ELICITATION OF PROTEIN – PROTEIN INTERACTIONS FROM BIOMEDICAL LITERATURE USING ASSOCIATION RULE DISCOVERY

Jarvie John Samuel

Thesis Prepared for the Degree of

MASTER OF SCIENCE

UNIVERSITY OF NORTH TEXAS

August 2010

APPROVED:

Xiaohui Yuan, Major Professor
Miguel Ruiz, Committee Member
Qunfeng Dong, Committee Member
Bill Buckles, Departmental Graduate Coordinator
Ian Parberry, Chair of the Department of Computer Science and Engineering
Costas Tsatsoulis, Dean of the College of Engineering
James D. Meernik, Acting Dean of the Robert B. Toulouse School of Graduate Studies

Samuel, Jarvie John. Elicitation of protein-protein interactions from biomedical literature using association rule discovery. Master of Science (Computer Science), August 2010, 82 pp., 22 tables, 22 illustrations, bibliography, 131 titles.

Extracting information from a stack of data is a tedious task and the scenario is no different in proteomics. Volumes of research papers are published about study of various proteins in several species, their interactions with other proteins and identification of protein(s) as possible biomarker in causing diseases. It is a challenging task for biologists to keep track of these developments manually by reading through the literatures. Several tools have been developed by computer linguists to assist identification, extraction and hypotheses generation of proteins and protein-protein interactions from biomedical publications and protein databases. However, they are confronted with the challenges of term variation, term ambiguity, access only to abstracts and inconsistencies in time-consuming manual curation of protein and protein-protein interaction repositories. This work attempts to attenuate the challenges by extracting protein-protein interactions in humans and elicit possible interactions using associative rule mining on full text, abstracts and captions from figures available from publicly available biomedical literature databases. Two such databases are used in our study: Directory of Open Access Journals (DOAJ) and PubMed Central (PMC). A corpus is built using articles based on search terms. A dataset of more than 38,000 protein-protein interactions from the Human Protein Reference Database (HPRD) is cross-referenced to validate discovered interactive pairs. A set of an optimal size of possible binary protein-protein interactions is generated to be made available for clinician or biological validation. A significant change in the number of new associations was found by altering the thresholds for support and confidence metrics. This study narrows down

the limitations for biologists in keeping pace with discovery of protein-protein interactions via manually reading the literature and their needs to validate each and every possible interaction.

ACKNOWLEDGEMENTS

I would like to thank my committee, Dr. Yuan, Dr. Ruiz and Dr. Dong, for their great guidance and constant support. In particular, I extend my sincere gratitude to my major professor, Dr. Yuan, without whose painstaking efforts, this research work would not have known the fruits it bears now. Thanks for the continued confidence in me throughout developing the various modules in research, his openness to ideas and his great patience in helping me consolidate the ideas and providing reviews of the drafts prepared at different stages of my research. I would also like to thank my peers in our Computer Vision and Intelligence Lab. I would like to thank Dr. Allen Clark, and all staff members at Office of Institutional Research, who supported me from the very first day at UNT. I would like to thank my parents and brothers for providing me the constant support and courage to pursue my research goals. I would like to thank my wife Lovely, who has been my greatest support and motivation.

Last but not least, my humble gratitude to God Almighty for His grace and blessings. I thank all those who prayed and wished me all success.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS iii

LIST OF TABLES vii

LIST OF FIGURES viii

Chapters

1. INTRODUCTION 1
   - 1.1 Background 1
   - 1.2 Genes, Proteins and Biomarker Discovery 4
   - 1.3 Information Retrieval 5
     - 1.3.1 Boolean Model 5
     - 1.3.2 Vector Space Model 5
     - 1.3.3 Information Retrieval Systems 5
   - 1.4 Information Extraction 7
     - 1.4.1 Entity Recognition 7
     - 1.4.2 Relationship Extraction 9
   - 1.5 Text Mining 10
     - 1.5.1 Importance of Text Mining 11
     - 1.5.2 Types, Inferences and Evaluations 11
     - 1.5.3 Association Rule Mining 12

2. RELATED WORK 14
   - 2.1 Information Retrieval 14
   - 2.2 Entity Recognition 15
     - 2.2.1 Synonyms and Abbreviations 16
   - 2.3 Relationship Extraction 16
   - 2.4 Hypotheses Generation 18
   - 2.5 Integration Framework 19
   - 2.6 Summary 20

3. METHODOLOGY 21
3.1 Overview 21
3.2 Building the Corpus 22
3.2.1 Web-based Search Interface 23
3.2.2 C-Engine Crawler 23
3.2.3 Extraction of Text and Images 24
3.3 Image Analysis 25
3.4 Identification of Protein Names 25
3.4.1 POS Tagging 26
3.4.1.1 Lingua::EN::Tagger 27
3.4.1.2 Methods Used 29
3.4.2 Processing Non-simple Sentences 29
3.4.3 Recognizing Entities 30
3.4.3.1 Identification by Word Appearance and Morphology 30
3.4.3.2 Dictionary of Proteins 31
3.4.3.3 Elimination by Comparison 31
3.5 Extraction of Relationship between Entities 31
3.6 Discovering Novel Associations using Text Mining 33
3.7 Generating Hypotheses - Strategy 35
4. EXPERIMENTS AND DISCUSSION 36
4.1 Experiments 36
4.2 Evaluations and Observations 45
4.2.1 Evaluating Explicit Associations 46
4.2.2 Evaluating Associations from Abstracts 48
4.2.3 Evaluating Implicit Associations 49
4.2.4 Visualization 51
4.3 Discussions 64
5. CONCLUSION AND FUTURE WORK 70
5.1 Conclusion 70
5.2 Future Work 71
BIBILIOGRAPHY 73

LIST OF TABLES

Page

Table 1.1 List of information retrieval systems 6

Table 1.2 List of information extraction systems 10

Table 2.1 List of text mining tools 19

Table 2.2 List of systems that employ integrated framework approach 20

Table 3.1 Types of POS tagging classes 26

Table 3.2 List of POS tagset used in Lingua::EN::Tagger 29

Table 3.3 Partial list of frequently used functional verbs 32

Table 4.1 List of query terms used for document retrieval from PMC and DOAJ 38

Table 4.2 Number of articles retrieved from PMC and DOAJ for each query term 38

Table 4.3 List of explicit protein-protein interactions for query term “ERBB2 breast cancer” 42

Table 4.4 A view of a transaction file 43

Table 4.5 Number of unknown associations for each query for articles retrieved from PMC 44

Table 4.6 Number of unknown associations for each query for articles retrieved from DOAJ 45

Table 4.7 A view of protein interaction pairs 45

Table 4.8 Number of explicit associations – extracted and matched 47

Table 4.9 Time taken for identification of protein names 48

Table 4.10 Comparison using abstracts and full texts 48

Table 4.11 Number of matched and novel associations between proteins from PMC - abstracts 49

Table 4.12 Number of matched and novel associations between proteins from PMC 50

Table 4.13 Number of matched and novel associations between proteins from DOAJ 51

Table 4.14 Decrease in number of unknown associations when minsup changes 67

Table 4.15 Decrease in number of unknown associations when minconf changes 68

LIST OF FIGURES

Page

Figure 3.1 System overview 21

Figure 3.2 Flow chart of processing steps involved for discovery of novel protein interactions ..22

Figure 4.1 Screenshot of C-Engine search portal 36

Figure 4.2 Screenshot of articles retrieved for a query term.....37

Figure 4.3 Application of POS tagging on parts of text article 40

Figure 4.4 Sliced view of the protein dictionary containing protein names and their synonyms..41

Figure 4.5 Network of explicit protein-protein interactions for query – ERBB2 breast cancer....52

Figure 4.6 Another visualization of explicit protein interactions for ERBB2 breast cancer 53

Figure 4.7 Network of explicit interactions with limited degree of freedom 54

Figure 4.8 Network of ERBB2 and HER2 – case of synonyms 55

Figure 4.9 Another visualization of ERBB2 and HER2 network – case of synonyms 56

Figure 4.10 Grid-network of explicit protein-protein interactions 57

Figure 4.11 Grid-network with reduced degree of freedom 58

Figure 4.12 Grid-network with only novel associations.....59

Figure 4.13 Grid-network – tracing explicit associations 1.....60

Figure 4.14 Grid-network – tracing explicit associations 2.....61

Figure 4.15 Grid-network – tracing explicit associations 3.....62

Figure 4.16 Extracting network of interest 1 63

Figure 4.17 Extracting network of interest 2 63

Figure 4.18 Support-confidence-association rule plots for different query terms (PMC).....65

Figure 4.19 Support-confidence-association rule plots for different query terms (DOAJ) 66

Figure 4.20 Number of unknown associations for general query against specific query term…..67

CHAPTER 1

INTRODUCTION

1.1 Background

Extracting information from a stack of data is a tedious task and is an issue that researchers face in the field of proteomics, which is the study of protein structures, functions and establishment of interactions among proteins. Volumes of published research papers document the progress and development about various genes and gene products in several species. This has been possible with the advent of an array of biological and computational techniques including multidimensional protein identification technology (MudPIT), protein microarray technology, and mass spectrometry [1], protein chips, and two-hybrid systems [2]. It is possible now to map the entire genome of a species within a time range from weeks to months [3]. Not only do the experiments produce a large result dataset, but they also contribute in consuming a large amount of time for biologists to manually sift through the data to identify interesting information.

Biologists are in search of identifying the gene expression patterns, such as which genes are expressed or suppressed during onset of a specific disease or in specific cell growth cycle, finding out which pair of genes or proteins interacts during biological processes to better identify possible biomarkers for diseases and drug discovery.

Apart from publishing results in papers, there has been a steady pace in creating biomedical databases, capturing protein-protein interactions and even genome data of several species such as the mustard weed (*Arabidopsis thaliana*) [4], yeast (*Saccharomyces cerevisiae*) [5] and human [4,6]. Some inherent challenges with these large biomedical databases are that they need to be:

- Scalable – Manual curation cannot be at par with the rate at which the results are

being generated

- Evolution – Results or associations documented would cease to be complete and consistent with the passage of time
- Annotator agreement – It is highly unlikely that two curators agree on manually curated results, notwithstanding the vast nomenclature formats used by biologists themselves [7]

It is a challenge in itself for a biologist to keep abreast of the developments by reading through the literature or querying for results from a database.

A simple search on Google Scholar using the search term “protein-protein interactions” gives more than 900,000 articles excluding patents. A more specific search for a protein associated with breast cancer, HER-2 (human epidermal growth factor receptor 2), returns a search result of more than 135,000 articles on Google Scholar. A synonym of the above protein, ErbB-2, retrieves more than 50,000 articles. The sheer number of publications limits the effective use to extract meaningful information. Hence, there has been a growing relevance on literature-mining tools by researchers. Several tools have been developed by computer linguists to assist in the identification, extraction, and hypotheses generation of proteins and protein-protein interactions from biomedical literature and protein databases.

However, in most cases, access is limited to only abstracts of these papers. Additionally, there are innate differences between conventional text mining methods and those when applied to biomedical literature – where authors use scientific jargon, non-standard terms and structures, and differences in naming conventions followed to refer a gene or gene product [8]. Employing text mining methods on full text articles helps identifying a broader list of associations that researchers wouldn’t gather from 20 lines of abstracts alone. Moreover, using full text articles

make it possible to extract secondary relations or supporting data associations mentioned in the article that abstracts might not include. The importance of including such supporting data lies in the fact that they may be keys for bridging two sub-classes of research that scientists might generally overlook. The challenges are further aggravated by the fact that rigorous efforts are needed from non-specialist users to integrate existing tools for clinical evaluation to achieve standardized formats or even machine compatibility [9].

As discussed in [130], extraction of protein-protein interaction from literature is the one of the most studied research areas in biomedical text mining. One of the significant aspects of this research is to address the aforementioned challenges by extracting protein-protein interaction and elicit possible interactions using associative rule mining of full text articles available publicly from biomedical literature databases such as Directory of Open Access Journals (DOAJ) and PubMed Central (PMC). A comprehensive compilation of possible binary protein-protein interactions generated using apriori algorithm for association rule mining is presented. A better realization of the methods is made possible by using captions from figures that are very specific to elaborate relations or definitions. Another aspect of this research is to validate the possible associations against a publicly available dataset of protein-protein interactive pairs. Known interactions are removed and a new set of interactive pairs are elicited, which are later made available for biological validation. In other words, the goal is not only identification of all possible implicit associations, but also to generate a set of optimal size of unknown associations that would be manageable for biological validation. This research presents a simple but efficient way of narrowing down the limitations for biologists in keeping pace with the discovery of protein interactions and their need to validate each and every possible interaction found in research articles or datasets.

The rest of this chapter provides a brief overview of some of the key steps in text mining. Section 1.2 gives an overview of the biological entities and biomarker discovery in biomedical parlance. Section 1.3 elaborates on related information retrieval methods. Once the articles are retrieved, the next crucial step is to identify all the instances of an entity that are of interest, which will aid in extracting relationships and other information related with the entity. Section 1.4 introduces information extraction using named entity recognition and relationship extraction. Section 1.5 offers a brief discussion about text mining, in particular association rule mining.

1.2 Genes, Proteins and Biomarker Discovery

The gene, as we know, is “the basic unit of heredity in a living organism” [10]. And gene products are the resultants of gene expression in the form of RNA or proteins. “How much of a gene product” is used to determine the nature of a gene? A biomarker is an indicator used to measure the presence or stage of a disease. Discovering biomarkers is essential to understanding the effectiveness of a drug therapy for a specific disease treated as proteins are mostly affected by diseases [11]. Disease-based biomarkers include oncogenes such as p53 for cancer, LDL for hypertension and cholesterol for heart disease [12]. Like many other systems, biological systems are not devoid of interactions; in particular, we are interested in protein interactions, which determine role in biological processes including cell signaling, diseases (e.g. cancer), and mutative action on another protein. Methods such as mass spectroscopy and immuno-histochemical staining [1] are used to discover novel biomarkers in a clinical environment. With the availability of computational linguistic tools, much research has focused on biomarker discovery using text mining from online sources of publications [99]. This research translates gene and gene products as entities that need to be identified and extracted from the biomedical corpus generated from DOAJ and PMC.

1.3 Information Retrieval

Information retrieval (IR), the first step in any literature mining process, involves retrieving relevant documents from a database based on search terms or keywords provided by the user. In most cases, the retrieved documents are ranked in order of relevance to the query term(s). Mostly successful methodologies to implement an IR are the Boolean model and vector space representation model.

1.3.1 Boolean Model

Boolean model is a simple IR-model based on set theory, in which queries are specified as Boolean expressions, such as ‘AND’ or ‘OR’, and thus allow the user to retrieve relevant documents pertaining to the combination of query terms specified. However, the Boolean model does not support partial matching of terms, and only straight-forward queries can be formulated by users using logical expressions. Another disadvantage is that there is no ranking of retrieved documents, resulting in either too few or too many documents.

1.3.2 Vector Space Model

In the vector space model, each document is represented as a term vector. Binary or non-binary weights are assigned to these term vectors. Non-binary weights accommodate for partial matching of terms too. The weights are determined using term frequency (TF) within a document as well as the inverse document frequency (IDF). Thus, in this vector space, both queries and documents are represented as weighted vectors and are used to compute the degree of similarity between the document and the query. Cosine similarity measure is a common metric used to compute similarity between documents.

1.3.3 Information Retrieval Systems

Ad hoc information retrieval systems include general-purpose search engines such as

Google and AltaVista, and biomedical systems (domain-specific) such as PubMed. PubMed system of the National Library of Medicine was the first major resource for online biomedical publications [13]. It comprises more than 19 million citations for biomedical articles from MEDLINE and life science journals. A list of information retrieval systems is summarized below in Table 1.1.

TABLE 1.1

LIST OF INFORMATION RETRIEVAL SYSTEMS

IR system	Description
EBIMed [100]	retrieves abstracts from Medline
GoPubMed [43]	knowledge based search engine for biomedical literature [101]
Google Scholar [102]	search portal for scholarly literature; ranked results based on term weights, publisher date, author and number of citations
CrossRef [47]	links citations across publishers using DOI
XplorMed [46]	retrieves association between words from an input set of Medline abstracts

This research utilizes the publicly available archives of life science journals such as PubMed Central (PMC) and Directory of Open Access Journals (DOAJ).

The PMC system is maintained by National Institutes of Health's (NIH) National Center for Biotechnology Information (NCBI). PMC allows complete and free access to the full text of the journal articles using a search interface and through "E-Utilities." The latter is a collection of programming utilities that allows retrieval of articles including abstracts, full texts, and citations to a query term in user requested formats, such as SGML and XML. One major reason to use PMC is its role in aggregating data from diverse sources and storing the information in one retrievable format [14]. An inherent problem with biomedical domain such as proteomics is the large set of synonyms – other common words used to describe the same concept – for biological entities such as genes and proteins. This might result in missing key or relevant publications if

query term disregards the related terms or synonyms. PubMed system handles this problem by query expansion of the search terms using biological thesauri, stemmed tokens and list of synonyms and abbreviations [9].

DOAJ was created with the goal towards increased user access to publications in various "scientific and scholarly journals." DOAJ currently has a repository of more than 350,000 articles on various sciences. It encompasses literature even from non-English journals. However, the scope of this research focuses only on biomedical literature in English. Almost all the publications are available in both pdf and html format. This research uses the articles in html format for text processing, while the pdf articles are downloaded for future use. DOAJ comprises about 4868 journals, out of which 182 journals are related specifically to biology and life sciences [15].

1.4 Information Extraction

Information extraction (IE) is one of the vital steps in the text mining process. Simply stated, it involves two steps: first, identification of the entities of interest; and secondly, the extraction of relationship between a pair of entities of given type.

1.4.1 Entity Recognition

It is a concept borrowed from natural language processing (NLP), where it is known as "semantic tagging" or "named entity recognition" (NER). The objective is to recognize all the "instances of a name for a specific type of entity," where the entities are names of genes and proteins. Instances of an entity name would mean synonyms, abbreviations, expansions and other gene or protein names for that particular entity in question. The principal motivation for entity recognition is that it forms the first step to identify the gene and protein names considering all its variants and being able to represent them in a standard format to extract relationships from text

corpus [17]. It is known that there is an “average of 5.5 different names for every gene of a human being” [7].

Entity recognition is a mapping mechanism of biological concepts and terms in the text. Genes and proteins can be mentioned in the text articles either as full names (*human epidermal growth factor receptor 2*) or as acronyms (ErbB-2) [23]. This identification is challenged by the absence of a complete dictionary for most types of genes and proteins. Effectiveness of a simple or approximate word matching to detect typographical variations, such as Erbb2 or ERBB-2, varies across entities [17]. Although there are standard naming conventions for gene and protein nomenclature, authors tend to use deviant patterns to highlight a particular gene or protein function [7]. For example, in [21], authors identify yeast genes “SRC1 and YDR458C as HEH1 and HEH2 to indicate the helix-extension-helix structure” of the yeast.

With increasing research catering only to identification of entities, several dictionaries have been compiled with almost all the known variations and abbreviations for genes and proteins. These dictionaries are mapped to the ontological concepts, which help to identify the relevance of the term from text. Examples of such ontologies are GO [18], HUGO [19] and UMLS [20]. They provide downloadable versions of gene and protein names along with unique identification numbers available in different formats. Most of these databases or files are updated on a regular basis to add, delete or update the concepts and descriptions about entities.

Another challenge in entity recognition is term variation. Term variation stands for expression of a term or concept in number of ways. Several biological entities have different names or synonyms. For example, ErbB-2 and HER2 indicate the same biomarker protein associated with breast cancer. About 33% of the term occurrences are variants, which mean there are many terms that indicate the same entity [22].

Several strategies to recognize entities of interest include lexicon-based methods, manual or automatic rule-based methods [17], machine learning techniques [2] or a combination of them. So, the result of entity recognition is a set of tags assigned to each term indicating the type or nature of entity, similar to the concept of part-of-speech (POS) tagging. An elaborate discussion about the methods used for entity recognition is mentioned in Section 2.2 of Chapter 2.

1.4.2 Relationship Extraction

This second step in information extraction deals with extracting the relationship between a pair of entities. In other words, relationship extraction methods identify structures from the text that contain the biological entities or terms using tagging, concept sets and pattern templates devised manually or automatically [2]. The nature of relationships can be very general, such as biochemical association of CD45 with the T cell receptor, or very precise associations, such as regulatory relationships (e.g. regulatory function of Ndd1, a cell-cycle regulator, on Mcm21, a kinetochore protein, during normal cell growth [24]).

One focus of relationship extraction is the detection of protein-protein interaction from literature. The extracted information is usually validated by the biologist in task. In most cases, the information is manually curated before being stored in a gene-protein interaction (GPI) database. Other examples of types of relationships include molecular interactions of proteins, conceptual relationship among diseases, genes and ontology terms, dedicated associations with protein phosphorylation and so on [9]. Different strategies are employed to extract relationships entities and are discussed in Section 2.3 of Chapter 2.

A list of information extraction systems is summarized in the following Table 1.2.

