In: Internet Policies and Issues, Volume 7
Editor: B.G. Kutais
ISBN: 978-1-61668-745-8

Chapter 11

BIOMEDICAL LITERATURE ANALYSIS: CURRENT STATE AND CHALLENGES

Maurice HT Ling[1, 2], Christophe Lefevre[3] and Kevin R. Nicholas[2, 3]
[1]School of Chemical and Life Sciences, Singapore Polytechnic, Singapore
[2]Department of Zoology, The University of Melbourne, Australia
[3]Institute for Technology Research and Innovation, Deakin University, Australia

ABSTRACT

Advances in molecular biology tools and techniques from the end of the last century had shifted the focus of biomedical research from the study of individual proteins and genes to the interactions within an entire biological systems. At the same time, advanced tools generates large sets of experimental data which required collaborations of groups of biologists to decipher. This resulted in a need to have a diverse research knowledge. However, the amount of published research information in the form of published articles is increasing exponentially, making it difficult to maintain a productive edge. Biomedical literature analysis is seen as a means to manage the increased amount of information – to gather relevant articles and extract relevant information from these articles. We review the central (information retrieval, information extraction and text mining) and allied (corpus collection, databases and system evaluation methods) domains of computational biomedical literature analysis to present the current state of biomedical literature analysis for protein-protein and protein-gene interactions and the challenges ahead.

1. INTRODUCTION

With rising emphasis in genomics, transcriptomics and proteomics from the end of the last century, the focus of biomedical research is shifting from the study of individual proteins and genes to entire biological systems, such as tissues or whole organisms. Experimental techniques used to study them, like mass spectrometry and microarrays, often generates large data sets. A group of biologists must then collaborate to make sense of this large set of

experimental data, and often requires connections with research areas outside their own core competencies, which exists as published literature in various research areas. This has resulted in a need to be versed in research areas other then the researcher's own specialty. In addition, the amount of information, in the form of published articles, is increasing exponentially, making it difficult for a researcher to keep abreast with relevant literature manually [1], even on specialized topics.

Due to these changes, literature processing tools are becoming essential to researchers [2] as it was estimated that only about 20% of biological knowledge exist in structured formats, such as in databases, while the remaining 80% are in natural language document [1, 3]. They include targeting relevant papers, known as information retrieval; identifying gene or protein or chemical entities; identifying abbreviations; extracting facts from the literature, known as information extraction; and in same instances, generating hypotheses. This review shall briefly examine the historical roots and current state of biomedical literature analysis. The computational procedures of 30 systems will be briefly described to illustrate some of the current methods used, and its related areas of importance before defining the objectives and organization of this thesis.

2. BRIEF HISTORY OF BIOMEDICAL LITERATURE ANALYSIS

Don Swanson initiated interest in biomedical literature analysis by analyzing publications semi-automatically and suggested links between separate areas of research, such as fish oil and Raynaud's syndrome [4], migraine and magnesium [5] in the mid-1980s.

At around the same time, the First Message Understanding Conference (MUC-1) was held in 1987, which explored formats for recording information in documents. In 1989, MUC-2 concentrated on template filling of information and formulated the details of precision and recall measures, which is still in use today. MUC-3 (1991) and MUC-4 (1992) were centered on compiling and completing template from information in terrorist reports, and therefore, had no direct bioinformatics relevance but benefited improved techniques.

In 1992, the First Text Retrieval Conference (TREC-1) was initiated by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense and used the idea of challenge evaluation tasks to tease out the state of the art of that time. Both TREC-1 and MUC-5 in 1993 were greatly influenced by the Tipster program (a U.S. Government program) which emphasized on evaluation-driven research [6], setting the tone for future conferences. TREC-2 in 1993 was critical for providing the baseline performance for main tasks, which was then expanded to having different tracks for various tasks in future TRECs. Each track in TREC was started based on interests and notably, a genomic track was started in 2003 [7] with a subsequent TREC in 2006. In short, MUC and TREC conferences had significantly advanced the field of information retrieval and extraction by their challenge tasks due to rigorous use of systematic common evaluations [8].

The earliest work in text mining for genomics was by Timothy Leek [9]. Fukuda et al. [10] pioneered protein name recognition (named entity recognition) in text in the 3rd Pacific Symposium on Biocomputing. Craven and Kumlein [11] and Blaschke et al. [12] independently published the first work on recognition of relationships between entities (proteins, genes, and small molecules). By 2000, the focus had shifted to the recognition of

relationships between entities (proteins, genes, and small molecules), with Shatkay and Wilbur [13] and GENIES [14] as one of the first systems. Following GENIES, the field of biomedical literature analysis for information retrieval and information extraction was extremely active with numerous systems being developed (reviewed in later sections). The Pacific Symposium on Biocomputing between the years 2001 and 2004 included predominantly presentations on various aspects of biomedical literature analysis and other related conferences, such as Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), saw an increase in the number of similar presentations. Specific workshops for biomedical literature analysis, such as BioLink 2004, were run in that period. It could be said that the years between 1998 and 2004 were the golden age of biomedical literature analysis.

By the end of 2004, the general emphasis of biomedical literature analysis had diminished at these conferences. This might be due to the asymptotic performance of various systems and a serious lack of benchmark data, such as biomedical corpora tagged for various purposes [15]. This might also meant that conventional means of analysis and technology transfer from more traditional fields, such as computational linguistics for understanding general text, had reached its maximum. The changes were more likely due to the rising interest in other fields, such as microarray and sequence analyses. The Fourth Asia Pacific Bioinformatics Conference (APBC 2006) saw more than half of the posters in the area of microarray and sequence analyses while only about 5% in biomedical literature analysis. However, the golden age of biomedical literature analysis of 1998 and 2004 had left us with preliminary resources and techniques that could be collated for other purposes.

Interestingly, although the field of biomedical literature analysis arose from Don Swanson's work in the mid-1980s [4, 5], hypothesis generation (text mining) was not adopted into mainstream biomedical literature analysis. This suggested that the field of biomedical knowledge is still largely uncharted with gems for many explorers to find in the years to come. With that optimistic thought, we shall explore the current areas of research in biomedical literature analysis.

3. CURRENT AREAS OF RESEARCH

Although the primary utility of biomedical literature processing is obtaining the relevant research articles (information retrieval; IR), extracting facts (information extraction; IE), and in some instances, drawing new hypotheses (text mining; TM) as shown in Figure 1, there are five concentrations of research efforts instead of the mentioned three. The other two are Named Entity Recognition (NER) and Abbreviation Recognition (AR), which are more domain specific [2]. Comparatively, IR, IE and TM tend to be less domain-specific.

Before focusing on each of the core research areas, it is crucial to understand some commonly used performance measures. Systems are typically measured in terms of precision (number of correct predictions divided by the total number of predictions) and recall (number of correct predictions divide by the total number of correct predictions in the test set) [2]. Precision (P) and recall (R) can be combined into a single F-score, defined as the harmonic mean of precision and recall, 2PR/(P+R) [16, 17]. A more extensive treatment of evaluation strategies will be given in Section 1.8.3.

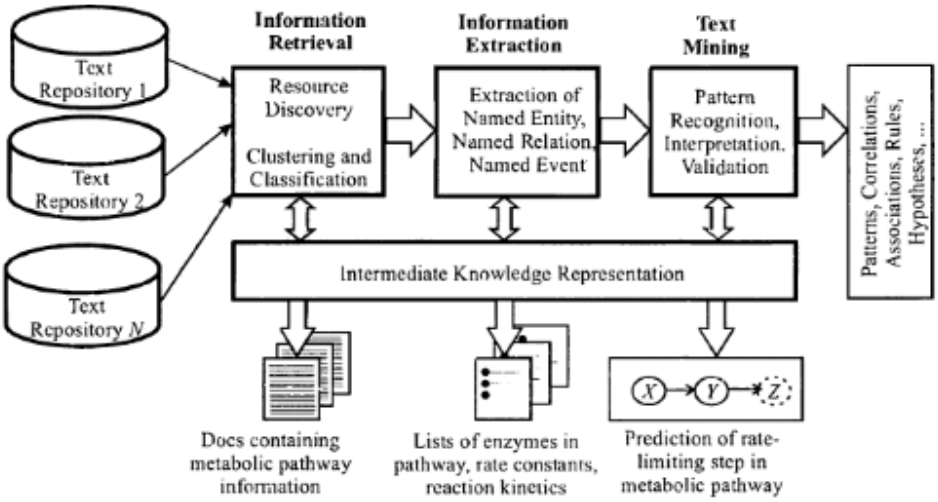

Figure 1. Stages of biomedical literature processing showing examples of possible output from each stage. Adapted from [16].

