ: "Automatic extraction of gene and protein synonyms from MEDLINE and journal articles". *Proc AMIA Symp* 2002:919-923.
[19] Jenssen T, Lagreid A, Komorowski J, Hovig E, "A literature network of human genes for high-throughput analysis of gene expression". *Nature Genetics* 2001, 28:21.
[20] Kazuhiro Seki and Javed Mostafa, "A Hybrid Approach to Protein Name Identification in Biomedical Texts", Laboratory for Applied Informatics Research, Indiana University.