TABLE 1.2

LIST OF INFORMATION EXTRACTION SYSTEMS

IR system	Description
iProLink [103]	uses protein name dictionaries, protein ontology and tagged corpora
PubGene [104]	search Medline abstracts; identifies interactions using co-citation
ABNER [105]	statistics based machine learning named entity recognition system
GAPSCORE [54]	computes a numeric score to assign entity name to a term
AliasServer [106]	handles multiple aliases that are used to identify proteins
Abbreviation Server [107]	handles multiple abbreviations to identify genes or protein names

1.5 Text Mining

The term “text mining” has become one of the most ubiquitous term in the field of biomedical natural language processing (BioNLP). Text mining differs from data mining where the former finds associations from unstructured data such as text while the latter discovers interactions from structured databases such as retail warehouses [35]. A straightforward definition would be “the discovery of new, hitherto unknown information by automatically extracting associations from diverse sources” [25]. It also is commonly known as “knowledge discovery” especially in the field of data mining where hidden information about associations among the itemsets in a transaction database is crucial. One application of data mining involves retail store transactions. In the field of BioNLP, text mining also is referred to as “hypotheses generation.” The name comes from the objective of text mining to infer indirect relationships from the already known fact, i.e. to generate a hypothesis. One of the most popular applications of text mining in BioNLP is discovery of novel protein-protein interactions.

1.5.1 Importance of Text Mining

Text mining is different from information extraction (IE). The latter is only able to extract relationships from text that already has been identified or highlight relations already present in the text [2]. Text mining, on the other hand, use the explicit associations extracted from multiple sources to infer previously anonymous relationships that are worthwhile to investigate [17].

The proliferation of biomedical literature has resulted in researchers keeping track of only a very small research publication and thus, not being even aware of the established results from "bibliographically disjointed" fields (topics of interest belonging to two sub areas of research domain) [26] to make logical deductions about an inferred association. Contemporary research involves cross-over of disciplines that were previously never thought to have any associations. There is a high probability that researchers are totally ignorant of the facts from either field [9]. The growing urge of researchers to keep pace with latest trends and discoveries in their fields of research and the immediate needs to make sense of the piles of data from high-throughput experiments compel the importance of text mining [23].

1.5.2 Types, Inferences and Evaluations

Text mining began out of the idea first postulated by Don Swanson called "complementary structures in disjoint literature (CSD)" [27]. This powerful model uses a very simple model to discover the hidden relationships from known facts. Swanson proposed "A interacts with B, and B interacts with C, therefore A may interact with C" which is now famously known as Swanson's ABC model [28]. Using this model, he found out previously unknown relationship between fish oil and Raynaud's syndrome [29] and role of magnesium deficiency for migraines [30] long before both these associations were clinically verified.

Text mining can be classified as using “closed” or “open” framework. In “closed” framework, a hypothesis is initially provided by the user and then the literature is mined to support this hypothesis. While, in “open” discovery problem, possible hitherto credible and undiscovered associations are revealed for a particular entity of interest [9].

Most of the novel relationships discovered might be too superficial or already known to be published explicitly, as demonstrated by text mining results on study of yeast interactions. This supports the argument to access full text articles than abstracts or citations alone, to know whether the implicit association has already been published or not. Another previously stated argument for using full text articles is that it helps discover associations even from “bibliographically disjointed” areas [9, 26].

Generated hypotheses need to be evaluated either manually or by automatic methods. Although time consuming, manually reviewing the literature is conducted to evaluate the significance and validity of the novel association [17]. Trivial associations can be eliminated by validating the results against known data sources encompassing similar types of associations or integrating databases that are preferably manually curated [9].

1.5.3 Association Rule Mining

This research presents the implementation of an association rule mining method for text mining to detect association among genes and proteins. It is one of the most-popular methods to discover relations between various entities in a database. The problem statement can be defined as: “Given a set of transactions T, in a database D, where $T \in D$, find all rules that will predict an item’s occurrence based on presence or absence of other item(s) in the transaction.” It is usually expressed as an implication:

$$X \rightarrow Y$$

where, X and Y are sets of items. The above expression means that a transaction T contains a set of items, X, which is likely to contain another set of items, Y. Association rule mining find many applications in areas such as “market basket analysis, web usage mining, bioinformatics” [31] and so on.

Two metrics that are used to evaluate the newly discovered association rules are support (S) and confidence (C). Support of X and Y is defined as the fraction of transactions that contain both X and Y; whereas confidence measures how frequently items in Y appear in transactions that contain X. Thus, our goal can be narrowed down to finding all rules that have support and confidence greater than a threshold specified by the user. The thresholds for support and confidence are called minsup and minconf, respectively.

A brute force approach would involve listing all possible association rules, compute the support and confidence for each rule and discard those rules that are below the threshold values. However, this approach is computationally expensive. There are many techniques that are not prohibitive for mining association rules. Most popular among them is the apriori algorithm and its variations. Apriori algorithm follows a two-step approach: a) frequent itemset generation, where all itemsets that have support greater than minsup are considered; and b) rule generation, where high confidence rules are generated from each frequent itemset that exceed minconf value. Further details about these steps are discussed in Chapter 3.

CHAPTER 2

RELATED WORK

There has been significant research in the field of literature mining in the past decade. Interestingly, there have been many contributions focusing not only on development of entire system for literature mining but also on specific problems from sub-areas involved such as entity recognition and relationship extraction. Section 2.1 describes related work in information retrieval. Methods employed by researchers for entity recognition and relationship extraction are discussed in Section 2.2 and Section 2.3, respectively. A brief overview on contemporary methods used for hypotheses generation is discussed in Section 2.4. Systems that integrate literature with data from other sources (e.g. experimental datasets and biological databases) are discussed in Section 2.5.

2.1 Information Retrieval

Most of the existing systems incorporate vector space model for retrieving relevant articles from the database based on the query term provided. Unlike other retrieval systems, biomedical information retrieval systems face unique challenges such as term variations, term ambiguity and multi-worded term for a particular entity. Solutions include removal of stopwords and using thesauri for query expansion. A list of stopwords may be compiled for use in filtering out those terms from the articles. A simple search for “ErbB2 role in signal transconduction” would retrieve documents from the database containing terms “ErbB2” and “signal transconduction” based on tf-idf methods of vector space model. However, there is a greater chance that documents containing synonyms of the term “ErbB2” – HER2 – or articles mentioning only the expansion of the term would not be retrieved. Moreover, the system should be able to understand that “signal transconduction” is a gene ontology term and expand the query to retrieve more relevant documents rather matching query terms alone [9]. PubMed is known to

use the above techniques in its retrieval system [36-39]. Another major development in information retrieval is the study of presentation of the retrieved articles. Migrating from a ranked list of retrieved documents as given by Google to improved visualization such as result-summarization, wheeled or networked display of results have been studied and developed [40-44]. Examples of the web-based information retrieval systems include E-BioSci [45], XplorMed [46], Crossref [47], NPG Search [48].

2.2 Entity Recognition

There is no contest to the fact that entity recognition is a crucial step for successful text mining of the underlying associations [49, 50]. Therefore, it is no surprise that a great amount of research in entity recognition is carried out to identify names of genes and proteins from articles [17]. Methods to differentiate whether the entity in question is a gene or a protein also are studied [9]. However, there is no significant impact on information extraction as there is only a narrow line of distinction between genes and proteins as cited in [57-59]. Basically, there are two strategies that have evolved for identification of entities – dictionary-based methods and rule based methods.

Dictionary-based methods rely on recognition of terms based on known patterns such as presence of letters followed by numbers, terms ending with specific suffixes; or terms which match certain biological concepts. Dictionary-based methods also rely upon interpreting information from “neighborhood of words” [51]. Building a dictionary of gene or protein names thus helps identify entities from free texts. [52] mentions a system that uses a lexicon of names and highly frequent terms that appear in conjunction with these names. However, the most-important challenge is the constant need to revise this dictionary for additions, deletions or updates of entity names, which requires curators’ attention for agreement and validation.

The most-popular identification strategy is employing concepts from natural language

processing (NLP), such as parts-of-speech (POS) tagging, or from statistics, such as assigning a confidence measure for a term to be of particular type. AbGene [53] is based on modifications of Brill-POS tagger, where names of putative genes and proteins are tagged. It also employs identification of entities using information from words adjacent to its location. As opposed to AbGene, [54] assigns a score to each term in a document based on term morphology and boundaries. Machine learning strategies also have been employed for identification of entities concurrently with NLP concepts. For example, a hidden Markov model (HMM) was used on GENIA corpus and gave a precision of more than 60% [55].

2.2.1 Synonyms and Abbreviations

It seems highly intuitive to have a gene or protein name synonym list either created manually or downloaded from curated databases [60-62]. However, this compiled list would soon become incomplete for want of frequent updates or changes available from the databases [63-64]. Therefore, automatic strategies have been investigated to extract synonyms from text using manual or automated pattern matching rules [65, 66]. Manual and automatic pattern matching rules based on position of characters and parentheses in abbreviation compared to expanded form [17], statistical determination [67] techniques are used to help identify entities expressed in the form of abbreviations.

In short, entity recognition techniques can be employed to create dictionaries of entity names, act as precursor for relationship extraction and be used for “cross-linking literature related to genes” [9]. IHOP [56] provides a network of genes and proteins to access abstracts in PubMed using hyperlinks and thus, act as web-based tool to bring information under a single window.

2.3 Relationship Extraction

One of the most common relationships that biologists like to extract is relation between

genes or proteins. These relations are important as they help in “gene expression analysis and database annotation” [69]. Stated simply, these relationships give more insights on the functions and type of functions among genes and proteins. Some of the strategies used for relationship extraction are briefly discussed below:

- Template-based methods – A template or pattern is created, usually in the form of regular expressions, to extract the interactions between entities. Templates can be created manually or automatically. Extracting interactions from long sentences by pattern matching using manually devised rules [70]; along with complex POS rules, syntactic and semantic constraints [71] and on GENIES corpus [72] have been investigated. Manually creating patterns for all interactions is unrealistic and time-consuming. Hence, researchers have come up with automatically generating patterns based on words in the neighborhood of the text containing entities [65-66]. [33] describes a dynamic programming algorithm to automatically create unique patterns by “aligning relevant sentences” and functional verbs.
- NLP-based methods – Here techniques borrowed from NLP are employed to decompose sentences from text using tokenization, tagging of words and building structures such as parse tree [73] and semantic labels [9] from which relationships are extracted. NLP methods can be further classified based on the type of parsing – Full parsing as in [75, 76] and Partial or shallow parsing as in [77]. Not many studies have been able to resolve relationships from multiple sentences – anaphora resolution [74]. However, the latter is of less significance as research [78, 79] mention that most of the relationships between entities are present in a single sentence.

- Statistical methods – Co-occurrence is one of the most widely used statistical method [81]. It extracts associations by identifying terms that have been observed to “co-locate” more than by chance. Although it is a much straightforward way of extracting relations, it suffers from following drawbacks – unable to extract associations from complex sentences and causality about the relationships [9]. It finds greater application in database curation [82].

[80] describes a system where binary protein-protein interactions are extracted from sentences. Some other examples of relationship extraction are finding association between genes or proteins and GO codes such as [83] , event of disease [84] and other biological or pathological tasks [85-87].

2.4 Hypotheses Generation

Generation of a hypothesis from observations and existing sources of information and associations is crucial in a fast, growing field of biology and life sciences. Furthermore, there exist inherent challenges in extracting all the implicit relations from the articles themselves for need of proper, easy-to-use tools [32]. Therapeutic applications of drugs [88, 89] and biomarker discovery are some of the immediate potentials of text mining.

Several strategies are used to generate hypotheses. Swanson’s ABC model of association discovery uses extensive manual comprehension of literature from diverse backgrounds [35]. His research led to the creation of Arrowsmith system [90], which provides a list of words that are common to the titles of two literature sets. However, recent trends show interest in automatically generating hypotheses from a given corpus or a set of known relationships. Feasibility of mining associations NLP techniques such as term frequency, co-occurrence relation between terms are demonstrated in [9, 91, 92].

Some of the contemporary and available text mining tools are mentioned below:

TABLE 2.1

LIST OF TEXT MINING TOOLS

Text Mining System	Description
LitLinker [108]	Use UMLS metathesaurus to identify interesting concepts; employ association rule mining
MANJAL [109]	Use information from semantics to extract explicit relationships
IRIDESCENT [110]	Use fuzzy logic to identify new associations; employs multiple dictionaries from different sources of knowledge

One of the flip sides of automatic generation of hypotheses is that list of potential associations become very large, and hence it is necessary to formulate some evaluations strategies. Although time-consuming, a manual search of literature to account for support of the novel hypotheses is suggested in [17]. Several evaluation task forces such as knowledge discovery and data-mining (KDD) [93] and conference challenges such as BioCreAtIvE (critical assessment of information extraction systems in biology) [94] have undertaken many initiatives to develop techniques and standards to evaluate various text mining tools [2].

Apart from discovery of novel relationships, text mining also can help in “assisted curation” – management of databases containing gene or protein names, interactions and other biomedical terms [95]. GOAnnotator [96] and PreBIND [49] are examples of applications of text mining for database curation [7].

2.5 Integration Framework

There has been rapid interest in bridging the gaps among associated fields of biology and computational sciences. One such study is to integrate sources of data from multiple formats into a unified framework, which results in a centralized repository of data as well as a one-stop

source on which potentially larger set of meaningful and novel associations can be derived [97]. It also provides a way to bring researchers working on different models of organisms to connect and share or be aware of the discoveries in their sub-areas of research [9]. Some of such studies [17] undertaken are summarized in the table below:

TABLE 2.2

LIST OF SYSTEMS THAT EMPLOY INTEGRATED FRAMEWORK APPROACH

System	**Description**
TXTGATE [111]	Information from multiple online databases are integrated to identify gene pairs
PubMatrix [112]	A visualization tool to display associations from multiple queries to PubMed
Textpresso [113]	A retrieval and mining tool developed for finding gene associations in *C.elegans* (worm)

2.6 Summary

It can be observed that most of the existing state-of-the-art systems do not use full text articles for relationship extraction and hypotheses generation. Although the current systems employ strategies for identification of protein names, they fail to utilize a combination of methods to maximize the identification of protein names. A number of methods are available for extracting explicit associations from abstracts, but they do not take advantage of those associations that are precisely mentioned from captions of figures and experimental results. Strategies to produce an optimal sized set of unknown implicit associations that is feasible for subsequent clinical validations are not discussed in the literature. This thesis attempts to address the above challenges.

CHAPTER 3

METHODOLOGY

3.1 Overview

This research focuses on identifying novel relationships among genes or proteins from publicly available full-text articles using NLP and association rule mining. Figure 3.1 depicts the overall structure of this research:

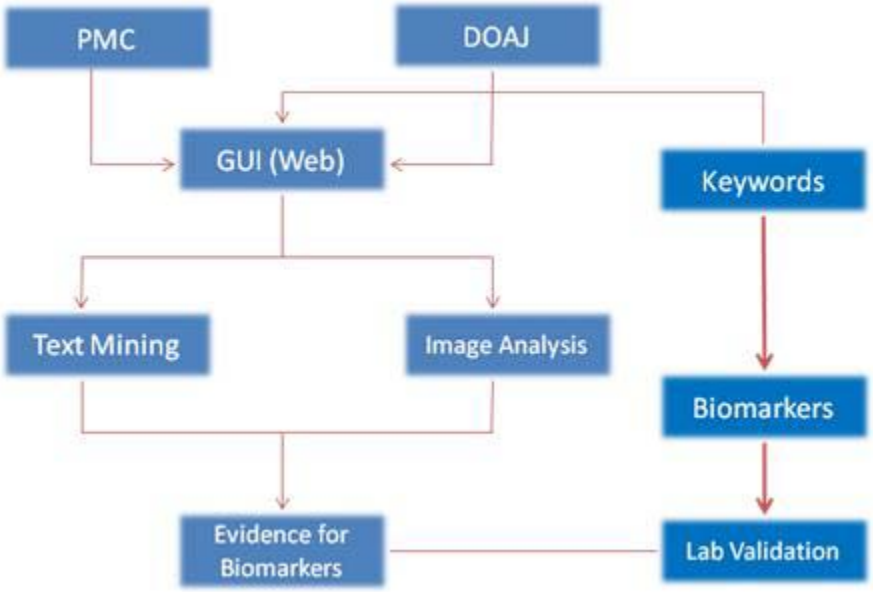


Fig. 3.1 System overview

Based on search keyword(s), journal databases such as PubMed Central and DOAJ are queried, and articles are retrieved in text, html or pdf formats and stored onto a local server. Text and images are then extracted from these documents. Implicit relationships between entities are discovered using text mining. Information from the images is extracted using techniques borrowed from image processing. Evidence for biomarkers or novel associations is hypothesized from results of text mining and image analysis. These hypotheses are then submitted to a clinician or biologist for clinical validation and verification of results by experiments.

Although finding evidence for biomarkers or new pairs of interaction is the ultimate goal of this research, biomarker discovery is preceded by a number of intermediate steps. Figure 3.2 is a flow chart that presents the various processing steps involved.

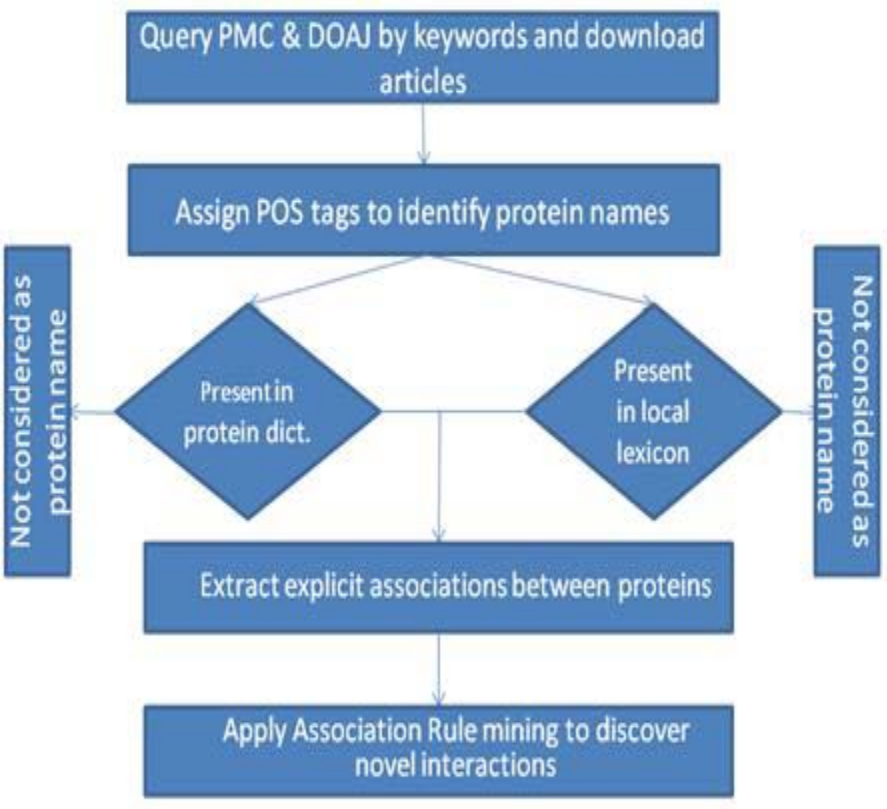


Fig. 3.2 Flow chart of processing steps involved for discovery of novel protein interactions

3.2 Building the Corpus

This research limits the task of information retrieval to downloading the corpus. No retrieval system based on Boolean model or vector space model is developed. Instead, publicly available search tools of PMC and DOAJ are utilized to retrieve articles. The PubMed system uses a mixture of both Boolean and vector space models [9], while the type of model used by DOAJ is not known or documented. A GUI called “C-Engine” is created to accept the queries from the user and pass them to PMC and DOAJ, and a listing of the results from the databases is

displayed and downloads the contents of the articles to a local server. This research uses the articles in the form of full-text, abstracts, summary and captions from figures inside the articles; whereas the state-of-the-art methods rely mostly on abstracts.

3.2.1 Web-based Search Interface

"C-Engine" is a simple user-friendly web based GUI designed to accept terms from the user to query PMC and DOAJ databases. Two text boxes are provided to accept query terms for which articles need to be fetched from PMC and DOAJ databases. Query terms could be general words (such as proteins, biomarkers, and *C.elegans*), names of proteins (such as ErbB2 and Ste4p) or phrases or sentences (such as "yeast cell cycle" and "Nab2p sequence binding"). The GUI also provides capability to combine Boolean operators to enhance or limit the number of articles returned related to the query terms. Either one or both of these databases can be used for retrieval of articles. Once the query terms are submitted using the "search" button, relevant articles from PMC and DOAJ are listed along with details of publication such as author, journal, year and volume of publication, if available, are provided.

3.2.2 C-Engine Crawler

When query terms are submitted using GUI of C-Engine, a modified web-crawler is invoked. A web-crawler [114], also known as spider, bot or robot, is a program that is used to gather pages from World Wide Web (WWW) in order to index them for the purpose of search engines. Examples of popular web crawlers are Googlebot [115], Msnbot [116] and Yahoo! Slurp [117]. Some of the characteristics of a web crawler are that it must ensure that only latest web pages are downloaded to disk for indexing. A crawler must be *polite* and honor the robot exclusion protocols, i.e., it should not crawl or index the page if the robot.txt file of the web page disallows or excludes the page to be indexed. Most importantly, crawlers must be resilient to

avoid loops and malicious pages while indexing the pages from web.

A set of seeds that consists of URLs is used as the starting point by the crawler. The crawler fetches the page mentioned in the seed-set, extracts the text and links from that page. The text contents are downloaded to a separate file while the child links are stored in a data structure, such as hash, to be later visited by the crawler. The crawler avoids a page or link visited recently to prevent falling into loops.