3.1. Information Retrieval: Finding the Papers

Information retrieval (IR) tools aim to identify text (full text, abstracts, sentences) pertaining to a certain topic of interest, which might be user-defined (ad hoc IR) or pre-defined categories, as in text categorization [18], or a hybrid of both [19]. The best-known biomedical IR system, PubMed, is an ad hoc IR system, which uses Boolean logic and a vector space model, and is available as a programmatic interface, PubMed EUtils.

Both Boolean logic and the vector space model are established IR methodologies [20]. Boolean logic builds on keyword queries where multiple keywords are chained using Boolean operators, such as "(mouse AND cytokines) NOT interferon". In the vector space model, each document is represented by document-term vector, calculated by a frequency-based weighing scheme, which is then compared to the query vector [21, 22]. This had also been used for thematic analysis of the AIDS literature [23].

Building on PubMed's IR system, more advanced tools like Textpresso [24] and MedMiner [25] employed named entity recognition (see Section 3.3) to identify text for protein and gene names. Shatkay et al. [26] used named entity and terms recognition before presenting the results to a machine-learned classifier to score biomedical text. Results from an IR system are usually presented as a long list which gives a poor overview. Based on the idea of index creation, words that are not commonly used, also known as discriminatory words, had been used by MedlineRanker [27] to rank retrieved citations. Recent developments had attempted to either summarize search results into other representations using Gene Ontology [28, 29] or to represent pair-wise searches as network [30], as it had been shown that biomedical ontologies can significantly enhance the quality of document clustering [31] which can be used as input for text mining. Biological ontologies had also been used to restrict the amount of returned documents in PathBinderH [32] and to assist in document navigation [33]. Query term matching [21] and expansion [34] is an important aspect and a current obstacle in biomedical IR due to multiple synonyms describing the same entity. For

example, 'yeast', also known as 'baker's yeast', can be both 'Saccharomyces cerevisiae' or 'Schizosaccharomyces pombe' or their abbreviated forms, 'S. cerevisiae' and 'S. pombe' respectively. Although matching of query terms to biomedical vocabularies like Medical Subject Headings [30], UMLS [35, 36] and OWL-DL ontology [37] is possible, the query term expansion is still largely an open problem [38].

3.1.1. Brief Descriptions of information retrieval systems

MedMiner [25] interrogates PubMed and GeneCard. User query of GeneCard database obtained a list of associated genes, which could be filtered with a user defined list of genes of interest and used to query PubMed. Results from PubMed were analyzed at sentence-level and only sentences with at least one gene and keyword remained. Final results were presented to the user after clustering based on rules or keywords.

Textpresso [24] was an example of a system able to extract multiple types of information from biomedical literature [38]. Text was POS tagged with Brill tagger [39] trained on C. elegans literature but was not further used. Instead, the same text was tagged by a set of 33 tags, forming the Textpresso Ontology, consisting of 14500 Regular Expressions. Exhaustive indexing on the ontology was carried out to facilitate text retrieval. A further work [19] expanded on Textpresso by automatic clustering of the results into document categories using phrase clustering and support vector machines. Pharmspresso [40] was also built on Textpresso [24] for studying the relationship between genetic variation and the variation in drug response phenotypes. When tested on 45 human-curated abstracts, Pharmspresso [40] identified 78% of target genes, 61% of target polymorphisms, and 74% of target drug concepts.

PathBinderH [32] intersected PubMed search results and NCBI Entrez Taxonomy. Both search results and taxonomical terms were user-defined. Only results that were indexed in the correct taxonomy were returned to the user.

Shatkay et al. [26] used MedPost [41] and YamCha [42] to identify terms in the source text to train a classifier to identify the required articles.

3.2. Abbreviation Recognition: Aliasing the Names

Growth in biomedical terminology parallels the growth of biomedical literature and many of these biomedical terminologies have abbreviations. There are two foreseeable uses in collecting terminologies and their abbreviations. Firstly, abbreviations can aid information retrieval by providing a means to expand search terms to cover terminological variants that have the same abbreviation. Secondly, a dictionary of terminologies and abbreviation may facilitate text processing of multi-word terms as described in the previous section. Abbreviation Recognition (AR) is pair recognition of a terminology (may be a phrase or an entity) and its corresponding abbreviation from free text.

Most of the progress in AR was made between the years 2001 and 2004, and had mostly adopted rules-based techniques with a statistical score. Liu and Friedman [43] were probably the only group that used mainly statistical co-location to determine abbreviations and their phrases. The main drawback of AR by statistics is the need for a large collection of text, known as a corpus (corpora for plural).

Rules-based methods used the knowledge of how abbreviations were being formed and had been well described by Jeffrey Chang (Stanford University's Abbreviation Server) [44, 45], which demonstrated 97% precision at 22% recall (F-score = 0.36) and 95% precision at 75% recall (F-score = 0.88). AbbRE [46] and Schwartz and Hearst [47] both used pattern matching rules and achieved 96% precision with 70% recall (F-score = 0.81), and 96% precision with 82% recall respectively (F-score = 0.88). SaRAD system [48] reported 95% precision with 85% recall (F-score = 0.9). AcroMed [49] reported 97% precision with 72% recall (F-score = 0.83) while ARGH [50] reported 96% precision with 93% recall (F-score = 0.94). The results (abbreviations and their full form) of Stanford University's Abbreviation Server, SaRAD, AcroMed and ARGH were available [51]. Acromine [52] used the observation that abbreviations can usually be expanded to their full form (reversing the idea of reducing full forms into abbreviations) reported precision of 99% with 82% to 95% recall (F-score = 0.89 – 0.97). Sohn et al. [53] uses a similar long-form to abbreviation matching algorithm to AbbRE [46] reported 96.5% precision with 83.2% recall. MBA [54] used text alignment for acronym-type abbreviations and statistics for non-acryonym-type abbreviations. It reported 88% recall and 91% precision (F-score = 0.89). BIOADI [55] exploited a set of textual features to describe the properties of potential abbreviation pairs and reported 96% precision with 85% recall (F-score = 0.90). Torii et al. [56] performed a meta-study to compare the results of a number of AR systems and found that they generally agree with each other in terms of results. In addition, Kuo et al. [55] evaluated Sohn et al. [53], Schwartz and Hearst [47] and BIOADI [55] using 2 different corpora and found that the performance of the systems agree with each other in terms of precision, recall and F-score.

3.3. Named Entity Recognition: Identifying the Players

The goal of Named Entity Recognition (NER) is to find entities, the names of physical or abstract objects, in a given text [57]; in particular, the names of chemicals, proteins and genes. Essentially, NER asks the question “What makes a name a name?” This question is generally answered by recognizing words that may refer to entities, followed by identifying the entities in question uniquely. NER is currently one of the most difficult tasks in biomedical text mining [38] and solving this problem will allow for more complex text mining tasks to be addressed [58] as it is a prerequisite for information extraction and advanced IR [38, 59, 60]. NER could also be expanded into recognizing other medically important terms, such as names of diseases [61].

One of the main reasons for the difficulty is the high degree of variations in terms that are not explicitly reflected in biomedical ontologies [62]. It is common that biological entities can have several names, for example, PTEN and MMAC1 refers to the same entity [2]. It was estimated that one-third of biological terms are variants [63]. In addition, biological and chemical entities may have multi-word names and variants of the names. For example, “Peroxisome Proliferator Activated Receptor”, “Peroxisome proliferator-activated receptor” and “Peroxisome-proliferator-activated receptor” refer to the same entity. Liu et al. [64] did a comprehensive study on this area. With new genes and proteins being discovered and named in the genomic era, it can be implied that there is no complete dictionary of named biological entities; hence, NER by simple text matching will not suffice [2, 65]. Due to the potential

utility and complexity of the problem, NER has held the attention of many researchers and it is not surprising that much work in biomedical NER had focused on recognizing gene and protein names in text.

The approaches can be classified as either lexicon-based, rules-based, statistics-based, or combinations of these means. Krauthammer et al. [66] adapted BLAST algorithm to identify entities from text with 78.8% precision and 71.7% recall (F-score = 0.75). A good example of a rules-based NER system is AbGene, which was based on a Brill tagger trained on 7000 manually-tagged sentences using a 'Gene' tag. It achieved a 85.7% precision and 66.7% recall (F-score = 0.75) [67]. In contrast, GAPSCORE, also a rules-based system, examines the appearance, morphology and context of the word before applying a classifier trained using these features. It achieved 74% precision and 81% recall (F-score = 0.77) in inexact matches, and 59% precision and 50% recall (F-score = 0.54) in exact matches [68]. On the other hand, Hanisch et al. [58] used a dictionary approach and achieved 95% precision and 90% recall (F-score = 0.92) while Egorov et al. [69] achieved 98% precision and 88% recall (F-score = 0.93). A further study report using curated dictionary demonstrated good performance, F-score between 0.8 to 0.9, across different organisms [70]. VTag [71], McDonald and Pereira [72] and ABNER [73] employed conditional random field, a statistical approach and achieved precisions between 58.2% to 85.4% and recall between 53.9% and 79.8%. Zhou et al. [74] combined several approaches using voting strategy to achieve a F-score of 0.83. Other groups had also attempted combinations of approaches to improve precision [75-78]. Jimeno et al. [61] demonstrated that although a curated dictionary method for NER is still superior (F-score of 0.59 to 0.69) compared to statistical methods (F-score of 0.28 to 0.57), voting strategy may provide a better performance than any single method (F-score of 0.54 to 0.83). Li et al. [79] reported an F-score of 0.743 by chaining 2 conditional random field models for named entity recognition followed by classification.

It is clear from recent developments that a dictionary approach (solely or in combination) tends to outperform lexicon-based or statistics-based approaches [61, 80, 81]. However, it is debatable how well NER must perform before rendered useful in biomedical text mining [82] as previous studies illustrated that performance (in terms of F-score) of biomedical information extraction in approximately equal to that of biological NER [50, 83, 84].

It is unlikely automated biomedical NER will approach that of human experts in the near future in terms of precision. However, current biomedical NER systems can be useful in providing an initial list of genes and protein names for human curation if precision is critical in the context of application.

A close relative to NER is an area of study known as gene normalization (GN). The main purpose of GN is to standardize a set of gene names using a thesaurus of gene synonymes and homologies [85] which will be useful in many aspects of biomedical literature analysis. GN is currently an area of active research and had warranted a specialized track in the latest BioCreative II challenge workshop in 2007 [86-88].

Intuitively, AR is an easier task than NER, which is supported by the observation that current AR systems uniformly perform consistently better compared to current NER system in term of their F-score [2]. Similarly to NER, AR techniques has been used to create lists of abbreviations, such as ADAM [89], which can be used for other applications.

3.4. Information Extraction: Getting the Facts

The most common aim of biomedical information extraction (IE) is finding relationships between two entities [90, 91], in this case, usually either genes, proteins or metabolites. The relationship in question may range from very general, like any form of biochemical association, to very specific, such as regulatory activation. In contrast to IR, IE systems tends to be less ad hoc but are more targeted towards specific relationships. Another difference is the granularity of text. IR systems identifies text of interest whereas IE systems work within the text to identify facts of interest, which can be subsequently verified by a curator reading the paper of interest. Biomedical information extraction are currently being developed by three different ways [92]; co-occurrence, template matching, and natural language processing.

3.4.1. Co-occurrence

Co-occurrence is fundamentally statistical and based on the tenet that multiple occurrences of the same pair of entities suggests that the pair of entities are related in some way [22, 93] and the confidence of such relatedness increases with more co-occurrences. In practice, most systems used a frequency-based scoring technique [38, 93-95] to ensure that the co-occurrence of two entities is higher than random chance [2]. Being a statistical probability, co-occurrence of entities within the same text alone will not give any insights into the type and nature of the relationship [96] unless it is used downstream to IR systems which pre-identified the type of relationships of interest [94, 97, 98]. Despite so, the advantages of co-occurrence techniques over natural language processing (NLP) are simplicity, easy to implement and efficient over large amounts of data [99].

Two of the most successful implementations of co-occurrence methods are PubGene [100] and CoPub Mapper [101]. Both had been shown to co-relate well with microarray results. A recent tool, PPI Finder [3], attempted to map queries into Gene Ontology [28, 29] before co-occurrence analysis and reported that only 28% of the co-occurred pairs in PubMed abstracts appeared in any of the commonly used human protein-protein interaction databases (HPRD, BioGRID and BIND). In addition, only 69% of the known protein-protein interactions in HPRD showed co-occurrences in the literature [3] suggesting large proportions of experimentally validated protein-protein interaction may not be reported in the literature.

A variant of co-occurrence which bootstrapped on PubMed had also been suggested [30, 102] and is commonly known as co-citation. Common experiences in using PubMed suggested that if two terms were used together with an 'AND' clause to search PubMed, it would return the intersection of the results compared to when each term was used separately. Translated statistically, the size of the intersection (proportion of co-cited documents) increases with more relatedness in the pair of entities used to interrogate PubMed. However, co-citation had not been evaluated against co-occurrence measures.

3.4.1.1. Brief descriptions of Co-occurrence systems

PubGene [100] used the idea that if 2 entities were mentioned in the same article, there will be some relationship, no matter how remote. By this, it simply counted the number of articles with occurrences of 2 entities in question and used the count as a relative strength of relatedness. If there was 1 article in 10 million mentioning both gene entities, there would be 60% chance of a true relationship between them. This was increased to 71% with 5 or more

articles in 10 million. A study testing PubGene's count based co-occurrence on statistical testing framework using Poisson distribution demonstrated that 1 co-occurred protein-pairs in 900 thousand abstracts is generally significant (p-value of less than 1%) [303] suggesting that PubGene's criteria of 1 co-occurred protein-pairs in 10 million abstracts is generally significant.

CoPub Mapper [101] calculated the article number-normalized occurrence of any 2 entities (number of articles with both concepts divided by total number of articles, divided by the product of number of articles with one concept each divided by the total number of articles), which was termed as mutual information measure. Mutual information measure was then converted to logarithmic scale and normalized on a scale of 0 to 100, known as the scaled log transformed relative score. The confidence of relationship between 2 entities was directly proportional to the scaled log transformed relative score.

MedInfoText [103] aims to extract the relationships between gene methylation and cancer from biomedical literature. It uses Lucene-based full text search engine for Perl, Plucene (http://search.cpan.org/dist/Plucene/), for term indexing. The relationships are extracted based on co-occurrences of terms in abstracts and sentences by measuring association rule interestingness [104] using support and confidence measures.

3.4.2. Natural language processing / template matching

There is significant overlap between Natural Language Processing (NLP) methods and Template Matching methods as both share many common components, such as ontologies and substantial use of Regular Expressions. Template Matching methods mainly uses the grammatical structure of English sentence to construct templates, which can be done either manually [105] or automated [106, 107], and using these templates to extract specific information from text. An example has shown extracting mutation information solely by regular expressions achieved 85% precision [108]. BioIE [109] uses rules with template matching. However, sole use of template matching is not widespread in biomedical text analysis. Instead, template matching is usually used either implicitly within NLP methods or explicitly to process the outputs of NLP. The main principles of NLP will be described to facilitate discussions into the tools that employed NLP.

Natural Language Processing (NLP) covers all aspects and stages of processing natural human speech or text into forms usable by automated systems. In the case of biomedical NLP, it can be reasonably assumed that only machine-accessible text in English language forms the majority of source materials. Hence, this section only covers the processing of English text. Given the long history of NLP, a large volume of text had been written and this section can only provide a broad coverage of the general techniques used in NLP, namely, tokenization, part of speech (POS) tagging, and shallow parsing.

The text is first broken up into its constituent atoms, known as tokens, by a process of tokenization. While the granularity of tokens may vary from chapters to phonemes (atomic units of sound), the most common form of tokenization for NLP is to break down each sentence into words and punctuations. Generally, in English text, words in a sentence are delimited by whitespace(s), except words prior to a punctuation, like a comma. This feature poses the main challenge of tokenization – distinguishing punctuations (especially period) signaling the end of either a sentence or phrase and that being part of the previous token, like in shorthand (for example, Mr., Dr.). Another challenge is the expansion of common contractions, such as "you've", which is usually done using a dictionary approach.

Part of Speech (POS) Tagging is the process of annotating a series of tokens, presumably a sentence, with semantics information (that is, the role of each token in a sentence) with a set of tags, such as Penn Treebank Tag Set [110]. There are two main approaches to POS tagging, rule-based and probabilistic. Probabilistic taggers estimate the probability of a sequence of POS tags for a given sequence of words, based on a probability model, such as the Hidden Markov Model [111, 112]. On the other hand, a rule-based tagger [39] uses contextual rules to assign tags to ambiguous words. Contextual rules, often known as context frame rules, are rules suggesting a tag based on the tag(s) before and after the unknown word. This is usually followed by morphological rules which looks at the appearance of the word. For example, words ending with "ing" is likely to be verbs.