The C-Engine crawler uses DOAJ URL and/or PMC URL as the seed-set, with the URLs affixed with the query terms or keywords that were provided by the user from the GUI. Unlike general-purpose web-spiders, the C-Engine crawler limits crawling within DOAJ and/or PMC web pages. In other words, all web links that do not belong to DOAJ or PMC are not stored into the hash data structure, which is used to keep track of the child links. The C-Engine crawler uses a LWP package [118] to extract the contents from the web page (including a list of URLs, title, author(s), and the year of publishing) pertaining to the relevant articles matching the query term given by DOAJ and/or PMC. The crawler then downloads to the local server the articles in the available html, pdf or text format. All images are preserved in the folder that stores the same name as the article. The entire list of crawled and indexed links is written to a URL file, which also indicates the order in which the crawler traversed through the web pages. A history file is created to keep track of the articles that are downloaded and used in building the biomedical corpus.

3.2.3 Extraction of Text and Images

Articles that are available in the html format are parsed through scripts to extract text and images. Pre-processing steps such as removal of SGML tags, special characters (for example, forward and backward slashes, brackets, and comparison operators) and multiple white spaces

are invoked to extract text contents and store them in a separate file by the name of the article. Images are extracted by referencing html tags: <img> and its attributes ‘src’. In the scenario, where links contained only partial references such as “protein_signalling.jpg,” the script keeps track of the root link and concatenates it with the image file name to recover the complete link and retrieve the corresponding image. Both text and images extracted from articles are saved into corresponding files of the same name.

The files, consisting of only the text contents from their local copies of articles, constitute the corpus necessary to identify genes or proteins, extract explicit interactions between the entities (genes or proteins) and use them to discover novel associations that form the goal of this research.

3.3 Image Analysis

The scope of this thesis does not include analysis of information from images. Information from images such as color, texture and shape are employed to find more similar images. In the context of the literature mining, the content of biomedical images depicts the results of experiments or data in a better way [119]. Locating salient points of interest, localized content-based image retrieval are some of the directions in which analysis and processing of images have taken shape. Images also are extracted from the articles in html format and stored in folders by the name of same articles.

3.4 Identification of Protein Names

As discussed in Chapter 2, one of the critical steps in text mining is to identify entities of interest. Unlike financial or news corpus, biomedical publications are plagued by several names to represent a concept, and equally large number of acronyms, and variations in how they are represented. Methods, such as those mentioned in Section 2.2, are some of the best strategies

employed to circumvent the challenges posed by lack of standardized naming conventions. Hence, recognizing the entities, such as gene or protein or disease names, is the first of two steps in information extraction.

Sentences in an article span one or multiple lines. In order to extract associations among entities buried in a sentence, it is important to format the file such that each line contains exactly one sentence. This is achieved by borrowing a concept from NLP, namely part-of-speech (POS) tagging. Section 3.4.1 describes POS tagging and how it uses Lingua-EN-Tagger [120] to identify protein names from text. After the words from the text are tagged, manually coded template matching is used to split complex or compound sentences into multiple sentences [121], which is discussed in Section 3.4.2. Other strategies that complemented the identification of protein names are presented in Section 3.4.3.

3.4.1 POS Tagging

POS Tagging, also known as grammatical tagging or word-category disambiguation [122] is defined as assigning parts of speech to words in a text. Its importance lies in the fact that it gives large amount of information about a word and neighboring words. Just as parts of speech in English language gives insight about the context of the word used, POS tagging is a NLP method to infer contextual information about the word by looking at the tags assigned for that word**.** POS tagging is done using a set of tags that corresponds to 50 or more "parts of speech" for NLP. POS tagging can be classified into open and closed classes.

TABLE 3.1

TYPES OF POS TAGGING CLASSES

Class Type	**Parts of Speech**
Open	Noun, Verb, Adjective, Adverbs
Closed	Preposition, Determinant, Pronoun, Conjunction, Aux. Verb, Participle, Number

Several algorithms are developed to implement POS tagging. Some of the commonly used strategies are summarized below:

- Rule based tagging – Rule based taggers use a disambiguation rules set that mostly is manually written to resolve ambiguities concerning assignment of parts of speech to an unknown word. Most rule based taggers have large datasets of words and rule sets. An example of a rule based tagger is the EngCG Tagger, which has primarily two stages. In the first stage, all possible parts of speech for a word are returned. Then, a set of constraints are applied to prune those parts of speech that are ambiguous to the current context of the word.
- Hidden Markov model tagging – Words are assigned appropriate tags depending on the probabilistic measures computed. [123] describes it as a "sequence classification task," where the task becomes assigning a sequence of POS tags.
- Transformation based tagging – It is based on transformation-based learning approach, where principles from rule-based and stochastic systems are used. It has a rule set that determines what tags are to be assigned to words and also learns rules automatically from the data. Most of the TBL taggers need to be trained using a pre-tagged corpus. One most widely used TBL tagger is the Brill POS tagger.

3.4.1.1 Lingua::EN::Tagger

Lingua::EN::Tagger is a perl module, available from CPAN, that is developed for POS tagging. Parts of speech tags are assigned to the text using POS tag information made on corpus available from Penn Treebank project [124]. It's a statistical based tagger (HMM), and assigns appropriate tags based on the tag assigned to the preceding word. By default, Lingua::EN::

Tagger assigns noun tags to unknown words. In short, the tagger can be used for the following purposes :

- A tagged set of text
- Breaking paragraphs into individual sentences
- Extraction of all noun and noun phrases
- Frequency count of all nouns
- Extraction of maximal noun phrases

A list of POS tagset is provided below in Table 3.2:

TABLE 3.2

LIST OF POS TAGSET USED IN LINGUA::EN::TAGGER

Tag	Parts of Speech
CC	Conjunction, coordinating
CD	Adjective, cardinal number
DET	Determiner
EX	Pronoun, existential there
FW	Foreign words
IN	Preposition / Conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LRB	Punctuation, left bracket
LS	Symbol, list item
MD	Verb, modal
NN	Noun
NNP	Noun, proper
NNPS	Noun, proper, plural
NNS	Noun, plural
PDT	Determiner, prequalifier
POS	Possessive
PP	Punctuation, sentence ender
PPC	Punctuation, comma
PPD	Punctuation, dollar sign
PPL	Punctuation, quotation mark left

PPR	Punctuation, quotation mark right
PPS	Punctuation, colon, semicolon, ellipsis
PRP	Determiner, possessive second
PRPS	Determiner, possessive
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Adverb, particle
RRB	Punctuation, right bracket
SYM	Symbol
TO	Preposition
UH	Interjection
VB	Verb, infinitive
VBD	Verb, past tense
VBG	Verb, gerund
VBN	Verb, past/passive participle
VBP	Verb, base present form
VBZ	Verb, present 3SG -s form
WDT	Determiner, question
WP	Pronoun, question
WPS	Determiner, possessive & question
WRB	Adverb, question

3.4.1.2 Methods Used

For each article, the text contents are first tagged using Lingua:: EN::Tagger. Unknown words such as protein names (e.g., ACK) are assigned noun tags by default. Other entity names such as ErbB2 or its variations such as ErbB-2, erbb2 or erbb-2 are assigned adjective tags for cardinality (<cd>). With the aid of regular expressions and tags assigned by Lingua::EN::Tagger, the text contents of each article are decomposed into individual sentences spanning each line. The end result is a collection of documents where each line corresponds to one sentence, and each word (including punctuations such as '.','!','?' and ',') is tagged.

3.4.2 Processing Non-simple Sentences

Depending on the number of occurrences of main and dependent clauses, a sentence can

be either complex or compound. Extracting associations among entities, such as proteins, from a complex or compound sentence requires complex NLP techniques. However, this need not be the case as mentioned in [121, 125]. Rules based on pattern matching and tags assigned to the words are used to decompose a complex sentence into two parts. For instance, if parts of a sentence are separated by two verbs, the sentence is decomposed into two parts, each having one verb connecting them. An example is shown as follows:

"IL-2 ... stimulating … with the G-CSF granulocyte ... and also stimulates

GM-CSF interferon gamma ..."

This will be split into:

Sentence 1: *"IL-2 ... stimulating … with the G-CSF granulocyte …"*

Sentence 2: *"IL-2 also stimulates GM-CSF interferon gamma ..."*

3.4.3 Recognizing Entities

Combinations of strategies have been shown to improve identification of entity names [34, 121]. Entity identification uses the following strategies to recognize names of proteins, their variations in acronyms and synonyms.

3.4.3.1 Identification by Word Appearance and Morphology

This work considers identification of protein names in the form of abbreviations and acronyms. Fukuda et. al. [57] described an exhaustive set of rules to identify protein names and their variations. Methods to extract single terms with upper-case and/or lower-case letters, numerical values by "core-term extraction method," are employed in this work. "Core terms" generally refer to potential candidates as genes or proteins. Simple regular expressions are used to consider qualification such as the presence of "_", length of the term less than eight characters. Heuristics, such as abbreviations, that are highly likely to be enclosed in parentheses also are

considered. The above rules are used to trim down the set of terms bearing noun and adjective tags (<nn>, <nnp>, <jj> and <cd>). Thus, words which are nouns and adjectives but are not essentially names of genes or proteins are removed.

3.4.3.2 Dictionary of Proteins

A local dictionary of protein names and its synonyms for human is created using datasets [6] available from a central repository of protein names and functions called UniProt knowledge database. The datasets comprise of entries for human chromosomes 1 to 22, X, Y and index of human protein with sequence variants.

Words tagged as nouns and adjectives by Lingua::EN::Tagger are compared against this dictionary for faster identification of entity names. Another advantage is that words that were missed are captured by this matching of words against protein names in the dictionary. This dictionary can be updated for addition or deletion of protein names. It always is recommended to keep this local dictionary updated to improve identification of protein names.

3.4.3.3 Elimination by Comparison

Here, common nouns (identified using tags) present in each text document are bumped against a standard English lexicon. It uses the heuristic that English lexicons, such as that available in "/usr/share/dict/words" folder of UNIX machine, do not contain names of proteins. Thus, all non-proper nouns are removed from the list of possible candidates of protein names. A global regular expression is used to match words between text and words in UNIX lexicon. Thus, using POS tagging and above strategies, almost all protein names and its variations are identified.

3.5 Extraction of Relationship between Entities

Relationship extraction among entities helps to understand the nature, type and associated

functions of these entities. This work is interested to find out the explicit protein-protein interactions from text articles. “A protein’s function can be interpreted by how that protein interacts with the multitude of other proteins, which proteins do they mostly interact and localization characteristics within a cell. And in most cases, since proteins are the targets of attack by diseases and points of interest of therapeutic drug effects against such diseases, protein-protein interactions are intensely studied” [126]. This work focuses on binary interactions, i.e., interaction between two proteins.

One of the most-common methods used in relationship extraction is using templates either manually or automatically created. This work uses a pattern matching rule to identify protein-protein interactions. From the previous step of entity recognition, names of proteins have already been identified in the texts. In biomedical publications, associations between proteins are refereed by a “functional verb” or a “keywords.” There are a host of functional verbs and sub-classes or synonyms depending on the nature of interactions. For this purpose, a dictionary of known functional verbs is created from those available from sources such as Reactome [127] and HPRD [128].

A set of frequently used functional verbs are listed below in Table 3.3:

TABLE 3.3

PARTIAL LIST OF FREQUENTLY USED FUNCTIONAL VERBS

acetylate, activate, adhere, affect, alter, antagonize, associate, attenuate, augment,...
impact, impair, inactivate, increase, induce, influence, inhibit, initiate, interact,
involve,...phosphorylate, potentiate, precede, prevent, produce, promote, raise,
reactivate, recognize, recruit, reduce, regulate,... transform, trigger, up regulate

Lines of text are parsed to identify the keywords. In order to avoid difficulties in matching, the keywords are stemmed, i.e., broken to their roots, and thus accommodating matches for all possible variations of the keywords. For example, any occurrences of words, such as activated, activating, activation and activate, will be stemmed to its root and matched. Porter stemmer [129] is used for stemming functional verbs. The matched keywords are identified in the text articles. Patterns such as [.*protein1.*keyword.*protein2.*] are used to extract lines with identified protein names and interaction keyword. Some examples of such associations extracted are:

Bnr1p ***interacts*** *with Rho4p*

Cdc28 ***phosphorylates*** *Swe1*

P21 ***inhibits*** *Cdk2*

P21 ***inhibits*** *Cdk3*

3.6 Discovering Novel Associations using Text Mining

Text mining is employed to find implicit relations between the entities. This work uses association rule mining to discover novel associations between two proteins. New rules or associations are discovered based on measures of support and confidence. This work uses apriori algorithm for implementing association rule algorithm. This section discusses apriori algorithm and how it is incorporated to discover novel interactions between proteins.

An itemset whose support is greater than or equal to the minimum support threshold is called a frequent itemset. One of the key concepts of apriori principle is: “subsets of a frequent itemset are also frequent.” This can be mathematically described as the anti-monotone property of support. In other words, support of an itemset should never exceed the support of its item subsets. The entire problem of association rule mining using apriori algorithm can be

summarized in two steps:

a) Frequent itemset generation, and

b) Rule generation

The following explains how to create a frequent itemset using apriori algorithm.

Step 1: Generate a frequent itemset of length 1 (i.e., 1-itemset).

Step 2: Using frequent itemsets from step 1, create candidate itemsets of length one more than the frequent itemset.

Step 3: Candidate itemsets having infrequent items of length 1 are eliminated.

Step 4: Support of each candidate is counted.

Step 5: Eliminate candidates that are infrequent.

Step 6: Repeat steps 2 to 6 until no frequent itemsets are identified.

One of the most computationally expensive steps in the algorithm is counting the support for each candidate. Usually the entire database has to be scanned to determine the support of each candidate itemset. A hash structure is used to store the candidates and thus, comparison is made against hash buckets than each transaction against every candidate.

New rules are generated from the frequent itemsets by merging two rules that share the same prefix in the rule consequent. Only those rules are kept for which even its subsets have higher confidence. Rules that fall below minimum confidence threshold are pruned.

A transaction file is created from the list of explicit interactions of proteins. This transaction file consists of breakdown of items in a transaction. For example, an interaction such as “p21 inhibits Cdk3” is considered as a transaction with items p21 and Cdk3. A frequent itemset of length 2 is created. A hash structure is created to store the candidate itemsets. Depending on the threshold of confidence, a set of novel associations between proteins are

generated.

3.7 Generating Hypotheses - Strategy

One of the most significant attributes of this thesis is using full text articles, along with abstracts of articles and captions from figures, to extract explicit interactions between proteins and thereby using them to discover possible new protein-protein interactions. The importance of this strategy lies in the facts discussed in Section 1.5.2, which is, using full text articles, brings to light relationships that are secondary to that particular article, but might be of major importance in a sub-class of the same research domain. Thus, it provides a higher probability of finding implicit associations among entities that were never thought to interact, had only abstracts been used.

It is true, that incorporating such large text contents for an article can churn out a relatively large set of implicit relations. Also, the likelihood that some of these implicit relations are previously known relations or they could be interactions that have no biological significance cannot be ignored. In light of these facts, the goal of this thesis is clearly not just to identify all possible implicit associations that are precise or biologically significant. On the contrary, the methods employed are used to maximize the set of known associations and minimize the set of unknown associations that were found by apriori algorithm. In other words, our objective is to determine the optimal threshold for support and confidence to achieve a set of manageable size of unknown interactions and thus, enhancing feasibility to verify the above implicit associations by biological experiments.

CHAPTER 4

EXPERIMENTS AND DISCUSSION

4.1 Experiments

C-Engine is a web interface for entering keywords or query terms to find associated protein-protein interactions. Figure 4.1 shows a screenshot of C-Engine.

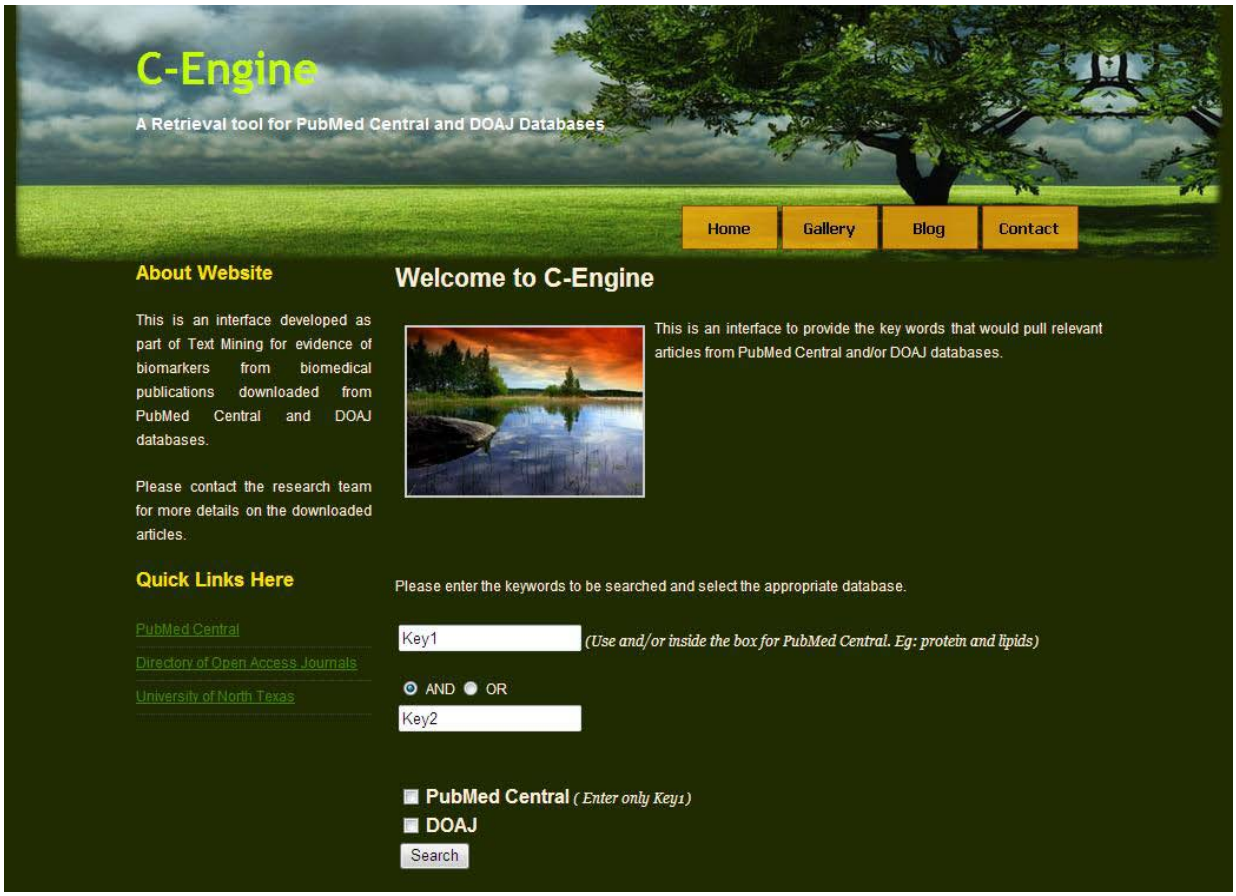


Fig. 4.1 Screenshot of C-Engine search portal

The fields “Key1” and “Key2” are used to accept keywords. Again, it is the user’s discretion to select the databases from which articles related to the query terms are to be retrieved. This thesis uses articles retrieved from DOAJ and PMC databases. Either one or both of these databases may be used for retrieval of articles. In all experiments for different query terms, the entire list of retrieved articles is used for finding explicit and implicit protein-protein

interactions. Once the user submits the query term(s) at the search interface, the C-Engine crawler is invoked, which a) produces a list of articles, along with the article name, publication date, authors and so on, that are displayed in the result page; and b) downloads the articles in html, pdf or text format that is available to the local server. Figure 4.2 is a screenshot of the results for a query term.