POS Tagging can be seen as a reduction scheme to map a potentially infinite amount of words into a small and definite set of tags, to facilitate further processing. Based on the sequence of POS tags, the source text is then broken up into non-overlapping phrases. This process is chunking, also known as shallow parsing, where phrases are tagged by a small number of grammatical phrase tags, such as, Noun Phrase, Verb Phrase, Prepositional Phrase, Adverb Phrase, Subordinate Phrase, Adjective Phrase, Conjunction Phrase, and List Marker. Chunking is generally useful as a preprocessing step for information extraction as the main effect of NLP is to process unstructured data (in the form of human text) into a more structured form (POS annotated phrases), suitable for information extraction [14, 113-115] or phrase extraction in itself, can be used for medical concept identification [116]. One of the most important output of chunking is the subject-verb-object(s) tuples. In linguistic typology, the English language, together with more than 75% of all languages in the world is classified under the subject-verb-object (SVO), also known as the agent-verb-object (AVO) typological system [117]. A sentence can be processed into more than one SVO tuples, which are useful for extracting relationships between entities [118-120]. In contrast to shallow parsing (chunking), full parsing or complete parsing is much more computationally intensive but has been shown to be useful in a real world biological application [121].

3.4.2.1. Brief Descriptions of Natural Language Processing/Template Matching Systems

GENIES [14] used GenBank and SwissProt to tag genes and protein names in text before processing using MedLEE [122], a specialized text processor for biomedical literature, which used rules to parse text into structured frames. The original lexicon in MedLEE was enlarged in GENIES.

MedScan [123] used a biomedical lexicon to tag text before tokenization and stemming words into their infinite form. Tokens were syntactically processed by a parser based on active chart parser algorithm [124]. Syntactic tokens were processed into semantics structured based on an established method [125].

MeKE [126] used Gene Ontology and LocusLink as basis for constructing function names and gene names ontologies to tag text. A pattern recognizer was trained to recognize sentences or phrase describing gene functions by sentence alignment. Extracted sentences were classified by the probability of containing protein-function relationships by a Naïve Bayes classifier.

PreBIND [94] collected a list of non-redundant protein names from the NCBI RefSeq database, which was used to scan for protein names in text. Extracted term features from text

(protein names, words, adjacent words) were used to train a linear support vector machine for identifying protein-protein binding interactions using yeast data.

Arizona Relation Parser [84] specialized a Brill tagger [39] with 100 PubMed abstracts and GENIA corpus [127]. POS tagged text was processed by a hybrid shallow parser which allowed for multilevel n-ary branching of up to 24-nary, meaning POS tags or phrases up to 24 tags or phrases away could be linked. Information from each allowed combination was extracted by rule templates.

BioRAT [128] used GATE [129] to provide utilities to perform POS tagging and information extraction. Text was POS tagged to remove uninformative words, such as determinant verbs, before passing through a series of template matching to extract protein-protein interactions using a set of handwritten templates and a manually curated list of entities forming the gazetteers. An evaluation between information extraction from abstracts and full text demonstrated that 58.7% of the interactions were extracted from full text but the precision from full text was 51.3%, 3.82% lower than that of abstracts (55.07%).

Chilibot [130] used TnT POS tagger [131] trained on GENIA corpus [127] followed by CASS for shallow parsing. Chunked text was used to construct named interactions among either biological concepts, genes, proteins, or drugs. It also showed that the connectivity of molecular networks extracted from the biological literature follows the power-law distribution.

GIS [132] consists of two modules: gene information screening and gene-gene relation extraction. In gene information screening, documents were downloaded from PubMed and informative sentences were selected before tagged using a domain-specific lexicon built from online dictionaries and suggestions by biomedical researchers. Gene-gene relation extraction had a identifier for gene-gene relations, presumably trained by machine learning methods. Identified relations were then evaluated to be positive, cooperative or negative.

MedTAKMI [119] was built on TAKMI, which used standard text processing (tokenization, POS tagging, shallow parsing) to process text into [133] subject-verb-object(s) tuples as an intermediate form for further information extraction. Besides using a dictionary of protein names, the difference between TAKMI and MedTAKMI was not clear. A number of tools for information extraction from subject-verb-object(s) tuples has been described [119] but none was evaluated. Nevertheless, MedTAKMI could be used as a curator's tool.

Karopka et al. [134] used GATE [129] and modified the ANNIE gazetteer of GATE to tag for gene names before tokenizing the sentences and POS tagging by GATE. Gene relations were extracted by 34 manually-written grammar rules in JAPE (Java Annotation Patterns Engine) language, which is essentially template matching.

Cooper and Kershenbaum [97] overlapped the results of text processing, and graphical and statistical analysis to extract protein-protein interactions. TALENT text mining system [135] was used for extracting protein-protein interactions from text by a method previously established for extracting relationships between noun phrases [136]. For each pair of proteins, a 3-hop neighbourhood graph was defined and the coherence of the graph, defined as “the ratio of the number of edges present to the possible number of edges”, was calculated but the threshold coherence for a positive result was not obvious. Positive results from both methods, led to an improvement of precision from 62% to 74%.

Santos et al. [118] used a shallow parser, CASS, to extract protein names (Wnt pathway proteins) from text by statistical comparison with a Wnt pathway corpus to avoid maintenance of a list of protein names. At the same time, they used Link parser to process

text to subject-verb-object(s) tuples before full parsing to extract interactions between the Wnt pathway components. However it was not clear which POS tagger was used before shallow parsing by CASS nor which grammar was used for Link parser.

Jang et al. [137] avoided the problem of multi-word protein names and complex sentences by simplification of sentences – protein names and specific noun phrases were substituted by pre-defined words, and parenthesis phrases which does not contain entity names were removed. This simplified sentence is POS tagged by a Brill tagger [39] trained on GENIA corpus [127], then shallow parsed. Protein-protein interactions were extracted by Regular Expression parsing of shallow parsed sentences.

CONAN [138] combined several known tools and used a set of decision criteria to evaluate the output of these tools and achieved 53% precision and 52% recall using LLL corpus [139]. CONAN used Krauthammer et al. [66], AbGene [67] and NLProt [140] for NER, and MuText [141] and PreBind [94] for interaction extraction.

GeneLibrarian [29] is a PubMed text summarization tool that uses a list of gene names as input, instead of keywords. It consists of 2 modules: GeneCluster and GeneSum. GeneCluster clusters the list of given gene names using Gene Ontology while GeneSum obtains text from PubMed based on the clusters and process the text linguistically by natural language processing methods. Finally, a 9-state finite state machine (FSA) is used to perform text summarization based on the part-of-speech tags of the processed text.

Rinaldi et al. [121] used a dependency parser, Pro3Gres [142], to parse processed text. It reported precision of 96% and recall of 63%. The source text were initially analyzed for specific terms, such as entity names, before splitting into sentences by MXTERMINATOR [143] and tokenized using Penn Treebank tokenizer [110]. The tokenized text was POS tagged by MXPOST [144] and lemmatized by morpha [145]. After which, the text was processed against GENIA Ontology [127] before shallow parsed by LTCHUNK [146]. The chunks were parsed for dependencies by Pro3Gres.

Feng et al. [147] aims to extract chemical-CYP3A4 interactions from biomedical text. They had used a combined rule-based and dictionary-based method to identify chemical names in text and had used GATE [129] for POS tagging and information extraction. An evaluation with 100 abstracts demonstrated 87.4% recall and 92.3% precision for chemical name identification and 85.2% recall and 92.0% precision for the extraction of chemical-CYP3A4 interactions.

Muscorian [148] used the 2-layered generalization-specialization paradigm suggested by Novichkova et al. [123] and achieved 85% precision and 30% recall on binding and activation relationships. A manually curated dictionary of entity names were assembled for abbreviation of multi-word entity names [45] in the abstracts before processing into subject-verb-object structures using MontyLingua [149], a generic text processing engine [150] formally used to process scientific text [151, 152]. This is followed by specific data mining from the subject-verb-object structures.

E3Miner [153] aims to extract the interactions between ubiquitin-protein ligase (E3) and its target proteins from biomedical text. E3 was identified using a specially constructed POS tagger and shallow parser. The target proteins were then identified using a rule-based method before processing for Gene Ontological terms. Using a set of 47 abstracts, a precision of 97% and a recall of 74% was indicated.