Results

Differential Proteome Analysis of the Preeclamptic Placenta Using Optimized Protein Extraction
Author Magnus Centlow Stefan R. Hansson Charlotte Welinder Journal Journal of Biomedicine and Biotechnology Year 2010 Vol 2010 Issue Pagesrecord No.
Combination PPAR&x03B3 and RXR Agonist Treatment in Melanoma Cells Functional Importance of S100A2
Author Joshua P. Klopper Vibha Sharma Reid Bissonnette Bryan R. Haugen Journal PPAR Research Year 2010 Vol 2010 Issue Pagesrecord No.
The Use of pp150 and gp116 Synthetic Peptides in the Detection of CMV Antibodies
Author F Nejatollahi I Alsharni M Moazen N Farahbakhsh Journal Iranian Red Crescent Medical Journal Year 2010 Vol 12 Issue 1 Pagesrecord No. 17-21
Amelioration of lead induced hepatotoxicity by Allium sativum extracts in Swiss albino mice
Author Sharma A Sharma V Kansal L Journal Libyan Journal of Medicine Year 2010 Vol 5 Issue Pagesrecord No. 091107
Haematology Blood Chemistry and Carcass Characteristics of Growing Rabbits Fed Grasshopper Meal as a Substitute for Fish Meal
Author A. A. Njidda and C. E. Isidahomen1 Journal Pakistan Veterinary Journal Year 2010 Vol 30 Issue 1 Pagesrecord No. 7-12
Effect of Lysine Supplementation in Low Protein Diets on the Performance of Growing Broilers
Author Saima M. Z. U. Khan M. A. Jabbar A. Mehmud1 M. M. Abbas and A. Mahmood1 Journal Pakistan Veterinary Journal Year 2010 Vol 30 Issue 1 Pagesrecord No. 17-20
CaracterizaciÃ³n fisicoquÃmica y comportamiento tÃ©rmico del aceite de â€œalmendraâ€ de guanÃ bana (Annona muricata L)
Author SolÀs-Fuentes J. A. Amador-HernÃ¡ndez C. HernÃ¡ndez-Medel M. R. DurÃ¡n-de-BazÃ°a M. C. Journal Grasas y Aceites Year 2010 Vol 61 Issue 1 Pagesrecord No. 58-65
Effect of natural antioxidants on the stability of ostrich meat during storage
Author Abou-Arab Esmat A. Abu-Salem Ferial M. Journal Grasas y Aceites Year 2010 Vol 61 Issue 1 Pagesrecord No. 102-108
Mass Spectrometry-Based Label-Free Quantitative Proteomics
Author Wenhong Zhu Jeffrey W. Smith Chun-Ming Huang Journal Journal of Biomedicine and Biotechnology Year 2010 Vol 2010 Issue Pagesrecord No.
Challenges for Biomarker Discovery in Body Fluids Using SELDI-TOF-MS
Author Muriel De Bock Dominique de Seny Marie-Alice Meuwis Jean-Paul Chapelle Edouard Louis Michel Malaise Marie-Paule Merville Marianne Fillet Journal Journal of Biomedicine and Biotechnology
Tight Junctions A Barrier to the Initiation and Progression of Breast Cancer
Author Kieran Brennan Gozie Offiah Elaine A. McSherry Ann M. Hopkins Journal Journal of Biomedicine and Biotechnology Year 2010 Vol 2010 Issue Pagesrecord No.
Potentials of phototrophic bacteria in treating pharmaceutical wastewater
Author E. I. Madukasi X. Dai C. He J. Zhou Journal International Journal of Environmental Science and Technology Year 2010 Vol 7 Issue 1 Pagesrecord No. 165-174
Silica Exposure and Serum Angiotensin Converting Enzyme Activity
Author RR Tiwari AB Kamik YK Sharma Journal International Journal of Occupational and Environmental Medicine Year 2010 Vol 1 Issue 1 Pagesrecord No. 21-28
Autoreactivity to sweat glands and nerves in clinical scabies infection
Author Ana Maria Abreu Velez A. Deo Klein III Michael S. Howard Journal North American Journal of Medical Sciences Year 2010 Vol 2 Issue 1 Pagesrecord No. 42-45
Etude des facteurs environnementaux influenÃ§ant la teneur en urÃ©e dans le lait de vache en Wallonie (Belgique)
Author Dufrasne I. Istasse L. Lambert R. Robaye V. Hornick JL. Journal Biotechnologie Agronomie SociÃ©tÃ© et Environnement Year 2010 Vol 14 Issue s1 Pagesrecord No. 59-66
Optimisation de la fertilisation azotÃ©e de cultures industrielles lÃ©gumiÃ¨res sous irrigation
Author Fonder N. Heens B. Xanthoulis D. Journal Biotechnologie Agronomie SociÃ©tÃ© et Environnement Year 2010 Vol 14 Issue s1 Pagesrecord No. 103-111
An Overview of Stress Response and Hypometabolic Strategies in Caenorhabditis elegans Conserved and Contrasting Signals with the Mammalian System
Author Benjamin Lant Kenneth B. Storey Journal International Journal of Biological Sciences Year 2010 Vol 6 Issue 1 Pagesrecord No. 9-50
Protein S100B and physical exercise
Author Cintia Alvim Stocchero Alexandre Pastoris Muller Ã Ivaro Reischak Oliveira Luis Valmor Portela Journal Revista Brasileira de Cineantropometria e Desempenho Humano Year 2010 Vol 12 Issue 1
Signaling molecules involved in immune responses in mussels
Author S Koutsogiannaki M Kaloyianni Journal Invertebrate Survival Journal Year 2010 Vol 7 Issue 1 Pagesrecord No. 11-21
The Effect of Oxamate on LDH-C4 Activity of Sperm in Rats
Author Saki J. Kadkhoaei Elyaderani M. Rahim F. Saki Gh. Kalantar Mahdavi S.R. Journal Qom University of Medical Sciences Journal Year 2010 Vol 3 Issue 4 Pagesrecord No. 3-10

Fig. 4.2 Screenshot of articles retrieved for a query term

Four different query terms are identified for the experiments. Each query term is used to search and retrieve articles from DOAJ and PMC separately. This is done in order to highlight the significance of the number of articles retrieved, the number of explicit associations extracted, the number of implicit associations discovered and its effectiveness in addressing problems of entity identification by respective search engines of DOAJ and PMC. Similarly, query terms are

identified that range from very general (English) words to very specific (protein names as symbols) terms, from single-word keywords to multi-word keywords, from cellular pathways to terms associated with diseases. Following are the query terms for which articles are retrieved from PMC and DOAJ and ultimately, novel relations between proteins are found.

TABLE 4.1

LIST OF QUERY TERMS USED FOR DOCUMENT RETRIEVAL FROM PMC AND DOAJ

protein	human protein	ERBB2 breast cancer	IL2 signaling

C-Engine crawler affixes each of the above query terms to PMC URL or DOAJ URL to form the seed-set, which serves as the initial webpage from which contents, such as text and web-links, are fetched. The LWP package of the C-Engine is used to extract text and links from the webpage. The C-Engine crawler downloads all articles and stores them in the local server. C-Engine keeps track of all the URLs for each downloaded article and maintains a log file with names of articles that are retrieved for a particular query. Table 4.2 shows the number of articles used for biomarker discovery for each query term from PMC and DOAJ.

TABLE 4.2

NUMBER OF ARTICLES RETRIEVED FROM PMC AND DOAJ FOR EACH QUERY TERM

Query Term	Number of Articles	
	PMC	**DOAJ**
Protein	7000	1213
Human Protein	10000	554
ERBB2 breast cancer	2238	38
IL2 signaling	291	4

Although both databases retrieve more number of articles than mentioned above, not all articles can be used for successive steps in text mining. This is because the articles retrieved includes html, pdf and text formats. Only html formats of articles are used in this thesis for text mining. Secondly, the retrieved articles consist of non-English publications. Such articles are removed as they do not contribute to the current scope of extracting implicit associations. Thirdly, it has been found that some retrieved files are empty. This could be caused by the limit posted by servers to control the number of articles fed to an individual IP address. Such practice is usually used to avoid malicious network jam. All such empty files are removed to build the required corpus of size as shown in Table 4.2. Thus, the number of articles for the query term "human protein" is greater than that of "protein" from PMC database is attributed to the reasons described above.

Articles in html format are parsed to extract text and images. SGML tags, multiple white spaces and special characters are removed using pre-processing steps in the script. The result is the creation of a text file with only text contents for each article. Each word and punctuation mark in these texts is then tagged with appropriate POS tags using Lingua::EN::Tagger as a pre-cursor to identify protein names. The tagged text contents are decomposed into individual sentence that span each line. Figure 4.3 illustrates application of POS tagging on parts of text. Words tagged with <nnp> and <cd> are potential names of proteins and are highlighted in yellow.

Complex or compound sentences are split into simple sentences using pattern matching rules and assignment of tags to each word. Manually written templates, such as presence of a conjunction tag connecting sentences that contain at least two nouns and a verb in each part, are used to decompose complex or compound statements.

Text containing sentences spanning multiple lines	FXR and SHP protein abundance was induced by FGF-19 and repressed after silencing. FGF-19 treatment led to a reduction in ASBT expression, whereas silencing increased ASBT levels.
Each sentence spans one line. Each word and punctuation marks are tagged	**<nnp>FXR</nnp><cc>and</cc><nnp>SHP</nnp><nn>protein</nn><nn>abundance</nn><vbd>was</vbd><vbn>induced</vbn><in>by</in><cd>FGF19</cd><cc>and</cc><jj>repressed</jj><in>after</in><vbg>silencing</vbg><pp>.</pp>** **<cd>FGF19</cd><nn>treatment</nn><vbd>led</vbd><to>to</to><det>a</det><nn>reduction</nn><in>in</in><nnp>ASBT</nnp><nn>expression</nn><ppc></ppc><in>whereas</in><nnp>silencing</nnp><vbd>increased</vbd><nnp>ASBT</nnp><nns>levels</nns><pp>.</pp>**

Fig. 4.3 Application of POS tagging on parts of text article

As mentioned in Section 3.4.3, a collection of methods is used to maximize identification of protein names that are in the form of symbols or acronyms. Rules set by Fukuda et. al. [57] for identification of protein symbols or acronyms are used to trim down list of words bearing noun and adjective (cardinal) tags. Thus, nouns or adjectives that are not names of proteins are removed. A dictionary of human protein names and their synonyms is created from datasets available from UniProt. A lexical order of proteins for human chromosomes 1 to 22 and sex chromosomes (X and Y) are maintained locally, which may be edited or updated at regular intervals. This protein dictionary is created using datasets available from UniProt that were last updated on 23rd of March, 2010. Protein names, such as FOR, IMPACT, HIS and NOT, that represent common English words, are deleted. Although such protein names are very few, they are removed to avoid ambiguities in extracting associations from text. Thus, this local protein

dictionary comprises of 19176 protein names (including their synonyms). Figure 4.4 shows a part of the protein dictionary used in the experiments.

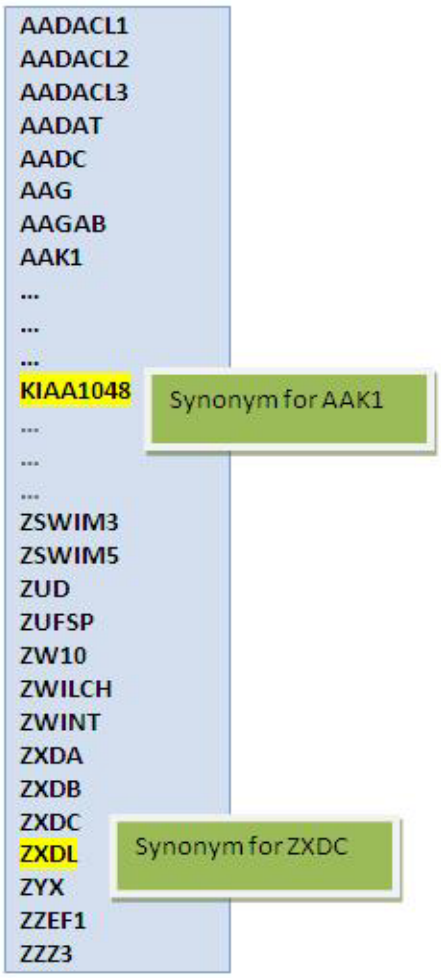


Fig. 4.4 Sliced view of the protein dictionary containing protein names and their synonyms

All words from the text that match with entries in the local protein dictionary are identified as protein names. In addition to those words labeled with noun or adjective tags, words recognized by the "elimination by comparison" method also are used for matching against the protein dictionary. The latter consists of those nouns that are not eliminated by comparison against an English lexicon, such as that available in the "/usr/share/dict/words" folder in UNIX systems. As a result, all protein names present in those articles are identified.

Binary protein-protein interactions, such as "P21 inhibits Cdk2", are extracted from the

text using template-based pattern matching. For this purpose, a list of functional verbs or keywords, such as activate, adhere, affect, inhibit, interact and regulate, is created from sources such as Reactome and HPRD. Porter stemming is used to decompose occurrences of all variations of functional verbs (e.g. inducing, induced) into its root. Patterns, such as presence of a functional verb between two protein names that are identified, are used for extracting binary interactions. Table 4.3 shows a list of explicit associations extracted from text articles for the query term "ERBB2 breast cancer".

TABLE 4.3

LIST OF EXPLICIT PROTEIN-PROTEIN INTERACTIONS FOR QUERY TERM "ERBB2 BREAST CANCER"

Article : Modeling ERBB receptor-regulated G1S transition to find novel targets for de novo trastuzumab resistance CDK2 increase CDK4 EGF signaling AKT1 ERBB2 activate AKT1 ERBB2 resulted EGFR CDK4 targeting MEK1 ERK1 targets ERBB1 ERBB2 resulted EGFR ERBB2 potentiates KIP1 Total interactions found: 8
Article : Site-specific relapse pattern of the triple negative tumors in Chinese breast cancer patients Total interactions found: 0
Article : Prolyl isomerase Pin1 is highly expressed in Her2-positive breast cancer and regulates erbB2 protein stability HER2 result ERBB2 Ren suppresses HER2 Total interactions found: 2

A list of all explicit binary protein-protein interactions and the number of articles that cite these associations are consolidated. The list of explicit associations is further used to create a transaction file, which treats each binary association as items belonging to one transaction. This transaction file is the input to discover novel associations between proteins. One of the most widely used association rule mining techniques, apriori algorithm, is employed to: a) create frequent itemsets of size-2, where all itemsets that have support greater than a threshold; and b) generate rules or possible new interactions between proteins using a threshold of confidence extracted from articles.

During candidate itemset generation, a hash structure is used to store the candidate items. It serves two purposes: a) an efficient way of storing candidate items; and b) a speedy access and faster computation of support for each candidate item than line-by-line comparison for every candidate item in each transaction. Size of transaction files varies depending on the number of explicit associations that were extracted from articles for each query. Table 4.4 shows a fraction of a transaction file.

TABLE 4.4

A VIEW OF A TRANSACTION FILE

Transaction File-id	**Entity name (Item name)**
1	AHR
1	ARNT
2	AHR
2	AHRR
3	AHR
3	ERBB2
4	AHR
4	ERBB2
…	…
184	ERBB1
184	ERBB3

185	FBS
185	SST

Experiments are run for varying threshold of support and confidence using each query term. Tables 4.5 and 4.6 show the number of unknown associations discovered for each query term run against PMC and DOAJ databases.

TABLE 4.5

NUMBER OF UNKNOWN ASSOCIATIONS FOR EACH QUERY TERM FOR ARTICLES RETRIEVED FROM PMC

Database Used : PMC				
Threshold for Support and Confidence (msup – minimum support, mconf – minimum confidence)	**Total number of unknown associations**			
	protein	**human protein**	**ERBB2 breast cancer**	**IL2 signaling**
msup = 1, mconf = 0.5	1247	1484	1125	501
msup = 1, mconf = 0.75	663	804	670	334
msup = 1, mconf = 1	632	764	654	331
msup = 2, mconf = 0.5	290	290	121	43
msup = 2, mconf = 0.75	83	101	54	20
msup = 2, mconf = 1	52	61	38	17
msup = 3, mconf = 0.5	172	171	54	14
msup = 3, mconf = 0.75	52	57	29	6
msup = 3, mconf = 1	33	17	13	3

TABLE 4.6

NUMBER OF UNKNOWN ASSOCIATIONS FOR EACH QUERY TERM FOR ARTICLES RETRIEVED FROM DOAJ

Database Used : DOAJ				
Threshold for Support and Confidence (msup – minimum support, mconf – minimum confidence)	**Total number of unknown associations**			
	protein	**human protein**	**ERBB2 breast cancer**	**IL2 signaling**
msup = 1, mconf = 0.5	624	560	79	51
msup = 1, mconf = 0.75	367	332	45	39
msup = 1, mconf = 1	352	316	42	39
msup = 2, mconf = 0.5	128	137	18	1
msup = 2, mconf = 0.75	55	55	6	1
msup = 2, mconf = 1	40	39	2	1
msup = 3, mconf = 0.5	67	72	13	0
msup = 3, mconf = 0.75	29	32	5	0
msup = 3, mconf = 1	14	16	2	0

4.2 Evaluations and Observations

In order to eliminate some of the already known protein-protein interactions from the list of implicit associations, a set of binary protein-protein interactions are created from a file available from the Human Protein Reference Database (HPRD), version 8, last updated on July 6, 2009. The set consists of symbols for interacting pairs of proteins. There are 38,088 known associations mentioned in this dataset. Table 4.7 shows a slice of the binary protein-protein interaction dataset.

TABLE 4.7

A VIEW OF PROTEIN INTERACTION PAIRS

AATF => TSG101
ABAT => ABAT
ABCA1 => ARHGEF11
ABCA1 =>

ARHGEF12
ABCA1 => DLG2
ABCA1 => DLG3
ABCA1 => DLG5
ABCA1 => GOPC
…
…
…
ZWINT => MIS12
ZWINT => ZW10
ZXDC => NR1H2
ZXDC => RORA
ZYG11B => CUL2
ZYG11B => TCEB1
ZYX => ACTN1
ZYX => VASP
ZYX => VAV1

4.2.1 Evaluating Explicit Associations

One of the most-common methods to evaluate an information retrieval task is to compare the results against a gold standard. Performance metrics, such as precision, recall and F-measure, are commonly used for evaluation in such cases. Precision (P) can be defined as the fraction of associations that were relevant. Recall (R) can be defined as the fraction of relevant associations that were retrieved. F-measure is defined as the harmonic mean of precision and recall.

However, finding performance metrics for this work is extremely hard for the following reasons. Firstly, it requires an expert to label all explicit associations present in a retrieved article and indicate their relevance. This is a tedious work, and to our best knowledge, there is no such reference data available. Therefore, it is not possible to compute recall. Secondly, as mentioned in [9], the values for these metrics depend on number of positive samples in the corpus to be evaluated. [130] shows different evaluation corpora provides large variation in the performance metrics and proves comparison against a tagged corpora yield better evaluation methods.

However, a manually tagged corpus for the articles used in this thesis is unavailable. Although precision is the fraction of relevant association in the articles retrieved, it may be modified as the fraction of explicit associations that found a match in the dataset of binary protein interactions. Table 4.8 shows computation of a measure for explicit associations extracted for each query term against the HPRD dataset of 38,088 interactions.

TABLE 4.8

NUMBER OF EXPLICIT ASSOCIATIONS – EXTRACTED AND MATCHED

Query Term	PMC			DOAJ		
	Number of Explicit associations extracted	**Number of Explicit associations matched**	**Ratio of matched associations to total HPRD interactions (%)**	**Number of Explicit associations extracted**	**Number of Explicit associations matched**	**Ratio of matched associations to total HPRD interactions (%)**
Protein	7320	321	0.84	1054	47	0.12
Human protein	10918	386	1.01	1055	21	0.06
ERBB2 breast cancer	3323	144	0.38	91	12	0.002
IL2 signaling	677	33	0.09	31	2	0.005

Another observation that follows is the time taken to identify protein names in an article. While the intermediate steps of text mining are rather efficient (an average of 10 minutes), identification of protein names may take hours to complete the step. Table 4.9 shows the approximate time taken for protein name identification.

TABLE 4.9

TIME TAKEN FOR IDENTIFICATION OF PROTEIN NAMES

Query	Identification of protein names	
	PMC	**DOAJ**
	Time taken	**Time taken**
Protein	~28 hrs	~8 hrs
Human protein	~27 hrs	~5 hrs
ERBB2 breast cancer	~7 hrs	~20 min
IL2 signaling	~30 min	~2 min

It is straightforward to deduct that larger the corpus, the more time it takes for the step to complete.

4.2.2 Evaluating Associations from Abstracts

In order to show the effectiveness of the text mining methods, it is essential to show how this method performs on abstracts alone. Abstracts from PMC for the same four query terms are downloaded, tagged and listed the number of explicit interactions. Table 4.10 compares the explicit interactions identified for the same query terms over the same documents crawled.

TABLE 4.10

COMPARISON USING ABSTRACTS AND FULL TEXTS

Query Term	PMC (Full Texts)		PMC (Abstracts)	
	Number of Explicit associations extracted	**Number of Explicit associations matched**	**Number of Explicit associations extracted**	**Number of Explicit associations matched**
Protein	7320	321	188	11
Human protein	10918	386	251	16
ERBB2 breast cancer	3323	144	72	8
IL2 signaling	677	33	7	0

Extending text mining methods on the explicit associations using abstracts, number of novel associations for the four query terms can be derived as shown in Table 4.11.

TABLE 4.11

NUMBER OF MATCHED AND NOVEL ASSOCIATIONS BETWEEN PROTEINS FROM PMC - ABSTRACTS

Database Used : PMC (Abstracts)								
Threshold for Support and Confidence	protein		human protein		ERBB2 breast cancer		IL2 signaling	
	Novel	**Matched**	**Novel**	**Matched**	**Novel**	**Matched**	**Novel**	**Matched**
msup = 1, mconf = 0.5	266	23	356	24	70	7	14	2
msup = 1, mconf = 0.75	202	18	275	18	56	6	10	1
msup = 1, mconf = 1	198	18	268	15	56	6	10	1
msup = 2, mconf = 0.5	33	4	66	30	22	2	0	0
msup = 2, mconf = 0.75	25	3	47	4	18	1	0	0
msup = 2, mconf = 1	21	3	40	1	18	1	0	0
msup = 3, mconf = 0.5	12	0	19	3	8	1	0	0
msup = 3, mconf = 0.75	10	0	17	3	6	0	0	0
msup = 3, mconf = 1	6	0	10	0	6	0	0	0

It can be observed that number of explicit interactions found is significantly less for all the four query terms. Moreover, as the query terms become more specific, the number of explicit associations using abstracts significantly drops or there are no associations. Hence, it can be confidently stated that abstracts are specific and associations extracted from full text articles are larger and contain secondary information.

4.2.3 Evaluating Implicit Associations

Bumping the dataset of known associations of interacting pairs of proteins results in

eliminating some of the hitherto relations that were identified as novel. Thus, a hypotheses set of novel binary protein-protein interactions is generated. Tables 4.12 and 4.13 show the number of unknown associations that had a match, when compared against local dataset of known protein associations and the number of candidate novel (unknown) associations.

TABLE 4.12

NUMBER OF MATCHED AND NOVEL ASSOCIATIONS BETWEEN PROTEINS FROM PMC

Database Used : PMC								
Threshold for Support and Confidence	protein		human protein		ERBB2 breast cancer		IL2 signaling	
	Novel	**Matched**	**Novel**	**Matched**	**Novel**	**Matched**	**Novel**	**Matched**
msup = 1, mconf = 0.5	1195	52	1419	65	1082	43	474	27
msup = 1, mconf = 0.75	639	24	775	29	655	15	317	17
msup = 1, mconf = 1	614	18	739	25	639	15	314	17
msup = 2, mconf = 0.5	266	24	260	30	111	10	39	4
msup = 2, mconf = 0.75	74	9	90	11	53	1	19	1
msup = 2, mconf = 1	49	3	54	7	37	1	16	1
msup = 3, mconf = 0.5	156	16	149	22	50	4	12	2
msup = 3, mconf = 0.75	44	8	48	9	29	0	6	0
msup = 3, mconf = 1	19	14	13	4	13	0	3	0

TABLE 4.13

NUMBER OF MATCHED AND NOVEL ASSOCIATIONS BETWEEN PROTEINS FROM DOAJ

Database Used : DOAJ								
Threshold for Support and Confidence	protein		human protein		ERBB2 breast cancer		IL2 signaling	
	Novel	**Matched**	**Novel**	**Matched**	**Novel**	**Matched**	**Novel**	**Matched**
msup = 1, mconf = 0.5	596	28	540	20	71	8	47	4
msup = 1, mconf = 0.75	349	18	321	11	42	3	36	3
msup = 1, mconf = 1	335	17	307	9	39	3	36	3
msup = 2, mconf = 0.5	123	5	131	6	12	6	1	0
msup = 2, mconf = 0.75	52	3	53	2	5	1	1	0
msup = 2, mconf = 1	38	2	39	0	2	0	1	0
msup = 3, mconf = 0.5	64	3	67	5	11	2	0	0
msup = 3, mconf = 0.75	27	2	30	2	5	0	0	0
msup = 3, mconf = 1	13	1	16	0	2	0	0	0

4.2.4 Visualization

Using visualization tools, such as NodeXL [131], different interactions among proteins can be easily identified than comprehension by data alone. Another advantage of using visualization tools is that it brings out the hidden network of interactions that would have been difficult to observe otherwise. Also, it can be used as a method to evaluate some of the novel interactions found using text mining.

Several visualizations of protein interactions for the query terms “ERBB2 breast cancer” are created and evaluated. Figures 4.5 and 4.6 show the network of explicit protein-protein interactions for the query term “ERBB2 breast cancer”. The red points indicate unique proteins

mined for the above search term and the black lines indicate the interaction among those proteins. Irrespective of different shapes of the network model, it is observed that the network is highly complex to comprehend and decipher hidden interactions.

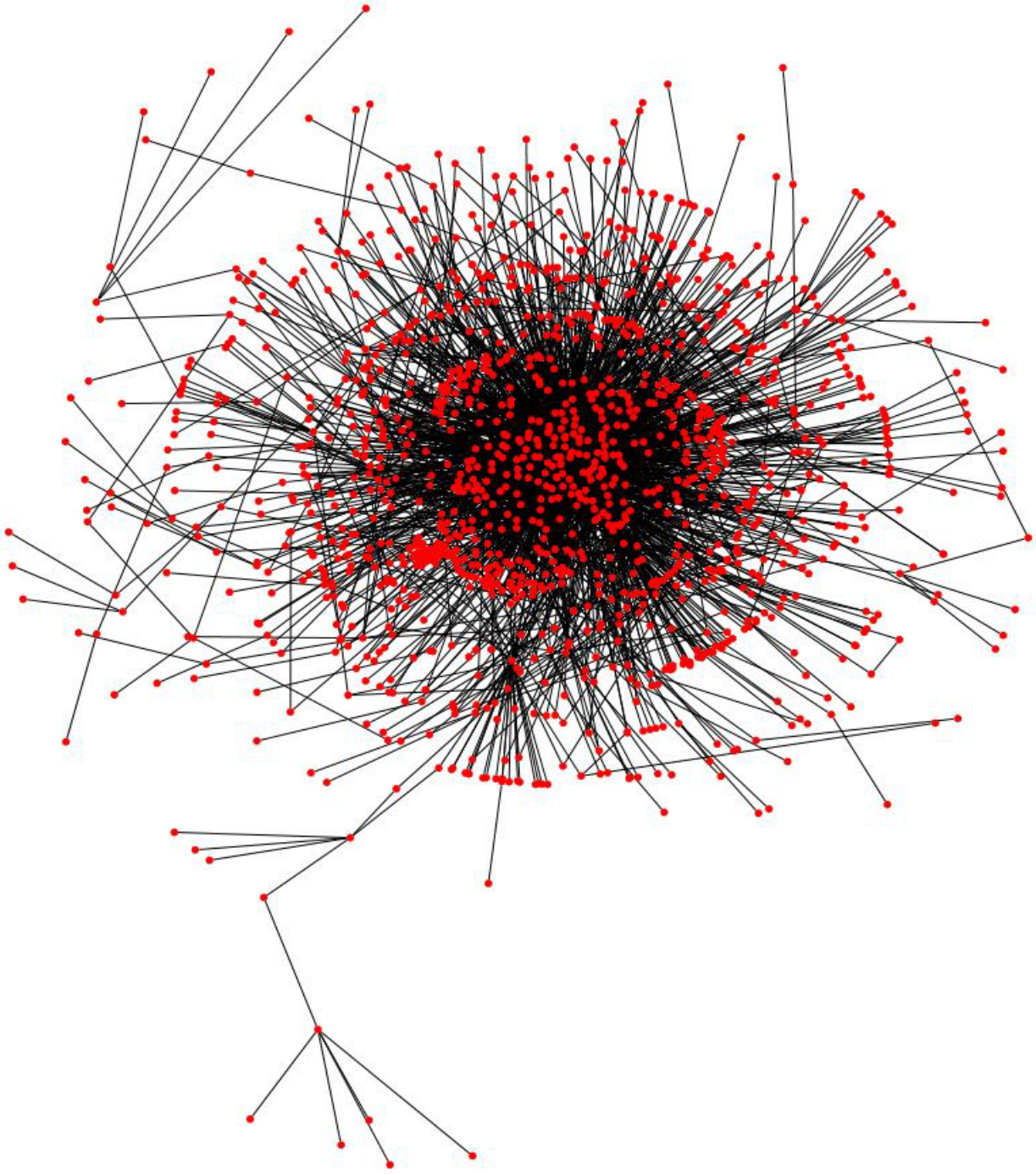


Fig. 4.5 Network of explicit protein-protein interactions for query – ERBB2 breast cancer

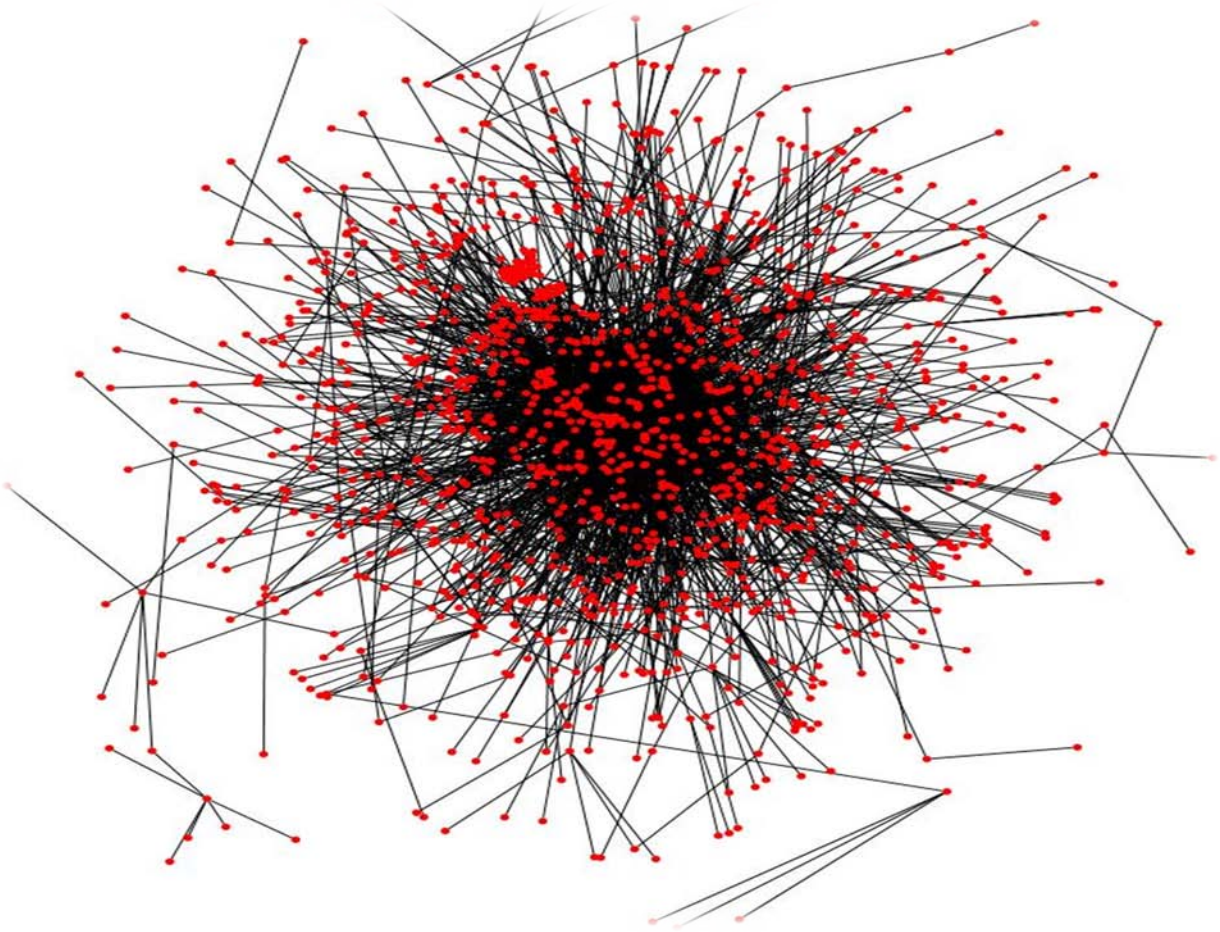


Fig. 4.6 Another visualization of explicit protein-protein interactions for ERBB2 breast cancer

Despite the complex view of the protein-protein interactions, there is not much information can be gathered from the above networks, except for the fact that most of the interactions among proteins limited to the center. A network of explicit protein interactions but with degree of freedom equal to 3 and above is shown in Figure 4.7. Degree of freedom is the number of interactions a protein has with other proteins. Biomarkers for breast cancer ERBB2 and its synonym HER2 are shown as solid yellow diamonds.

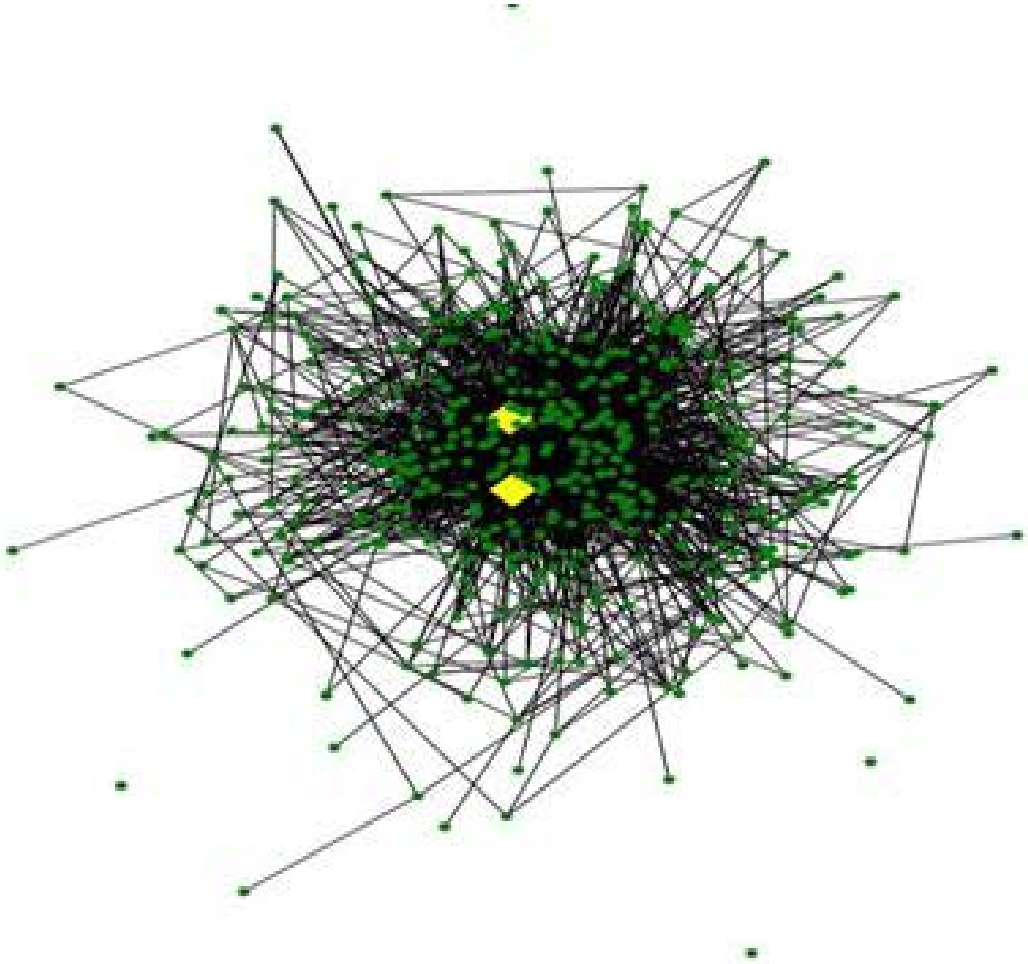


Fig. 4.7 Network of explicit interactions with limited degree of freedom

Figure 4.8 shows the network of ERBB2 and HER2 and shows the effect of term variations in biomedical literature. An interesting observation is found when all the interactions for proteins ERBB2 and HER2 are mapped. Although ERBB2 and HER2 refer to the same entity, it is observed that some of the proteins are found to have correlation only with ERBB2 or HER2 alone and remaining proteins have been correlated with both the synonyms. It is possible that this visualization provides an insight into interacting pairs that were not identified earlier with breast cancer.

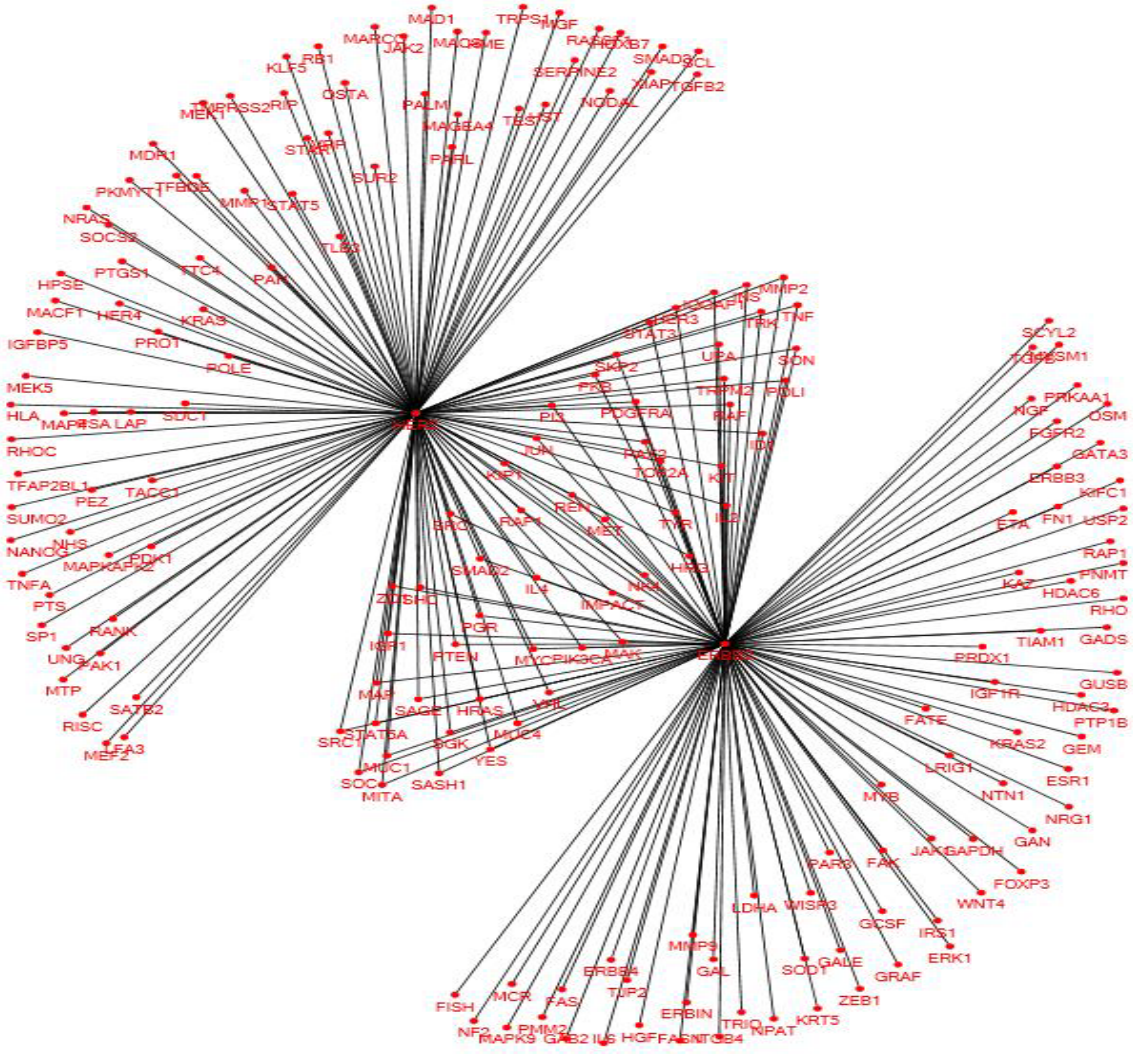


Fig. 4.8 Network of ERBB2 and HER2 – case of synonyms

Another visualization of the network of ERBB2 and HER2 proteins validates the above inference. Figure 4.9 shows a different network of both synonyms.

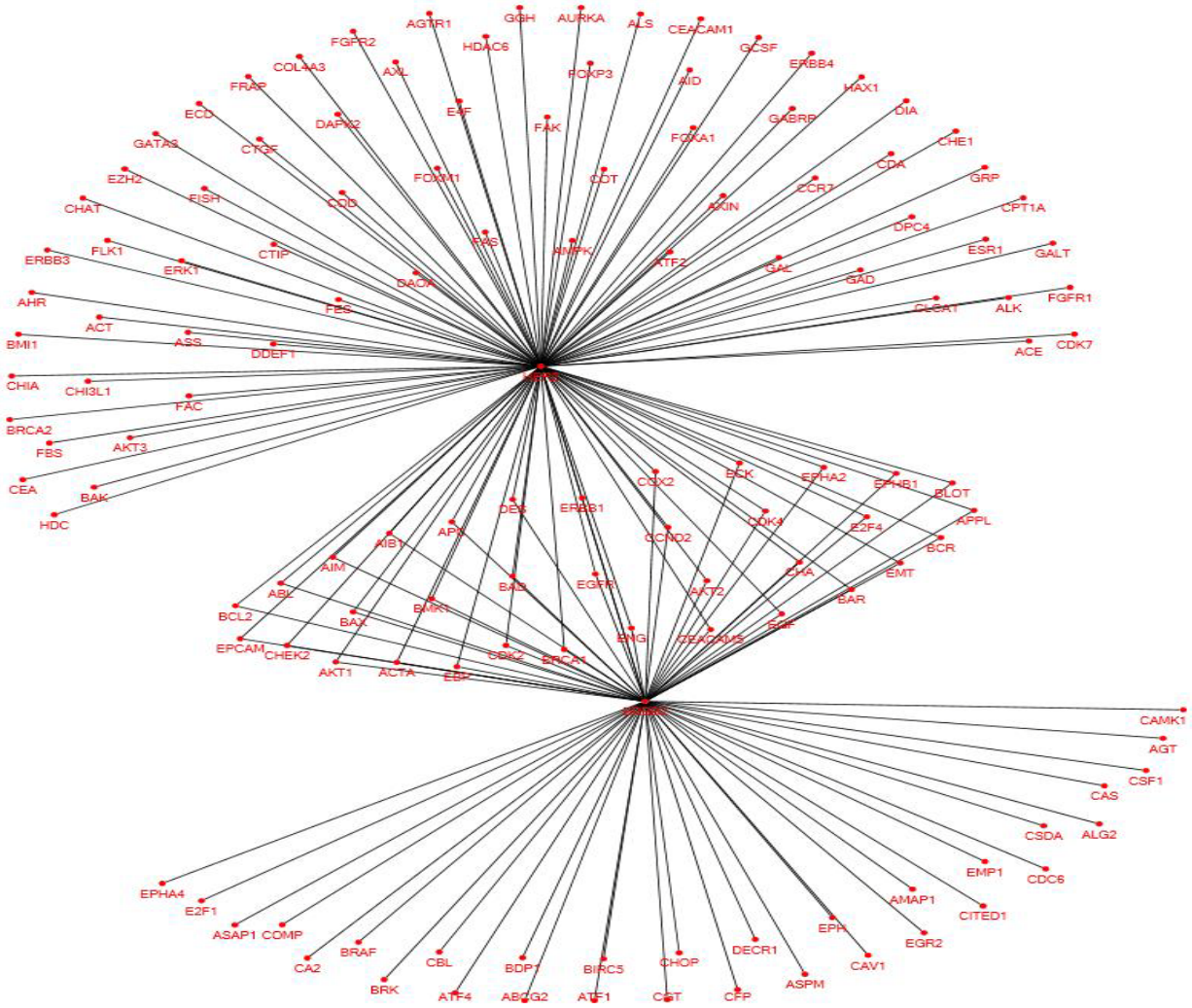


Fig. 4.9 Another visualization of ERBB2 and HER2 network – case of synonyms

Different types of networks can be obtained through visualization tools. A grid structure is used here to demonstrate the effectiveness of the text mining methods. Below is a grid network of interacting proteins for the query term "ERBB2 breast cancer" and the novel interactions identified using text mining are mapped onto this network. The goal is to back trace the hidden or novel associations using the grid network. Figure 4.10 shows the grid network of interacting pairs with novel associations marked in red. Biomarkers ERBB2 and HER2 are shown as solid red diamonds.