PIE [154] used a 2-phase method: a term co-occurrence based method to identify potential abstracts that may contain protein-protein interactions, followed by NLP processing

using a POS tagger trained on GENIA corpus [127] to extract protein-protein interactions. PIE was tested on BioCreAtIvE I corpus [155] and reported 84% precision.

Barnickel et al. [156] used artificial neural networks for semantic role labelling of sentences at a rate of 25 to 390 milliseconds per sentence for relationship extraction. It reported a precision of 71% with 43% recall.

Jiao and Wild [157] used a set of maximum entropy based learning models for POS tagging, NER, dependency parsing, and relation extraction to extract cytochrome P-450 protein and chemical interactions. It reported an overall precision of 68.4% and recall of 72.2%.

3.4.3. Applications of biomedical information extraction

There are two main schools of thought in current biomedical IE, one school (call it the "Specialist" for further argument) takes the view that biomedical texts are specialized text (very much like the use of Legalese in legal documents) requiring highly domain-specific tools. This opinion had sparked off the development of biomedical-specific POS tag sets (such as SPECIALIST tag set [158]), POS taggers (such as MedPost [41]), ontologies and NLP systems (such as MedLEE [159]). Another school (call it the "Generalist") takes the view that biomedical texts are not sufficiently specialized to require a re-development of existing tools but either re-use or adapt generic NLP tools for biomedical text processing. This triggered the use of generic NLP systems, such as TAKMI [133], Link Grammar [160], and GATE [129], for biomedical IE.

Regardless of opinions, the focus of biomedical IE has been on a few types of relationships, namely, physical protein-protein interactions (PPIs) [14, 94, 97, 161], non-physical PPIs [162-164], relationships between proteins and diseases and terms [101, 165-167], gene regulation [113, 168], protein phosphorylation [153, 169], alternate transcription [170], and functions of transcription factors [171].

Most of the earlier work in biomedical IE belongs to the Specialist's School and one of the significant contributions was the GENIES system [14], which modified MedLEE's lexicon preprocessor and parser. GENIES was only evaluated using one article and reported an overall precision of 96%. MedLEE [122] was also adapted to process pathology reports for breast cancer study [172]. The MeKE system used a lexicon of gene and protein names from LocusLink to construct an ontology for pattern matching within text into structured data [126]. Novichkova et al. [123] agreed that NLP can be used to process text into semantic structures and developed MedScan, which incorporated a biomedical lexicon. Further work by Daraselia demonstrated 91% precision with 21% recall (F-score = 0.34) in extracting protein interactions from text using MedScan [163]. The output of MedScan were assembled and constructed into a set of tissue-specific pathways, ResNetCore database [173]. The Arizona Relation Parser [84] attempted to improve MedScan's low recall by re-training Brill Tagger [39] with Brown Corpus, Wall Street Journal, PubMed abstracts, and added GENIA lexicon [127]. This was followed by a hybrid grammar, template matching and semantic filtering. It reported 89% precision with 35% recall (F-score = 0.5). GIS [132] uses a domain-specific lexicon but instead of NLP, it employs a machine learning approach and reported 84% precision with 77% recall (F-score = 0.80). Jang et al. [137] had trained Brill tagger [39] on GENIA corpus [127] and a purpose-built protein-protein interaction extractor system which achieved 81% precision and 43% recall (F-score = 0.56). E3Miner [153] built a specialized POS tagger and shallow parser to achieve 97% precision and 74% recall (F-score

= 0.84). PIE [154] used a GENIA [127] trained POS tagger and achieved 84% precision. Jiao and Wild [157] used separate maximum entropy based learning models for each component and reported 68.4% precision and 72.2% recall (F-score = 0.70). Barnickel et al. [156] used artificial neural networks for semantic role labelling of sentences for relationship extraction and reported 71% precision with 43% recall (F-score = 0.54).

In the Generalist's School, BioRAT [128] is one of the earliest systems to modify GATE [129] to extract protein-protein interactions and reported 48% precision with 39% recall (F-score = 0.43). TAKMI [133], originally developed to process customers call logs in IT helpdesks, was used to develop MedTAKMI [119] which uses term frequency to support search results. Karopka et al. [134] modified the ANNIE system of GATE [129] and modified precision and recall measures to account for partial correct extractions, and reported 92.8% precision with 30% recall (F-score = 0.45). Cooper and Kershenbaum [97] used a generic TALENT text mining system [135] with graphical and statistical approaches to mine for protein-protein interactions and reported 74% precision. A Link grammar parser [160] was used to mine the Wnt pathway and reported 90% precision with 64% recall (F-score = 0.75) [118]. Rinaldi et al. [121] used a myraid of text processing tools with GENIA Ontology [127] and reported 96% precision with 63% recall (F-score = 0.76). Feng et al. [147] used a purpose-built rule-dictionary hybrid for chemical name identification and GATE [129] for information extraction and reported 85% recall and 92% precision (F-score = 0.87). Muscorian [148] used MontyLingua [149], a generic text processing engine [150], in the 2-layered generalization-specialization paradigm [123] and achieved 90% precision and 30% recall (F-score = 0.45).

From the research results gathered from both school of thought (tabulated in Table 1), it is still not possible to demonstrate superiority of one approach over the other in terms of system performance. Intuitively, it might be easier to modify an existing system for a specific application than to develop one from scratch. In addition, it has not been demonstrated that systems developed from the Specialist's school of thought can be adapted to extract other biomedical relationship of interests as only a few systems have been designed to extract multiple relationships [38]. On the other hand, Generalists view generic NLP systems as a processing tool to convert unstructured text into structured forms, such as tuples; thus, is inherently more readily adapted for different problems. It is probably reasonable to comment that by adapting an existing system for use in biomedical text mining usually implies that the system has been used in a different context. For example, TAKMI was used in 3 different areas; analyzing customer call logs [133], generating frequently-asked-questions candidates [174], and in MedTAKMI [119].

However, adapting a generic system may require intense effort in formulating rules and templates (GATE and Link Grammar) for the specific problem domain or re-training parts of the system, especially POS tagger, which may require a prior manual tagging of training corpus [38]. For instance, Chilibot [130] used TnT tagger [131] trained on the GENIA corpus [127] but succeeded in using CASS parser (http://www.vinartus.net/spa), un-modified, for chunking. This might be an obstacle to adapt a previously adapted generic NLP system for a biomedical problem to another biomedical problem. Moreover, there is no certainty of rewards in this effort as Miyao et al. [175] had shown that combining text processing components may be synergistic.

Table 1. Summary of performances of biomedical literature analysis systems. 'NG' means that the particular performance metric was not given in the study.

Specialist Systems				Generalist Systems			
Study	Precision	Recall	F-Score	Study	Precision	Recall	F-Score
[14]	0.96	NG	--	[128]	0.48	0.39	0.43
[163]	0.91	0.21	0.34	[134]	0.93	0.30	0.45
[84]	0.89	0.35	0.50	[97]	0.74	NG	--
[132]	0.84	0.77	0.80	[118]	0.90	0.64	0.75
[137]	0.81	0.43	0.56	[121]	0.96	0.63	0.76
[153]	0.97	0.74	0.84	[147]	0.92	0.85	0.87
[154]	0.84	NG	--	[148]	0.90	0.30	0.45
[157]	0.68	0.72	0.70				
[156]	0.71	0.43	0.54				

It is inherent in the process of evaluating systems using corpora that human experts are the only absolute performer. That is, human experts are performing at 100% precision and 100% recall. Therefore, it should be conceivable that learning from the output errors by artificial intelligence and machine learning methods could be used to improve information extraction systems. The first of such biomedical information extraction systems which uses support vector machines and neural networks to mimic human expert curation had surfaced [176] with precision ranging from less than 30% to more than 90% over 68 extraction tasks. In addition, biomedical information extraction had been shown to be able to improve curation efficiency of protein-protein interactions into database [177].