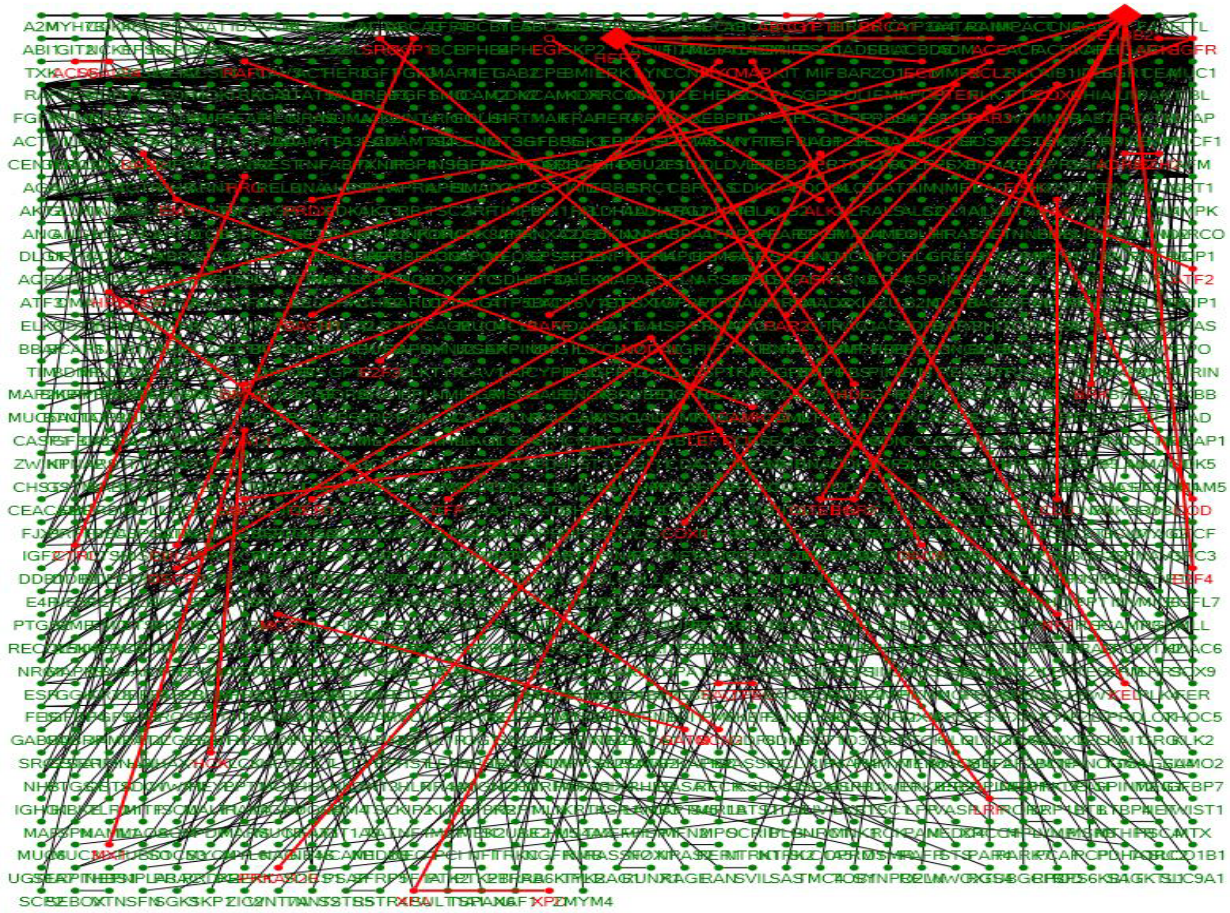


Fig. 4.10 Grid-network of explicit protein-protein interactions

In order to reduce the complexity of the network, a grid network of proteins with degree of freedom greater than or equal to 3 is shown in Figure 4.11.

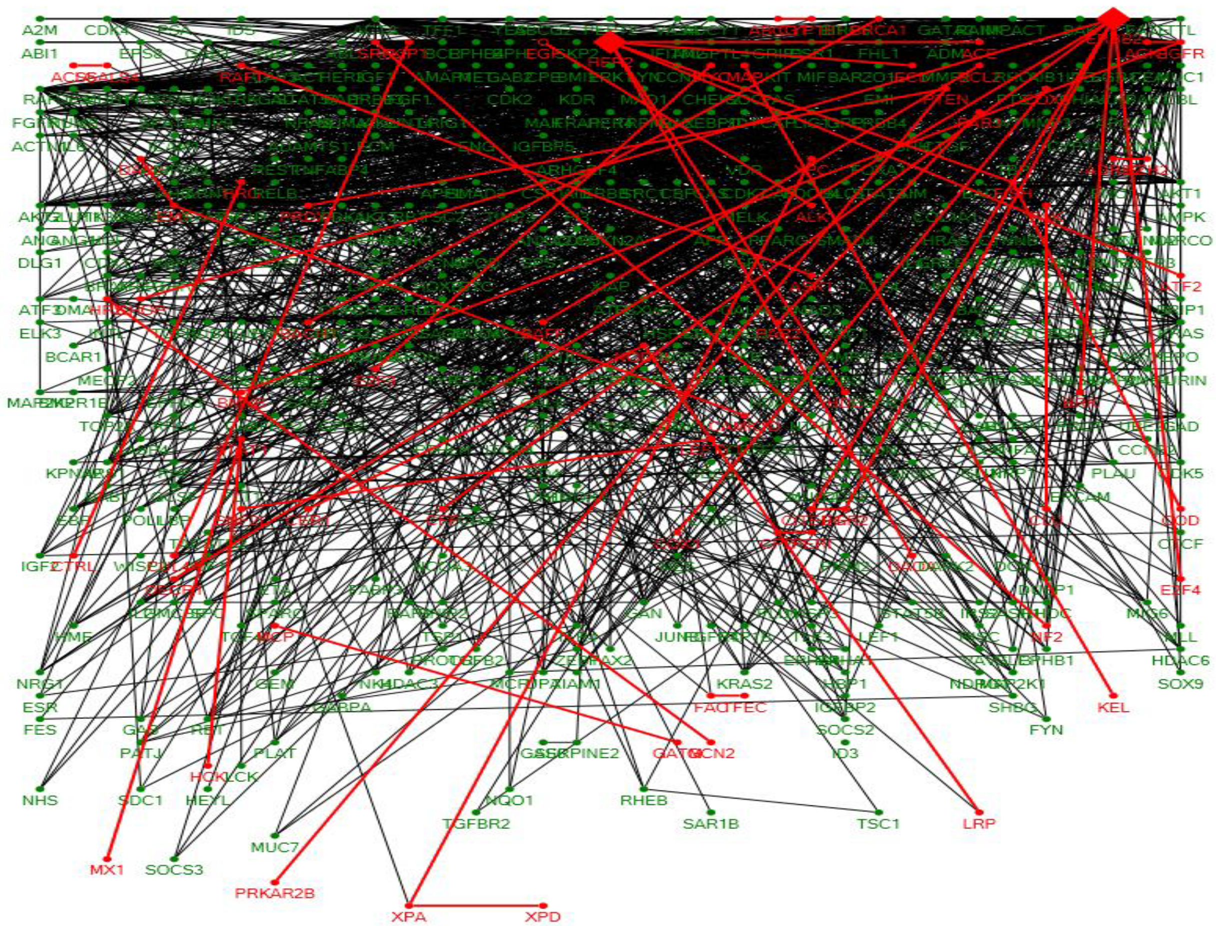


Fig. 4.11 Grid-network with reduced degree of freedom

A grid of novel association among proteins alone is shown in Figure 4.12. Upon this network, explicit interactions are back tracked step by step as shown in Figures 4.13, 4.14 and 4.15. Finally, a network of interest is identified and verified for its novelty.

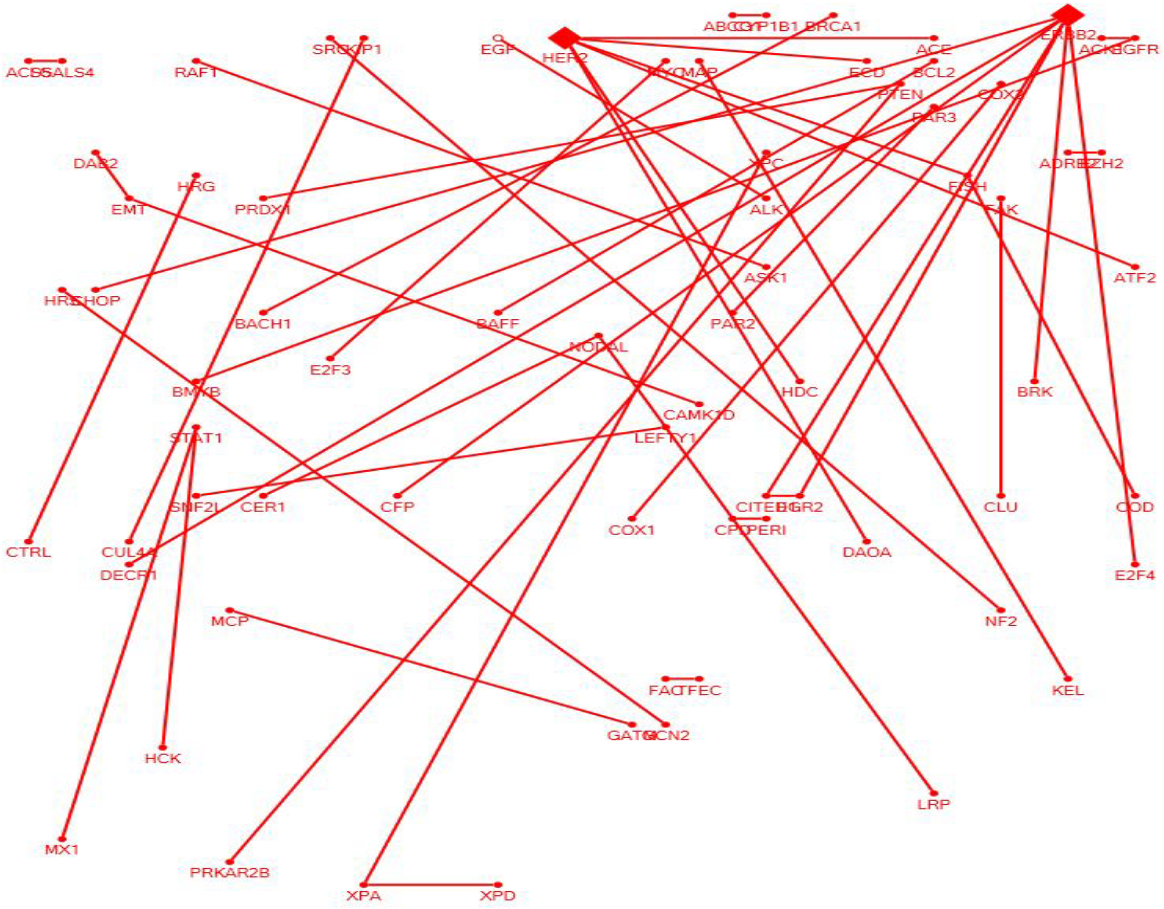


Fig. 4.12 Grid-network with only novel associations

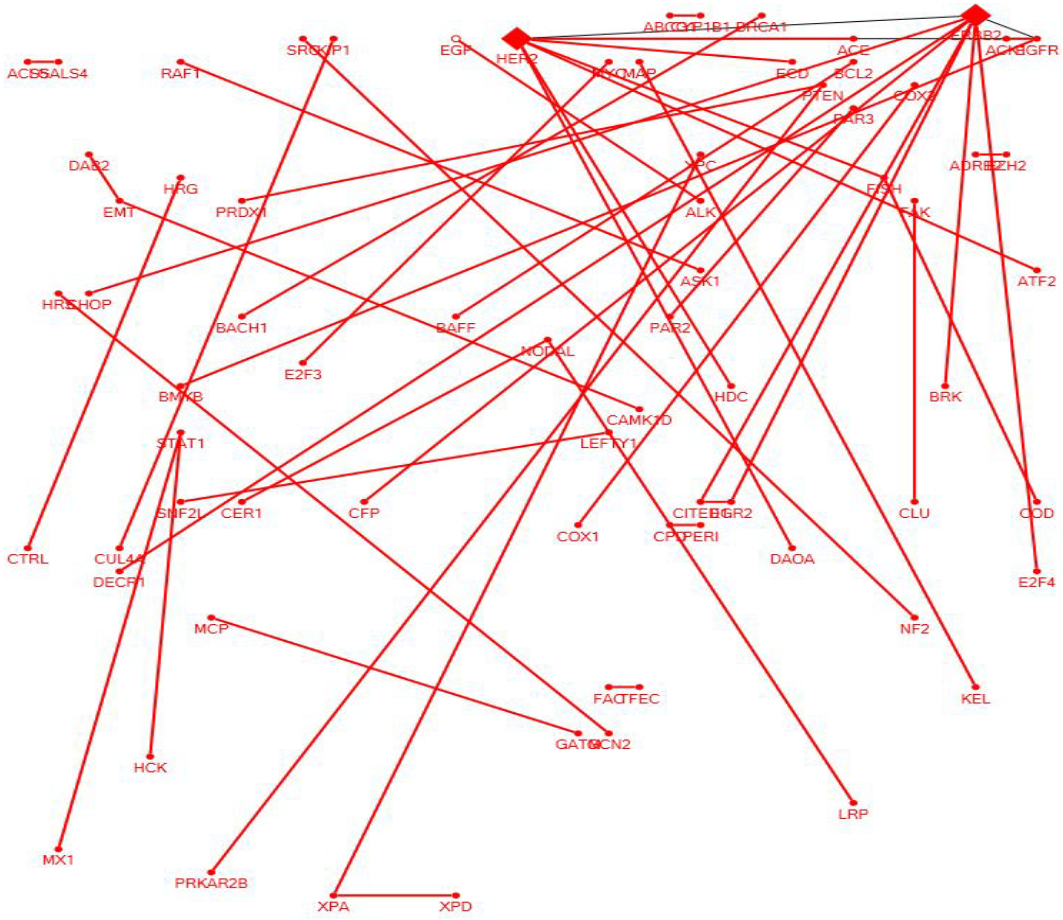


Fig. 4.13 Grid-network – tracing explicit associations 1

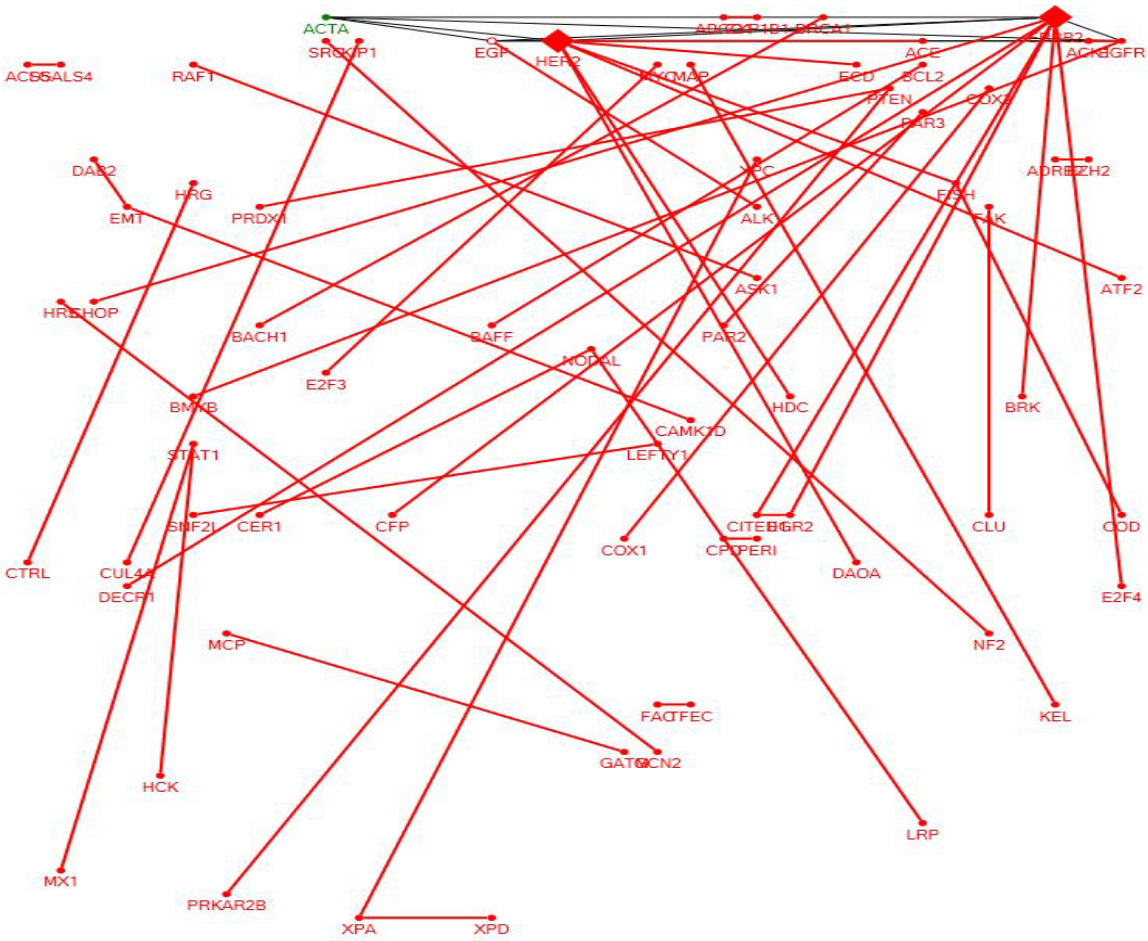


Fig. 4.14 Grid-network – tracing explicit associations 2

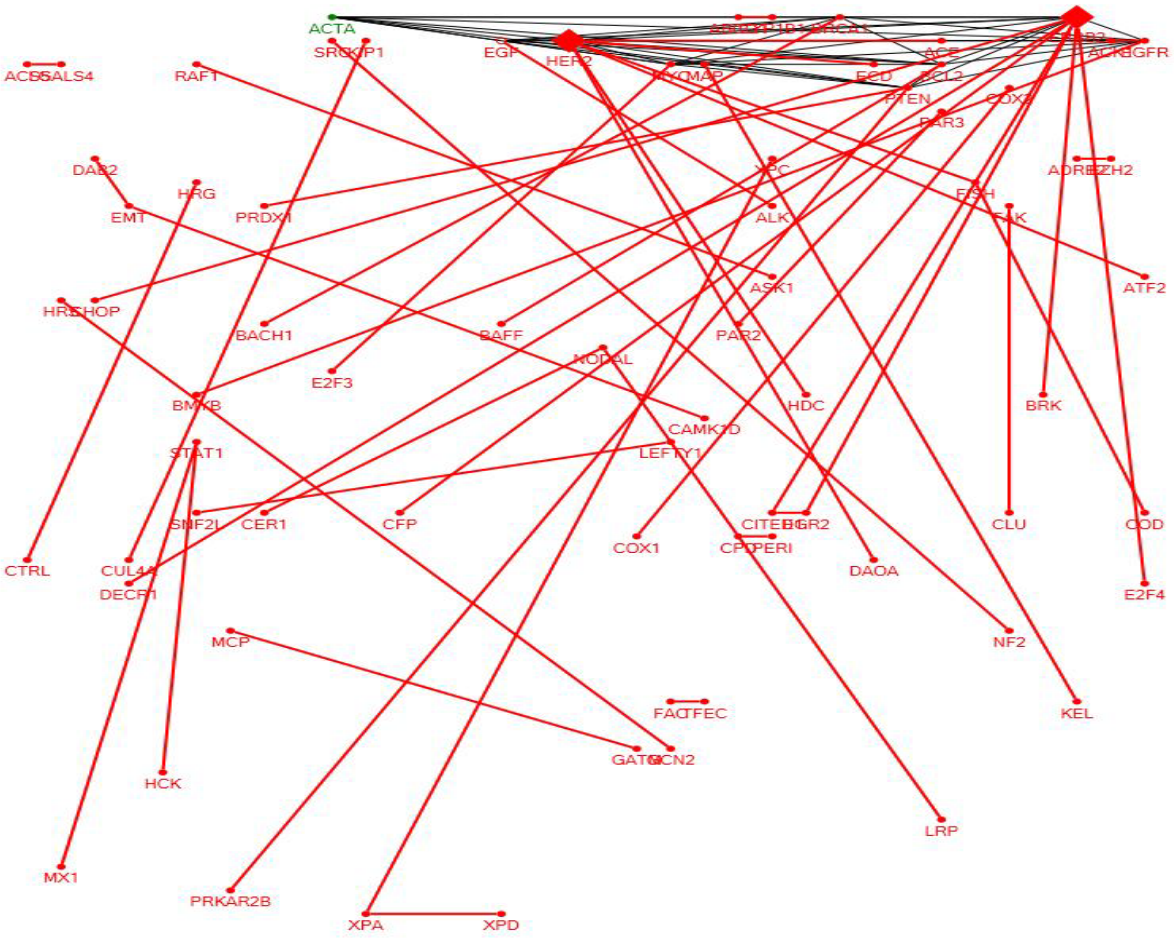


Fig. 4.15 Grid-network – tracing explicit associations 3

Following figures 4.16 and 4.17 are examples of types of networks of interested identified combining explicit and implicit associations among proteins of interest.