3.5. Text Mining: Finding Hypotheses

While the main premise of IR and IE is deductive reasoning (the conclusion is of no greater generality than the premises), text mining (TM) is fundamentally inductive reasoning (the conclusion is of greater generality than the premises). In other words, TM aims at finding or induce new information and hypotheses from existing knowledge from the literature. One of the pioneers of biomedical TM is Don Swanson who suggested in the mid-80s that there were connections between fish oil and Raynaud's syndrome [4], migraine and magnesium [5], arginine intake and the level of somatomedin C in blood [178] This had triggered the advancement of biomedical IR/IE, NER and AR, which are all precursors to TM. The method Swanson used is essentially Hypothetical Syllogism (if p then q; if q then r; therefore, if p then r), which is an extension of Modus Ponens. In biomedical TM, it is commonly referred to as Swanson's ABC model [179]. Using this discovery model, Weeber et al. [180] had attempted to automate it and found new potential uses for thalidomide. More recently, the potential therapeutic use of turmeric on spinal cord injuries was suggested [181].

Despite its potential and history, biomedical TM is still at its infancy [182]. In order for hypothesis generation systems to be a standard tool of biologists, a fundamental question needs addressing – how to evaluate an untested set of hypotheses? [2] A way to circumvent

this problem may be using statistical measurements from IR/IE to provide a means of prioritizing the potential of each hypotheses, as shown as Anne 2 [183]. In spite of this inherent problem, there may be use of TM to evaluate and score several possible hypotheses from experimental or clinical research [184]. TM is also known by other authors as "literature based discovery" [180, 185, 186] or "knowledge discovery" [187].

4. RELATED AREAS OF IMPORTANCE

Notwithstanding the development in previously discussed areas, there are three other key areas that are important within the literature analysis pipeline, namely, corpora, which forms the gold standard for evaluating systems; databases, which may be used to evaluate systems or are themselves resulting from literature analysis systems; evaluation strategies, the definition of performance metrics and their calculations; assisted microarray analysis using output from biomedical text analyses; and visualization tools for viewing large interaction maps.

4.1. Corpora

A corpus (corpora for plural) is a collection of literature which has been either tagged, annotated or categorized for specific purpose(s). Essentially, a corpus is a defined data set of literature. The importance of corpora to literature analysis tools cannot be over-emphasized, analogously, it is as important as antibodies to protein studies. However, there are not many corpora of biomedical literature for various purposes as they often require manual annotations with high-level agreement among annotators [188, 189], known to be labour-intensive to create [38] and need to reflect a biologist's interpretation of the text [154]. The main value of corpora is that it provides a known finite source of positives which is essential for calculating recall measure and error analyses.

Categorically, the following biomedical corpus for different purposes are as follows: For protein and gene name recognition (NER), there are Yapex (used in [68]) and GeneTag [190], PennBioIE [71] corpora. For abbreviation recognition, there are Medstract [191] and AB3P [53] corpora. For part-of-speech tagging, there are GENIA [127, 192, 193], PennBioIE [71] and MedPost [41]. GENIA team had also expanded the annotation into biomedical events to reflect a biologist's understanding of the text [154]. For relationship extraction, there is a dataset used for Learning Logic in Language 2005 (www.cs.york.ac.uk/aig/lll/) [139, 194] and BioCreAtIvE corpus [155]. BioScope corpus [195] represents the first attempt to incorporate uncertainty or negative information into a corpus.

There is a general sentiment that progress in biomedical literature analysis suffers from the lack of corpora [15] which is relatively obvious when one starts to list down the corpora available for each purpose. There is no biomedical corpus for shallow parsing (chunking) or citation retrieval from PubMed (information retrieval) and minimal choices for relationship extraction. Moreover, testing using different corpora (if any) can result in F-score varying as much as 19% [196]. Hence, researchers had resorted to compare their system output with that

in curated databases [197] as these databases represent high-quality molecular interaction data [198].

4.2. Databases

The main database for biomedical literature is PubMed where most source materials for literature analysis work is derived form. Possibly, the largest repository for biochemical information is KEGG which provides links to GenBank and a number of other publically available databases. Databases are repositories of source text (PubMed), curated tools for comparison (comparing system output against KEGG in Zhang et al., [197], Lee et al. [199] compared their system against SwissProt and Maguitman et al. [200] tested their system against Pfam), or are themselves the results from literature analysis, such as DIP [201]. It is generally true that databases can benefit from literature analysis efforts [202]. Large institutional initiatives, such as KEGG and BIND [203], which are mainly manually maintained and curated, cater to the general research community.

Bioinformatics has a section of the periodical catering to the publications of databases and Nucleic Acid Research releases issues periodically with a database focus (known as database issue). The latest edition of Nucleic Acid Research database issue (Volume 36) features 98 databases. These have almost became a de facto source of new databases. As the availability of corpora is scarce [204], a cursory knowledge of database availability might assist in evaluation efforts.

In terms of individual proteins, there are databases for proteins in specific organelles [205, 206]; prokaryotic proteins [207]; proteins of specific biochemical events [208-211]; proteins of specific domains or characteristics [212-222]; protein anomalies [223]; transcription factors [224, 225]; crystal structures [226]; proteomics resources [227] and specific classes of proteins, such as lectin [228] and centrosomal proteins [229].

Some databases focused on protein-protein interactions, which may be all types of interactions [201, 230-239] or specific interactions [240-244]. For genes, there are databases for genes of specific characteristics [211, 216, 245-256]; cleavage sites [257]; genetic variations [253, 258-262], promoters or regulatory elements [263-267]; genes of specific organisms [268] organelle genomes [206, 269]; comparative genomics [270]; entire genomic resource [271-274]; and expressed sequence tags [275-277].

Other databases includes those for managing experimental results [278-280]; haptens [281]; orientation of proteins on cellular membranes [282]; protein localization [283]; and therapeutically important pathways [284]. Currently, the largest and most extensive database for microarray and other high-throughput data storage is the Gene Expression Omnibus (GEO) [285].

Scanning the wide variety of databases, it is clear that the challenge is not the creation of databases but on the use of these databases, especially integrating them into a composite (federation) of biomedical databases and querying them [286-289].

4.3. Evaluation Strategies

During the development of a literature analysis tool, it is critical to have an estimation of the reliability, which are usually compared to a standard of desired results [92], usually in a form of tagged, annotated or categorized corpus. The most common evaluation strategy and metrices (measurements), such as precision and recall, originated from the Second Message Understanding Conference in 1989.

Given a corpus and a specific query, the results can be partitioned into true positives (TP; items correctly labeled as positive), false positives (FP; items incorrectly labeled as positive) true negatives (TN; items correctly labeled as negative), and false negative (FN; items incorrectly labeled as negative). With these four items, a few metrics can be established, the most common being precision, also known as positive predictive value, is defined as TP/(TP+FP); recall is TP/(TP+FN). Precision and recall are typically inversely related [290].

Precision and recall are commonly used because of their simplicity to evaluate against an established standard (annotated corpus). However, in the absence of a standard, precision can still be evaluated comparing the output of a system with its input. Karopka et al. [134] modified precision and recall measures to account for partially correct extractions. Precision and recall are important because the inverse of precision is a measure of false positives (1 – precision) of the system output and the inverse of recall measures false negatives (1 – recall) or proportion of lost information as a result of processing.

It is important to note that in IE, it is not possible to define true negatives (TN) as there is no theoretical bounds of the number of 'facts' that can be generated from a piece of text [291]. Therefore, a number of measures that required TN, such as accuracy ((TP+TN)/(TP+FP+TN+FN)), error rate ((FP+FN)/(TP+FP+TN+FN); which is 1 – accuracy, negative predictive value (TN/(FN+TN)), prevalence ((TP+FN)/(TP+FP+TN+FN)), and specificity (TN/(TN+FP)) are impossible to calculate. In addition, receiver operating characteristics (ROC), which had been used extensively in evaluating system performance [292-295], cannot be calculated for IE as it requires specificity.

Notwithstanding variations using different corpora for evaluating different systems and using different criteria for assigning results into each of the three bins (TP, FP, FN), it will be difficult to compare two systems each characterized by precision and recall. Given presence of a single decision parameter (non-categorical variable), it is possible to obtain and compare the respective relative operating characteristic curves (aROC) [296]. However, aROC is not possible for systems without a single decision parameter. Thus, precision (P) and recall (R) are reduced to a single F-score, defined as 2PR/(P+R), which is the harmonic mean of precision and recall [16, 17], and is always between 0 and 1 where 1 means that the system produces neither FP or FN. F-score assigns the same weight to both precision and recall, that is, both are equally important. A more general form of F-score allows for different weight be assigned to precision and recall [16]. Hirschman et al. [291] proposed a variant of F-score, simple matching coefficient (SMC), which is defined as TP/(TP+FN+FP).

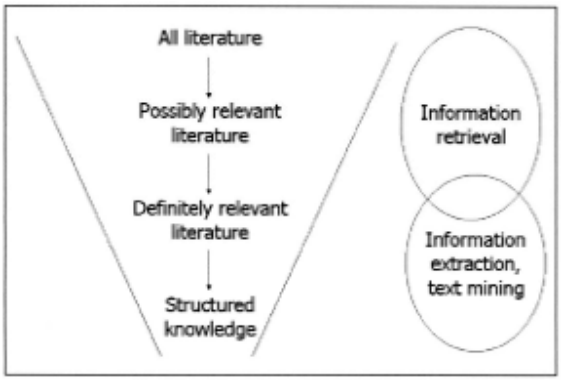

Figure 2. The funnel of knowledge-based information. Demonstrating how the structuring and understanding of knowledge progresses from all biomedical literature becomes more refined and relevant. With larger amounts of literature, information retrieval techniques are essential and once the relevant literature has been defined, information extraction and text mining are required [297]

5. CHALLENGES OF BIOMEDICAL LITERATURE ANALYSIS

The challenge for the field of biomedical literature analysis is to manage and process large amounts of literature. The rate of publication of literature has exceeded the capacity to manually review it, and therefore, there is an increasing need for IE in this post-genomic era where the focus of biomedical research is shifting from the study of individual proteins and genes to entire biological systems.

Biomedical literature analysis (see Figure 2) consists of getting source text from text repository (information retrieval; IR), finding information of interests within the text (information extraction; IE) and in the process, may require the recognition of entities like proteins (named entity recognition; NER) and evaluation of abbreviations (abbreviations recognition; AR). Extracted information may either be deposited into specialized databases as structured knowledge or may support knowledge discovery or support hypothesis generation by text mining (TM). Testing and evaluation of each step in the analysis requires the presence of a defined data set (corpus) for certain metrics, like recall measure, to be estimated. Alternatively, evaluation can be done by comparing with a suitable, existing database. Inherent in this process is the definition of an appropriate evaluation strategy to allow for comparisons across similar systems.

The main repository of biomedical textual data is PubMed (MedLine) which provided only one means for searching relevant information. Although tools such as Textpresso [24] and MedMiner [25] represented an alternative, they are essentially bootstrapping on PubMed's IR. Despite its universal use, neither the implementation details (source code) of PubMed's IR engine is known [30, 298] nor is the engine thoroughly evaluated for precision and recall. All research to date that retrieved data from PubMed as source had assumed that the data is pristine but this is neither possible nor verified. As a result, certain research questions in biomedical literature analysis which depended on the accuracy of PubMed, such as, what does the collective research knowledge to date tells us about the proteomic and metabolomic differences between mouse and rat, is fundamentally impossible at this stage.

Information extraction is heavily reliant on precise NER [59] and AR. In spite of on-going debate as to the required performance of automated NER [82] and AR systems before being useful in biomedical literature analysis, it is clear that near-human precision is unlikely in the near future. As a result of non-optimal performance (human performance is assumed optimal), IE is generally limited to a priori approach, that is, the user has to know the list of proteins or entities whose relationship he is interested in, as opposed to an a posteriori approach, such as finding relationships of possibly new entities or entities that are unknown to the user at time of search. Drawing analogies from genomic studies, a priori approach is analogous to microarray technology (the spots on the chip are pre-determined before actual experimentation) while a posteriori approach is likened to Massive Parallel Signature Sequencing (MPSS) [299]. Nevertheless, current or in future, improved NER and AR systems can be of great use in assisting human curators to assemble a near complete dictionary [300], such as ADAM [89]. Systems that learn from human curation efforts, in attempt to improve its performance, had recently surfaced [176]. A system that performed automated curation of extracted interactions from text at the graph-level has also emerged [301]. Another similar problem is the need to recognize variants of the same name, commonly known as gene normalization [86].

Relationships of paired entities by co-occurrence statistics usually requires large volumes of initial text as the term frequency of each entity (number of occurrences of a term divided by the total number of documents) is low, presumably less than 5% of the document set (corpus). This is likely to restrict it use on "small" corpus of less than 100000 and PubGene had used more than 10 million abstracts to generate its gene network [100]. There are three advantages of co-occurrence methods when compared to NLP. Firstly, co-occurrence is basically statistical correlation and is easier to understand by more biologists than NLP techniques. Secondly, most NLP systems work at the sentence level; thus, cannot extract relationships either spanning more than one sentence or in complex sentences, where information may be split across multiple SVO tuples. On the other hand, co-occurrence can be easily deployed on different granularity of text. But it is necessary to note that the fundamental assumption of co-occurrence is independent observations, which is assumable for abstracts as full text usually refer to prior papers (in its introduction and discussion), thus, independent observations cannot be assumed. Lastly, it is known that co-occurrence methods generally has a higher recall but lower precision as compared to NLP means [302], and given that IE by NLP generally suffer from poor recall, it may be possible to improve the overall performance (improving recall substantially while suffering a small decline in precision) by simultaneously employing both co-occurrence and NLP. This notion had been supported by a recent study [303] demonstrating that NLP extracted interactions is generally a proper subset of co-occurred pairs, suggesting that NLP can be used to annotate co-occurred pairs.

As previously described, biomedical IE is driven by the Specialist (biomedical text are highly domain-specific and require specially developed NLP tools) and Generalist (generic or existing non-biomedically focused NLP tools can be adapted for biomedical use) schools of thought. However, both directions require either formulation of rules and templates or re-training parts of the system. Both tasks are manually intensive, require manually tagged corpus [38]. Furthermore, there is no certainty in a better system and combining existing tools may be synergistic [175]. These approaches generally do not fall within the expertise of biologists which are the very people using the systems [204]. Early studies by Grover [304, 305] suggested that native generic NLP tools may be used in biomedical text. Recently, a

study by Ling et al. [148] had use an un-modified, generic NLP system for biomedical literature analysis and reported comparable precision. Although it has generally been assumed that some modifications must be made to generic NLP systems (especially the POS tagger) for it to be used on biomedical text, further analysis by Ling et al. [306] revealed that POS tagging accuracy may not negatively impact on the transformation to subject-verb-object structures due to complementary POS tag use in shallow parsing. However, Ling et al. [148] examined the extraction of 2 interactions from published abstracts and the extrapolation of these results to other interactions [24] requires further studies.

All biomedical literature analysis systems require some form of evaluation, either as performance metrics, such as precision and recall, or as statistical confidence of results (like in BLAST), in attempt to make evaluation across systems. However, this approach faces a number of challenges. Firstly, evaluation by performance often requires tagged corpora which are in severe shortage [15] and most available corpora do not provide programmatic tools to use them readily; hence, developers across the globe have to implement access routines in a particular computer programming language for each new corpus. Secondly, the current evaluation of individual systems make it difficult for comparison even though performance metrics may be known. This is due to different approaches used to obtain the performance metrics. For example, GENIES [14] was evaluated using only one paper; BioRAT [128] was evaluated against an existing database, DIP [201]; E3Miner [153] was evaluated against 47 abstracts. In order to be statistically sound, all systems should preferentially be evaluated against a common set of data, which was accomplished in challenges, such as TREC (trec.nist.gov/) [307], BioCreative and LLL (www.cs.york.ac.uk/aig/lll/) [194]. Alternatively, a common dataset can be established for communal use [196, 204, 308], like that of UCI Machine Learning Repository [309].

One of the main reasons this technology is slowly adopted by biologists is because they are not trained in computer science to integrate the tools effectively [310]. Therefore, it is necessary to present clear benefits of using these tools [2, 88, 204, 311-313]. It is almost a tradition in data analysis and mining to create a system that allows users to set their own parameters in accordance to the task at hand or to evaluate a system independent of meeting user needs [2]. However, the biologists using the system, who are generally clueless about the nature of each parameters, faces a daunting task of setting these parameters. Therefore, it is necessary to involve the biologists in the process of creating new tools or adapting or aggregating existing tools to help biomedical researcher to solve real world problems [2, 314, 315] as these needs remains unmet [316]. Hence, a recent trend is to use literature analysis to provide and update evidence data for Gene Ontology annotations [317-322], combining literature analysis with ontology information for query answering [323], extracting concepts from text [324] or to access the information needs of specific areas of research [315]. At the same time, literature analysis has also been used to group genes based on their functions [325], extracting medically important terms from text [61] and further the development of new ontologies [326].