Fig. 4.16 Extracting network of interest 1

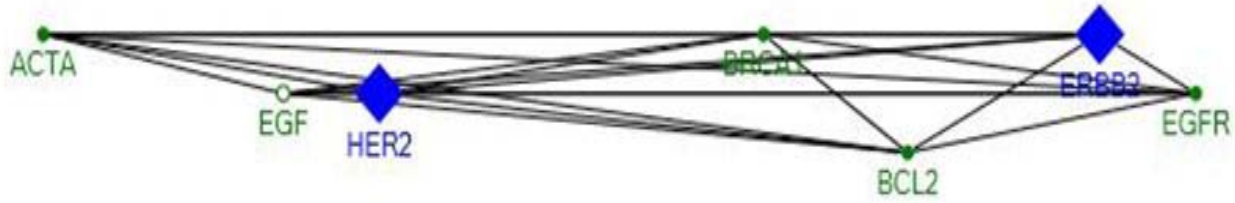


Fig. 4.17 Extracting network of interest 2

Another advantage of using visualization is to validate generated hypotheses using an extension of Swanson's ABC model [27]. Consider the generated hypothesis: ACE => HER2 as observed in Figure 4.14**.** Using Swanson's ABC model, it can be validated as follows:

ACE => AGTR2 [Proof: Entry in HPRD binary protein interactions database]

AGTR2 => ERBB3 [Proof: Entry in HPRD binary protein interactions database]

ERBB3 => ERBB2 [Proof: Entry in HPRD binary protein interactions database]

ERBB2 = HER2 [Proof: Synonym]

Hence, the novel hypothesis is valid.

Consider another example of a generated hypothesis: ACK1 => EGF as observed in Figure 4.15**.** Using Swanson's model, it can be validated as follows:

ACK1 = TNK2 [Proof: Synonym]

TNK2 => EGFR [Proof: Entry in HPRD binary protein interactions database]

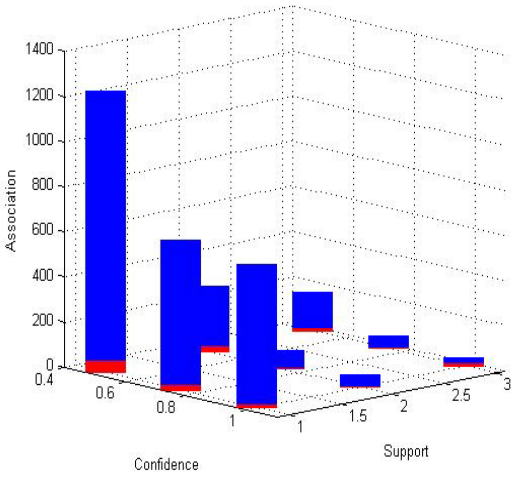
EGFR => EGF [Proof: Entry in HPRD binary protein interactions database]

Hence, the novel hypotheses are valid.

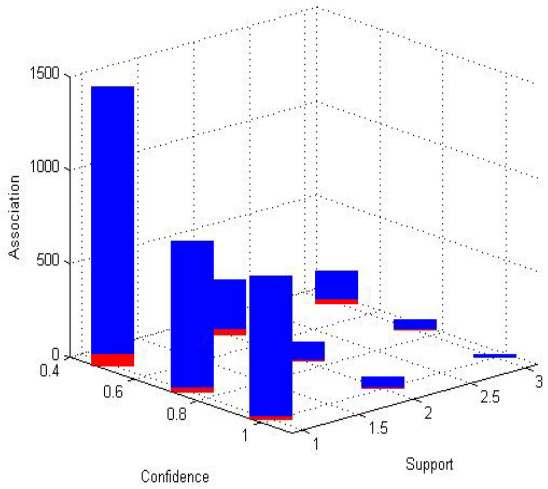
4.3 Discussions

Figures 4.18 and 4.19, generated from the above tables, show the variation in the number of unknown associations by varying support thresholds for transaction file generated for each query term. The red bar indicates the number of associations that are identified as already known interactions, with varying thresholds for support and confidence.

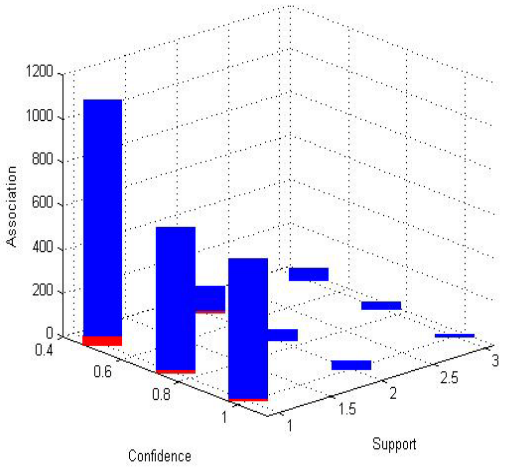
a) Database Used : PMC



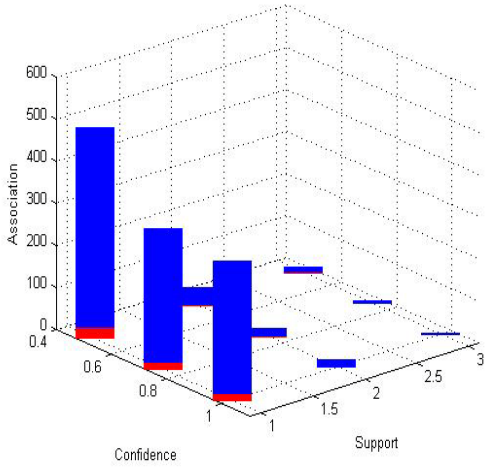
Support-confidence-number of associations for query term “protein”



Support-confidence-number of associations for query term “human protein”



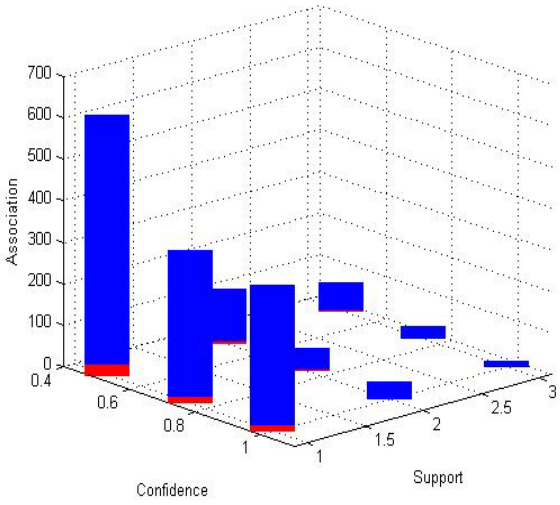
Support-confidence-number of associations for query term “ERBB2 breast cancer”



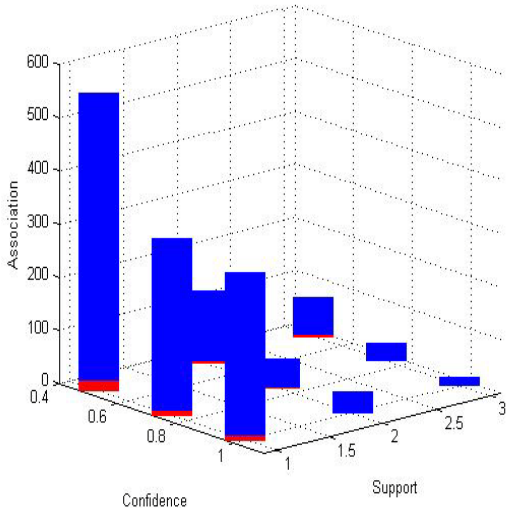
Support-confidence-number of associations for query term “IL2 signaling”

Fig. 4.18 Support-confidence-association rule plots for different query terms (PMC)

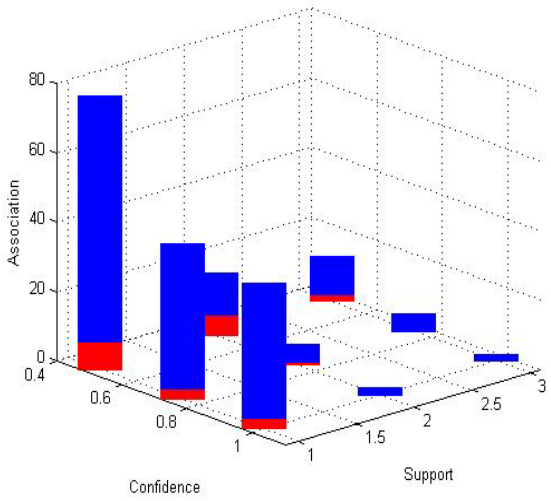
b) Database used : DOAJ



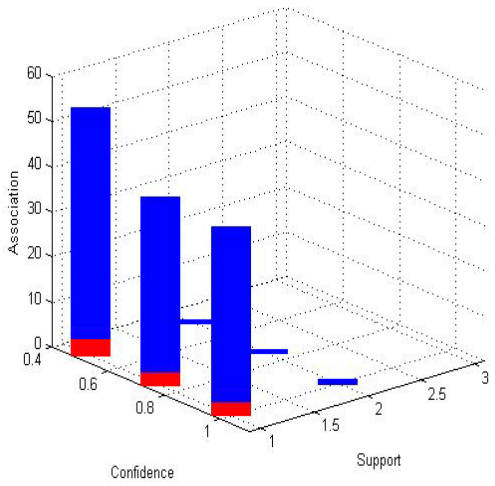
Support-confidence-number of associations for query term “protein”



Support-confidence-number of associations for query term “human protein”



Support-confidence-number of associations for query term “ERBB2 breast cancer”



Support-confidence-number of associations for query term “IL2 signaling”

Fig. 4.19 Support-confidence-association rule plots for different query terms (DOAJ)

TABLE 4.14

DECREASE IN NUMBER OF UNKNOWN ASSOCIATIONS WHEN MINSUP CHANGES FROM 1 TO 2

Query Term	Number of unknown associations			
	PMC		**DOAJ**	
	minsup = 1	**minsup = 2**	**minsup = 1**	**minsup = 2**
Protein	639	74	349	52
Human protein	775	90	321	53
ERBB2 breast cancer	655	53	42	5
IL2 signaling	317	19	36	1

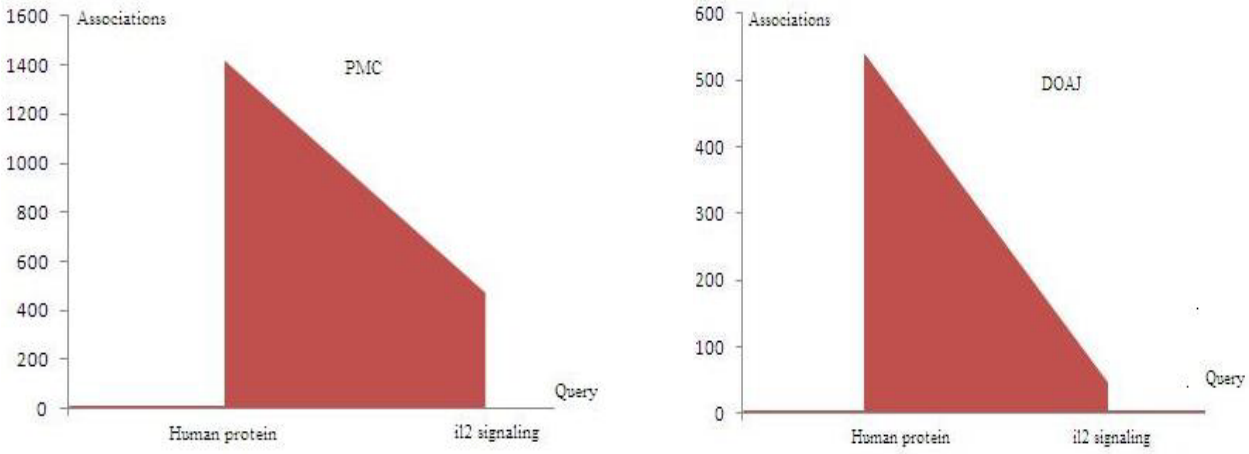


Fig. 4.20 Number of unknown associations for general query against specific query term

TABLE 4.15

DECREASE IN NUMBER OF UNKNOWN ASSOCIATIONS WHEN MINCONF CHANGES FROM 0.5 TO 0.75

Query Term	Number of unknown associations for minsup = 1			
	PMC		**DOAJ**	
	minconf = 0.5	**minconf = 0.75**	**minconf = 0.5**	**minconf = 0.75**
Protein	1195	639	596	349
Human protein	1419	775	540	321
ERBB2 breast cancer	1082	655	71	42
IL2 signaling	474	317	47	36

The following can be observed from the experiments and above graphs:

- The larger the corpus of articles to be mined, the bigger the number of unknown associations extracted. A significantly larger number of articles are retrieved from PMC than DOAJ, thereby having less number of unknown associations for DOAJ compared to PMC for the same query term. This seems advantageous as it provides a smaller set for clinical validation. However, there are no novel associations found for “IL2 signaling pathway” when articles are retrieved for higher support thresholds from DOAJ. Therefore, there is a trade-off in choosing the size of the corpus.
- The number of novel associations is very sensitive to changes in threshold of support than that of confidence. Table 4.14 shows there is an average of 10-fold decrease in number of unknown associations when threshold for support is increased from 1 to 2. This can be attributed to the large number of explicit associations (almost 70%, when considering articles from PMC) that have a support count equal to 1, i.e., such interactions are mentioned in only one article.

Hence, greater the support, explicit associations that are rarely mentioned may be avoided and thereby reducing the set of novel associations to a smaller size.

- There is a significant drop of nearly 50% (Figure 4.20) in the number of unknown associations when the query terms used are very specific to cellular pathways or diseases.
- There is almost 50% (Table 4.15) decrease in the number of novel associations when confidence is increased from 0.5 to 0.75. However, there is less or no prominent change in the number of such associations when confidence is increased from 0.75 to 1. This would seem a very crucial observation when users need to set confidence threshold for a pre-determined threshold of support.
- Although the number of known associations eliminated from the list of unknown associations seems to be small compared to the novel interactions generated, they prove to be benchmark for selecting optimal thresholds for support and confidence. Appropriate thresholds for support and confidence will enable to generate an optimal set of unknown interactions of manageable size for clinical validation and maximize the likelihood of at least some of them being biological significant.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

Based on the search terms provided at the C-Engine web portal, the crawler successfully downloads all articles in html, pdf or text formats from publication databases, such as PMC and DOAJ, to the local server. Effective identification of protein names in form of symbols or acronyms is made possible by employing a combination of strategies such as Fukuda rule sets, referencing a dictionary of protein names that is maintained locally, and elimination of words (non-proper nouns) that are present in English lexicon. Once the proteins are successfully identified from the text, manually devised templates are used for pattern matching along with a list of functional verbs, to extract binary protein-protein interactions from text. The list of explicit protein associations are used to create a transaction file that is used as input for association rule mining by apriori algorithm. Depending on the thresholds of support and confidence, the number of implicit associations varies. This list of hypotheses on associations can be reduced by eliminating known interactions from it. Since our objective is to elicit novel protein-protein interactions, the generally large set of implicit associations can be trimmed down by choosing an optimal value for minimum support and confidence, for a particular query term. Thus, a manageable set of novel relations are obtained, which can be validated in biology laboratories.

The following can be concluded from this research work:

- This thesis demonstrates the effectiveness of using full text articles, abstracts and captions from figures as opposed to other existing methods where only abstracts of articles are employed for explicit and implicit association discovery.

- It increases the likelihood that associations from the same or different sub-classes of research are discovered. All datasets, such as protein dictionary, function verbs list and binary protein interaction file, are able to identify maximum number of proteins and their interactions.
- From the experiments, it can be learnt that the number of implicit associations generated are more sensitive to changes in thresholds for support than confidence. Hence, it can be concluded that most of the query terms in experiments gave a manageable size of novel associations for combination of minimum support at 1 and minimum confidence at 0.75 and minimum support at 2 and minimum confidence at 0.5.
- Visualization techniques provide insights into nature of interactions that are usually not identified by tabular representation of data alone. Backtracking explicit associations from novel associations show the novelty of these implicit associations.

With subsequent laboratory validation of unknown associations from these optimal sets, it is possible that a hitherto anonymous relation could prove as an evidence for biomarker. This biomarker discovery would be very helpful to understand the effectiveness of a drug therapy for disease and so on.

5.2 Future Work

One important challenge in methodology, especially in entity identification, is the time taken to scan the entire corpus and recognize the protein names. And, it is not advisable to locally store multiple copies of the same article refereed by multiple queries. Also, it is a waste of space by storing multiple copies of locally tagged files, list of proteins identified and so on. A

solution could be using the history file to find out if that particular article has been stored, tagged its entities identified, relationships extracted or not.

Another possible extension of this work is to identify expansion of protein names along with symbols or acronyms for larger identification of entities and relationship extraction. Manually or automatically tagged corpora of full text articles could be used for text mining, and it helps in better evaluations of the entire system. A strict adherence to standardized nomenclature of protein names henceforth could ease the challenges in entity identification. Automatic template generation is suggested for extracting associations from sentences that have multiple phrases connecting the entities. A visualization tool would better help demonstrate pictorially the implicit associations from a network of explicit interactions.

Finally, it is the collective participation of biologists, computer linguists and database curators that would bring far-reaching changes in highly potential field of biomarker discovery and validation.

BIBLIOGRAPHY

[1] K. Chandramouli, P. Qian, “Proteomics: Challenges, Techniques and Possibilities to Overcome Biological Sample Complexity.” Human Genomics and Proteomics 2009;2009:1.DOI: 10.4061/2009/239204

[2] M. Krallinger, R. A. Erhardt, and A. Valencia, "Text-mining approaches in molecular biology and biomedicine." Drug Discov Today, vol. 10, no. 6, pp. 439-445, March 2005. [Online]. Available: http://dx.doi.org/10.1016/S1359-6446(05)03376-3

[3] G. Myers, "Whole-Genome DNA Sequencing," Computing in Science and Engineering, vol. 1, no. 3, pp. 33-43, May/June 1999, doi:10.1109/5992.764214

[4] Uniprot Documentation entries and gene designations for Arabidopsis thaliana - http://ca.expasy.org/cgi-bin/lists?arath.txt

[5] Saccharomyces Geneome Database – http://www.yeastgenome.org/

[6] Uniprot Documentation entries and gene designations for Homo Sapiens – http://www.ebi.ac.uk/uniprot/Documentation/

[7] R. Winnenburg, T. Wächter, C. Plake, A. Doms, M. Schroeder, "Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies?", Brief Bioinform. 2008 Nov;9(6):466-78. Epub 2008 Dec 6. Review.

[8] H. Shatkay, "Hairpins in bookstacks: Information retrieval from biomedical text," Brief Bioinform, vol. 6, no. 3, pp. 222-238, January 2005. [Online]. Available: http://dx.doi.org/10.1093/bib/6.3.222

[9] L. J. J. Jensen, J. Saric, and P. Bork, "Literature mining for the biologist: from information retrieval to biological discovery." Nature reviews. Genetics, vol. 7, no. 2, pp. 119-129, February 2006. [Online]. Available: http://dx.doi.org/10.1038/nrg1768

[10] H Pearson, "Genetics: what is a gene?". 2006, Nature 441 (7092): 398–401. doi:10.1038/441398a. PMID 16724031

[11] N. Rifai, M. A. Gillette, S. A. Carr, "Protein biomarker discovery and validation: the long and uncertain path to clinical utility", Nature Biotechnology 24, 971 - 983 (2006) Published online: 9 August 2006 | doi:10.1038/nbt1235

[12] Article on Biomarkers – http://www.biomarkersconsortium.org

[13] L. Hunter and K. B. Cohen, "Biomedical language processing: what's beyond pubmed?" Molecular cell, vol. 21, no. 5, pp. 589-594, March 2006. [Online]. Available: http://dx.doi.org/10.1016/j.molcel.2006.02.012

[14] PubMed Central system – http://www.ncbi.nlm.nih.gov/pmc/

[15] Directory of Open Access Journals search portal – http://www.doaj.org/doaj?func=searchArticles

[16] J. Wilbur, L. Smith, L. Tanabe, "BioCreative 2. Gene mention task." In: Proceedings of the Second BioCreative Challenge Evaluation Workshop. Madrid, Spain: Fundacion CNIO Carlos III, 2007;7–16.

[17] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," Brief Bioinform, vol. 6, no. 1, pp. 57-71, January 2005. [Online]. Available: http://dx.doi.org/10.1093/bib/6.1.57

[18] Gene Ontology Downloads – http://www.geneontology.org/GO.downloads.database.shtml

[19] HUGO Gene Nomenclature Committee – http://www.genenames.org/

[20] UMLS implementation Resources – http://130.14.16.110/research/umls/implementation_resources/index.html

[21] M. King, C. Lusk, G. Blobel, "Karyopherin-mediated import of integral inner nuclear membrane proteins." Nature 2006; 442:1003–7.

[22] I. Spasic, S. Ananiadou, J. McNaught, and A. Kumar, "Text mining and ontologies in biomedicine: Making sense of raw text," Brief Bioinform, vol. 6, no. 3, pp. 239-251, January 2005. [Online]. Available: http://dx.doi.org/10.1093/bib/6.3.239

[23] M. Krallinger and A. Valencia, "Text-mining and information-retrieval services for molecular biology," Genome Biology, vol. 6, no. 7, pp. 224+, 2005. [Online]. Available: http://dx.doi.org/10.1186/gb-2005-6-7-224

[24] H Yu, NM Luscombe, J Qian, M Gerstein, "Genomic analysis of gene expression relationships in transcriptional regulatory networks.", 2003 Trends Genet 19: 422-7.

[25] M. A. Hearst, "Untangling text data mining," in Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, vol. 43, no. 3. Morristown, NJ, USA: Association for Computational Linguistics, July 1999, pp. 3-10. [Online]. Available: http://dx.doi.org/10.3115/1034678.1034679

[26] M. Berardi , M. Lapi , P. Leo , C. Loglisci, "Mining generalized association rules on biomedical literature", Proceedings of the 18th international conference on Innovations in Applied Artificial Intelligence, p.500-509, June 22-24, 2005, Bari, Italy

[27] D. R. Swanson, "Complementary structures in disjoint science literatures," in SIGIR '91: Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, 1991, pp. 280-289. [Online]. Available: http://dx.doi.org/10.1145/122860.122889

[28] M. Weeber, R. Vos, H. Klein, L. T. De Jong-Van Den Berg, A. R. Aronson, and G.

Molema, "Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide." J Am. Med. Inform. Assoc., vol. 10, no. 3, pp. 252-259, 2003. [Online]. Available: http://dx.doi.org/10.1197/jamia.M1158

[29] D. R. Swanson, "Fish oil, raynaud's syndrome, and undiscovered public knowledge." Perspect Biol Med, vol. 30, no. 1, pp. 7-18, 1986. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/3797213

[30] D. R. Swanson, "Migraine and magnesium: eleven neglected connections." Perspect Biol Med, vol. 31, no. 4, pp. 526-557, 1988. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/3075738

[31] R. Agrawal, T. Imielinski, A. Swami, "Mining Association Rules Between Sets of Items in Large Databases", SIGMOD Conference 1993: 207-216

[32] M. Weeber, J. A. Kors, and B. Mons, "Online tools to support literature-based discovery in the life sciences," Brief Bioinform, vol. 6, no. 3, pp. 277-286, January 2005. [Online]. Available: http://dx.doi.org/10.1093/bib/6.3.277

[33] M. Huang, X. Zhu, Y. Hao, D. G. Payan, K. Qu, and M. Li, "Discovering patterns to extract protein-protein interactions from full texts." Bioinformatics, vol. 20, no. 18, pp. 3604-3612, December 2004. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/15284092

[34] S. K. Ng and M. Wong, "Toward routine automatic pathway discovery from on-line scientific text abstracts." Genome informatics. Workshop on Genome Informatics, vol. 10, pp. 104-112, 1999. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/11072347

[35] P. Srinivasan, "Text mining: Generating hypotheses from medline," Journal of the American Society for Information Science and Technology, vol. 55, no. 5, pp. 396-413, March 2003. [Online]. Available: http://dx.doi.org/10.1002/asi.10389

[36] G. Bhalotia, P. I. Nakov, A. S. Schwartz, and M. A. Hearst, "Biotext team report for the trec 2003 genomics track," in Proceedings of the Twelfth Text REtrieval Conference, 2003, pp. 612-621. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.10.9081

[37] W. Hersh & R. T. Bhuptiraju, TREC genomics track overview [online], http://trec.nist.gov/pubs/trec12/papers/GENOMICS.OVERVIEW3.pdf (2003).

[38] W. Hersh, R. Bhupatiraju, L. Ross, P. Roberts, A. Cohen, and D. Kraemer, "Enhancing access to the bibliome: the trec 2004 genomics track," Journal of Biomedical Discovery and Collaboration, vol. 1, pp. 3+, March 2006. [Online]. Available: http://dx.doi.org/10.1186%25252F1747-5333-1-3

[39] N. Stokes, Y. Li, L. Cavedon, and J. Zobel, "Exploring criteria for successful query expansion in the genomic domain," Information Retrieval. [Online]. Available: http://dx.doi.org/10.1007/s10791-008-9073-9

[40] R. Hoffmann, M. Krallinger, E. Andres, J. Tamames, C. Blaschke, and A. Valencia, "Text mining for metabolic pathways, signaling cascades, and protein networks." Sci STKE, vol. 2005, no. 283, May 2005. [Online]. Available: http://dx.doi.org/10.1126/stke.2832005pe21

[41] R. Hoffmann and A. Valencia, "A gene network for navigating the literature." Nat Genet, vol. 36, no. 7, July 2004. [Online]. Available: http://dx.doi.org/10.1038/ng0704-664

[42] S. Mukherjea, "Information retrieval and knowledge discovery utilising a biomedical semantic web," Brief Bioinform, vol. 6, no. 3, pp. 252-262, January 2005. [Online]. Available: http://dx.doi.org/10.1093/bib/6.3.252

[43] GoPubMed – http://www.gopubmed.org/web/gopubmed/

[44] L. Tanabe, U. Scherf, L. H. Smith, J. K. Lee, L. Hunter, and J. N. Weinstein, "Medminer: an internet text-mining tool for biomedical information, with application to gene expression profiling." Biotechniques, vol. 27, no. 6, December 1999. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/10631500

[45] E-BioSci –.http://www.e-biosci.org

[46] C. Perez-Iratxeta, P. Bork, and M. A. Andrade, "Xplormed: a tool for exploring medline abstracts." Trends Biochem Sci, vol. 26, no. 9, pp. 573-575, September 2001. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/11551795

[47] Pentz and Ed, "Crossref at the crossroads," Learned Publishing, vol. 19, no. 4, pp. 250-258, October 2006. [Online]. Available: http://dx.doi.org/10.1087/095315106778690760

[48] NPG Search – http://www.nonlin-processes-geophys.net/library_search.html

[49] I. Donaldson, J. Martin, B. de Bruijn, C. Wolting, V. Lay, B. Tuekam, S. Zhang, B. Baskin, G. Bader, K. Michalickova, T. Pawson, and C. Hogue, "Prebind and textomy - mining the biomedical literature for protein-protein interactions using a support vector machine," BMC Bioinformatics, vol. 4, no. 1, pp. 11+, March 2003. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-4-11

[50] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky, "Genies: a natural-language processing system for the extraction of molecular pathways from journal articles," Comput. Appl. Biosci., vol. 17, no. suppl_1, pp. S74-82, June 2001. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/17.suppl_1.S74

[51] W. A. Baumgartner Jr, Z. Lu, H. L. Johnson, J. G. Caporaso, J. Paquette, A. Lindemann, E. K. White, O. Medvedeva, K. B. Cohen and L. Hunter, "Concept recognition for extracting protein interaction relations from biomedical text," Genome Biology 2008

[52] B. de Bruijn and J. Martin, "Getting to the (c)ore of knowledge: mining biomedical literature." Int J Med Inform, vol. 67, no. 1-3, pp. 7-18, December 2002. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/12460628

[53] L. Tanabe and W. J. Wilbur, "Tagging gene and protein names in full text articles," in Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain. Morristown, NJ, USA: Association for Computational Linguistics, 2002, pp. 9-13. [Online]. Available: http://dx.doi.org/10.3115/1118149.1118151

[54] J. T. Chang, H. Schütze, and R. B. Altman, "Gapscore: finding gene and protein names one word at a time." Bioinformatics, vol. 20, no. 2, pp. 216-225, January 2004. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btg393

[55] G. Zhou , J. Zhang , J. Su , D. Shen , C. Tan, Recognizing names in biomedical texts: a machine learning approach, Bioinformatics, v.20 n.7, p.1178-1190, May 2004 [doi>10.1093/bioinformatics/bth060]

[56] IHOP – http://www.ihop-net.org/UniPub/iHOP/

[57] K. Fukuda, A. Tamura, T. Tsunoda, and T. Takagi, "Toward information extraction: identifying protein names from biological papers." Pac Symp. Biocomput., pp. 707-718, 1998. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/9697224

[58] N. Collier , C. Nobata , J. Tsujii, "Extracting the names of genes and gene products with a hidden Markov model," Proceedings of the 18th conference on Computational linguistics, p.201-207, July 31-August 04, 2000, Saarbrücken, Germany [doi>10.3115/990820.990850]

[59] S. Mika and B. Rost, "Protein names precisely peeled off free text." Bioinformatics, vol. 20 Suppl. 1, August 2004. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/bth904

[60] O. Miotto, T. W. W. Tan, and V. Brusic, "Supporting the curation of biological databases with reusable text mining." Genome Informatics. International Conference on Genome Informatics, vol. 16, no. 2, pp. 32-44, 2005. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/16901087

[61] J.-H. Chiang, H.-C. Yu, and H.-J. Hsu, "Gis: a biomedical text-mining system for gene information discovery," Bioinformatics, vol. 20, no. 1, pp. 120-121, January 2004. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btg369

[62] A. S. Reddy, D. S. Amarnath, R. S. Bapi, G. M. Sastry, and G. N. Sastry, "Protein ligand interaction database (plid)." Computational biology and chemistry, vol. 32, no. 5, pp. 387-390, October 2008. [Online]. Available: http://dx.doi.org/10.1016/j.compbiolchem.2008.03.017

[63] HUGO – http://www.hugo-international.org/aboutus.php

[64] HUGO--a UN for the human genome, Nat. Genet. 2003 Jun;34(2):115-6.

[65] H. Yu and E. Agichtein, "Extracting synonymous gene and protein terms from biological literature." Bioinformatics, vol. 19 Suppl 1, 2003. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btg1047

[66] K. Frantzi, S. Ananiadou, and H. Mima, "Automatic recognition of multi-word terms: the c-value/nc-value method," International Journal on Digital Libraries, vol. 3, no. 2, pp. 115-130, August 2000. [Online]. Available: http://dx.doi.org/10.1007/s007999900023

[67] P. Zweigenbaum, D. Demner-Fushman, H. Yu, and K. B. Cohen, "Frontiers of biomedical text mining: current progress," Brief Bioinform, vol. 8, no. 5, pp. 358-375, September 2007. [Online]. Available: http://dx.doi.org/10.1093/bib/bbm045

[68] C. Glymour, D. Madigan, D. Pregibon, and P. Smyth, "Statistical themes and lessons for data mining," Data Mining and Knowledge Discovery, vol. 1, no. 1, pp. 11-28, 1997. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.54.3561

[69] S. Raychaudhuri, H. Schütze, and R. B. Altman, "Using text analysis to identify functionally coherent gene groups." Genome Research, vol. 12, no. 10, pp. 1582-1590, October 2002. [Online]. Available: http://dx.doi.org/10.1101/gr.116402

[70] T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi, "Automated extraction of information on protein-protein interactions from the biological literature." Bioinformatics, vol. 17, no. 2, pp. 155-161, February 2001. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/17.2.155

[71] D. M. McDonald, H. Chen, H. Su, and B. B. Marshall, "Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser." Bioinformatics, vol. 20, no. 18, pp. 3370-3378, December 2004. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/15256411

[72] A. Rzhetsky, I. Iossifov, T. Koike, M. Krauthammer, P. Kra, M. Morris, H. Yu, P. A. A. Duboué, W. Weng, W. J. Wilbur, V. Hatzivassiloglou, and C. Friedman, "Geneways: a system for extracting, analyzing, visualizing, and integrating molecular pathway data." Journal of Biomedical Informatics, vol. 37, no. 1, pp. 43-53, February 2004. [Online]. Available: http://dx.doi.org/10.1016/j.jbi.2003.10.001

[73] J.-W. W. Fan and C. Friedman, "Semantic classification of biomedical concepts using distributional similarity." Journal of the American Medical Informatics Association: JAMIA, vol. 14, no. 4, pp. 467-477, 2007. [Online]. Available: http://dx.doi.org/10.1197/jamia.M2314

[74] M. Narayanaswamy, K. E. Ravikumar, and K. Vijay-Shanker, "Beyond the clause: extraction of phosphorylation information from medline abstracts." Bioinformatics, vol. 21 Suppl. 1, June 2005. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/15961474

[75] A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii, "Event extraction from biomedical papers using a full parser." Pac Symp. Biocomput., pp. 408-419, 2001. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/11262959

[76] D. Gildea and J. Hockenmaier, "Identifying semantic roles using combinatory categorial grammar," in Proceedings of the 2003 conference on Empirical methods in natural language processing. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 57-64. [Online]. Available: http://dx.doi.org/10.3115/1119355.1119363

[77] G. Leroy and H. Chen, "Filling preposition-based templates to capture information from medical abstracts." Pac Symp. Biocomput., pp. 350-361, 2002. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/11928489

[78] J. W. Cooper and A. Kershenbaum, "Discovery of protein-protein interactions using a combination of linguistic, statistical and graphical information." BMC Bioinformatics, vol. 6, no. 1, June 2005. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-6-143

[79] J. Ding, D. Berleant, D. Nettleton, and E. Wurtele, "Mining medline: abstracts, sentences, or phrases?" Pac Symp. Biocomput., pp. 326-337, 2002. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/11928487

[80] D. Hanisch, K. Fundel, H. T. Mevissen, R. Zimmer, and J. Fluck, "Prominer: rule-based protein and gene entity recognition," BMC Bioinformatics, vol. 6, no. Suppl 1, pp. S14+, 2005. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-6-S1-S14

[81] R. K. Lindsay and M. D. Gordon, "Literature-based discovery by lexical statistics," J. Am. Soc. Inf. Sci., vol. 50, no. 7, pp. 574-587, May 1999. [Online]. Available: http://dx.doi.org/10.1002/(SICI)1097-4571(1999)50:7%3C574::AID-ASI3%3E3.0.CO;2-Q

[82] A. Ramani, R. Bunescu, R. Mooney, and E. Marcotte, "Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome," Genome Biology, vol. 6, no. 5, pp. R40+, 2005. [Online]. Available: http://dx.doi.org/10.1186/gb-2005-6-5-r40

[83] J.-H. Chiang and H.-C. Yu, "Meke: discovering the functions of gene products from biomedical literature via sentence alignment," Bioinformatics, vol. 19, no. 11, pp. 1417-1422, July 2003. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btg160

[84] P. Srinivasan and M. Wedemeyer, "Mining concept profiles with the vector model or where on earth are diseases being studied," 2003. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.10.3011

[85] E. Eskin and E. Agichtein, "Combining text mining and sequence analysis to discover protein functional regions." Pac Symp. Biocomput., pp. 288-299, 2004. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/14992511

[86] D. Scutt, G. A. Lancaster, and J. T. Manning, "Breast asymmetry and predisposition to breast cancer." Breast Cancer Res., vol. 8, no. 2, 2006. [Online]. Available: http://dx.doi.org/10.1186/bcr1388

[87] N. Collier, A. Nazarenko, R. Baud, and P. Ruch, "Recent advances in natural language processing for biomedical applications." International Journal of Medical Informatics, vol. 75, no. 6, pp. 413-417, June 2006. [Online]. Available: http://dx.doi.org/10.1016/j.ijmedinf.2005.06.008

[88] M. Weeber, R. Vos, H. Klein, L. T. De Jong-Van Den Berg, A. R. Aronson, and G. Molema, "Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide." J. Am. Med. Inform. Assoc., vol. 10, no. 3, pp. 252-259, 2003. [Online]. Available: http://dx.doi.org/10.1197/jamia.M1158

[89] P. Srinivasan and B. Libbus, "Mining medline for implicit links between dietary substances and diseases." Bioinformatics, vol. 20 Suppl. 1, August 2004. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/15262811

[90] Arrowsmith – http://arrowsmith.psych.uic.edu/arrowsmith_uic/

[91] M. D. Gordon and R. K. Lindsay, "Toward discovery support systems: a replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between raynaud's and fish oil," J. Am. Soc. Inf. Sci., vol. 47, no. 2, pp. 116-128, February 1996. [Online]. Available: http://dx.doi.org/10.1002/(SICI)1097-4571(199602)47:2%3C116::AID-ASI3%3E3.3.CO;2-P

[92] M. Yetisgen-Yildiz and W. Pratt, "Evaluation of literature-based discovery systems," 2008, pp. 101-113. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-68690-3_7

[93] B. Boeckmann, A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider, "The swiss-prot protein knowledgebase and its supplement trembl in 2003." Nucleic Acids Res., vol. 31, no. 1, pp. 365-370, January 2003. [Online]. Available: http://dx.doi.org/10.1093/nar/gkg095

[94] C. Blaschke, E. A. Leon, M. Krallinger, and A. Valencia, "Evaluation of biocreative assessment of task 2." BMC Bioinformatics, vol. 6 Suppl. 1, 2005. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-6-S1-S16

[95] K. Van Auken, J. Jaffery, J. Chan, H.-M. M. Müller, and P. W. Sternberg, "Semi-automated curation of protein subcellular localization: a text mining-based approach to gene ontology (go) cellular component curation." BMC Bioinformatics, vol. 10, no. 1, pp. 228+, July 2009. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-10-228

[96] F. M. Couto, M. J. Silva, V. Lee, E. Dimmer, E. Camon, R. Apweiler, H. Kirsch, and D. Rebholz-Schuhmann, "Goannotator: linking protein go annotations to evidence text," Journal of Biomedical Discovery and Collaboration, vol. 1, pp. 19+, December 2006.

[Online]. Available: http://dx.doi.org/10.1186/1747-5333-1-19

[97] M. Scherf, A. Epple, and T. Werner, "The next generation of literature analysis: integration of genomic analysis into text mining." Brief Bioinform, vol. 6, no. 3, pp. 287-297, September 2005. [Online]. Available: http://dx.doi.org/10.1093/bib/6.3.287

[98] M. D. Yandell and W. H. Majoros, "Genomics and natural language processing," Nat. Rev. Genet., vol. 3, no. 8, pp. 601-610, 2002. [Online]. Available: http://dx.doi.org/10.1038/nrg861

[99] PolySearch – http://wishart.biology.ualberta.ca/polysearch/

[100] EBIMed – http://www.ebi.ac.uk/Rebholz-srv/ebimed/

[101] GoPubMed – http://www.gopubmed.org/web/gopubmed/

[102] Google Scholar – http://scholar.google.com/

[103] iProLink – http://pir.georgetown.edu/iprolink/

[104] PubGene – http://www.pubgene.org/

[105] ABNER – http://pages.cs.wisc.edu/~bsettles/abner/

[106] AliasServer – http://cbi.labri.fr/outils/alias/

[107] Abbreviation Server – http://bionlp.stanford.edu/abbreviation/

[108] LitLinker – http://litlinker.ischool.washington.edu/

[109] Manjal – http://sulu.info-science.uiowa.edu/Manjal.html/

[110] IRIDESCENT – http://www.etexxbio.com/

[111] TXTGATE – http://tomcat.esat.kuleuven.be/txtgate/

[112] K. G. Becker, D. A. Hosack, G. Dennis, R. A. Lempicki, T. J. Bright, C. Cheadle, and J. Engel, "Pubmatrix: a tool for multiplex literature mining." BMC Bioinformatics, vol. 4, December 2003. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-4-61

[113] Textpresso – http://www.textpresso.org/

[114] C.D. Manning, P.Raghavan, and H.SchutzeDing, Introduction to Information Retrieval,Cambridge University,Press. 2008, ISBN: 521865719

[115] Googlebot – http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=80553

[116] Msnbotl – http://help.live.com/help.aspx?mkt=en-us&project=wl_webmasters

[117] Yahoo! Slurp – http://help.yahoo.com/l/us/yahoo/search/webcrawler/

[118] LWP Package – http://search.cpan.org/dist/libwww-perl/

[119] M. Krauthammer, “Advanced Literature Mining through Image processing and Analysis,” 2009-2012, Grant 1R01LM009956-01A1, National Library of Medicine

[120] Lingua::EN::Tagger – http://search.cpan.org/~ACOBURN/Lingua-EN-Tagger/Tagger.pm

[121] D. Zhou and Y. He, "Extracting interactions between proteins from the literature." Journal of Biomedical Informatics, vol. 41, no. 2, pp. 393-407, April 2008. [Online]. Available: http://dx.doi.org/10.1016/j.jbi.2007.11.008

[122] E. Charniak, "Statistical Techniques for Natural Language Parsing". 1997, AI Magazine 18(4):33–44

[123] D. Jurafsky, and J. H. Martin, “Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition,” Prentice Hall, Second Edition

[124] Penn Treebank project – http://www.cis.upenn.edu/~treebank/

[125] T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi, "Automated extraction of information on protein-protein interactions from the biological literature." Bioinformatics, vol. 17, no. 2, pp. 155-161, February 2001. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/17.2.155

[126] H. Seitz, M.Werther, “Protein-Protein Interaction.” 2008, Springer, ISBN: 354068817X

[127] Reactome – http://www.reactome.org/

[128] HPRD – http://www.hprd.org/

[129] Porter Stemmer – http://tartarus.org/~martin/PorterStemmer/

[130] R. Kabiljo, A. B. Clegg, and A. J. Shepherd, "A realistic assessment of methods for extracting gene/protein interactions from free text." BMC Bioinformatics, vol. 10, no. 1, pp. 233+, July 2009. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-10-233

[131] NodeXL – http://nodexl.codeplex.com/