6. CONCLUSION

The golden age of biomedical literature analysis of 1998 to 2009 had left us with a number of disjoint sets of tools: systems for specific purposes in the process, such as MedPost; systems for specific biological purposes, like microGENIES; various ontologies and lexicons, like Textpresso Ontology, GENIA Ontology; visualization tools, etc. Although it seems that technology transfer from more traditional fields, such as computational linguistics for understanding general text, had reached its maximum in that period, creative use of these techniques, picking up and re-structuring the pieces left behind, and targeting the resulting systems to the actual needs of biologists, could bring forth the next golden age of biomedical literature analysis.

ACKNOWLEDGEMENT

We wish to thank Professor Thomas Rindflesch, National Institute of Health, USA; Professor Jonathan Wren, Associate Editor for Bioinformatics, for his comments on improving the initial drafts.

REFERENCES

[1] Hunter, L. & Cohen, K. B. (2006). Biomedical language processing: what's beyond PubMed? *Molecular Cell*, *21*, 589-594.

[2] Cohen, A. M. & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, *6*, 57-71.

[3] He, M., Wang, Y. & Li, W. (2009). PPI finder: a mining tool for human protein-protein interactions. *PLoS ONE, 4*, e4554.

[4] Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, *30*, 7-18.

[5] Swanson, D. R. (1988). Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, *31*, 526-557.

[6] Prange, J. D. (1996). *Evaluation driven research: the foundation of the TIPSTER text program.* Tipster Text Program Phase II, May 6-8, 1996.

[7] Hersh, W., Bhupatiraju, R. T. & Corley, S. (2004). Enhancing access to the Bibliome: the TREC Genomics Track. *Medinfo*, *11*, 773-777.

[8] Hirschman L. (1998). The evolution of evaluation: lessons from the Message Understanding Conferences. *Information Processing and Management, 37*, 383-402.

[9] Leek TR. Information extraction using Hidden Markov Model. *Department of Computer Science*. University of California, San Diego (1997)..

[10] Fukuda K., Tsunoda T., Tamura A., Takagi T. (1998). Toward information extraction: identifying protein names from biological papers. *Proceedings of the Pacific Symposium on Biocomputing* (PSB'98): 705 - 716.

[11] Craven M., Kumlien J. (1999). Constructing biological knowledge bases by extracting information from text sources. *Proc Int Conf Intell Syst Mol Biol,* 77-86.

[12] Blaschke C., Andrade, M. A., Ouzounis, C. & Valencia, A. (1999). Automatic extraction of biological information from scientific text: protein-protein interactions. *Proc Int Conf Intell Syst Mol Biol*, 60-67.

[13] Shatkay H. & Wilbur, W. J. (2000). *Finding themes in Medline documents: probabilistic similarity search.* IEEE Conference on Advances in Digital Libraries., pp. 183-192.

[14] Friedman C., Kra, P., Yu, H., Krauthammer, M. & Rzhetsky, A. (2001). GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*., *17*, S74-S82.

[15] Leser, U. & Hakenberg, J. (2005). What makes a gene name? Named entity recognition in the biomedical literature. *Briefings on Bioinformatics.*, *6*, 357-369.

[16] Natarajan, J., Berrar, D., Hack, C. J. & Dubitzky, W. (2005). Knowledge discovery in biology and biotechnology texts: a review of techniques, evaluation strategies, and applications. *Critical Reviews in Biotechnology*, *25*, 31-52.

[17] Tsai, R. T., Wu, S. H., Chou, W. C., Lin, Y. C., He, D. & Hsiang, J. et al. (2006). Various criteria in the evaluation of biomedical named entity recognition. *BMC Bioinformatics*, *7*, 92.

[18] Han, B., Obradovic, Z., Hu, Z. Z., Wu, C. H. & Vucetic S. (2006). Substring selection for biomedical document classification. *Bioinformatics*.

[19] Chen, D., Muller, H. M. & Sternberg, P. W. (2006). Automatic document classification of biological literature. BMC *Bioinformatics*, *7*, 370.

[20] Gerard, S., Edward, A. F. & Harry, W. (1983). Extended Boolean information retrieval. Communications of the *ACM*, *26*, 1022-1036.

[21] Aronson, A. R., Bodenreider, O., Chang, H. F., Humphrey, S. M., Mork, J. G. & Nelson, S. J. et al. (2000). The NLM Indexing Initiative. *Proc AMIA Symp,* 17-21.

[22] Wilbur, W. J. & Yang, Y. (1996). An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Comput Biol Med*, *26*, 209-222.

[23] Wilbur, W. J. (2002). A thematic analysis of the AIDS literature. *Pacific Symposium on Biocomputing*., *7*, 386-397.

[24] Muller, H. M., Kenny, E. E. & Sternberg, P. W. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biology*., *2*, e309.

[25] Tanabe, L., Scherf, U., Smith, L. H., Lee, J. K., Hunter, L. & Weinstein, J. N. (1999). MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques*., *27*, 1210-1214, 1216-1217.

[26] Shatkay, H., Pan, F., Rzhetsky, A. & Wilbur, W. J. (2008). Multi-dimensional classification of biomedical text: toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, *24*, 2086-2093.

[27] Fontaine, J. F., Barbosa-Silva, A., Schaefer, M., Huska, M. R., Muro, E. M. & Andrade-Navarro, M. A. (2009). MedlineRanker: flexible ranking of biomedical literature. *Nucleic Acids Res*., *37*, W141-146.

[28] Doms, A. & Schroeder, M. (2005). GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Research*, *33*, W783-786.

[29] Chiang, J. H., Shin, J. W., Liu, H. H. & Chin, C. L. (2006). GeneLibrarian: an effective gene-information summarization and visualization system. *BMC Bioinformatics*, *7*, 392.

[30] Simon, M. L., Patrick, M., Kimberly, F. J. & Jennifer, S. (2004). MedlineR: an open source library in R for Medline literature data mining. *Bioinformatics*, *20*, 3659.

[31] Yoo, I., Hu, X. & Song, I. Y. (2007). Biomedical ontology improves biomedical literature clustering performance: a comparison study. International Journal of *Bioinformatics Research and Applications*., *3*, 414-428.

[32] Ding, J., Viswanathan, K., Berleant, D., Hughes, L., Wurtele, E. S. & Ashlock, D. et al. (2005). Using the biological taxonomy to access biological literature with PathBinderH. *Bioinformatics*., *21*, 2560-2562.

[33] Baker, C. J., Kanagasabai, R., Ang, W. T., Veeramani, A., Low, H. S. & Wenk, M. R. (2008). Towards ontology-driven navigation of the lipid bibliosphere. *BMC Bioinformatics*, *9,* Suppl *1*, S5.

[34] Aronson, A. R. & Rindflesch, T. C. (1997). Query expansion using the UMLS Metathesaurus. Proc AMIA *Annu Fall Symp*, 485-489.

[35] Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*, 17-21.

[36] Hersh, W., Price, S. & Donohoe, L. (2000). Assessing thesaurus-based query expansion using the UMLS Metathesaurus. *Proc AMIA Symp*, 344-348.

[37] Witte, R., Kappler, T. & Baker, C. J. (2007). Enhanced semantic access to the protein engineering literature using ontologies populated by text mining. International Journal of Bioinformatics *Research and Applications*., *3*, 389-413.

[38] Jensen, L. J., Saric, J. & Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nature Review Genetics*., *7*, 119-129.

[39] Brill, E. (1995). Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging. *Computational Linguistics*., *21*, 543-565.

[40] Garten, Y. & Altman, R. B. (2009). Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC Bioinformatics*., *10*, Suppl *2*, S6.

[41] Smith, L., Rindflesch, T. & Wilbur, W. J. (2004). MedPost: a part-of-speech tagger for bioMedical text. *Bioinformatics*., *20*, 2320-2321.

[42] Kudo, T., Matsumoto, Y. (2000). Use of support vector learning for chunk identification. *4th Conference on CoNLL-2000 and LLL*-142-144.

[43] Liu, H. & Friedman, C. (2003). Mining terminological knowledge in large biomedical corpora. *Pacific Symposium on Biocomputing*., *8*, 415-426.

[44] Chang, J. T. (2003). *Using machine learning to extract drug and gene relationships from text.* pp. 183. Stanford University.

[45] Chang, J. T., Schutze, H. & Altman, R. B. (2002). Creating an online dictionary of abbreviations from MEDLINE. *Journal of the American Medical Informatics Association*., *9*, 612-620.

[46] Yu, H., Hripcsak, G. & Friedman, C. (2002). Mapping abbreviations to full forms in biomedical articles. *Journal of the American Medical Informatics Association*., *9*, 262-272.

[47] Schwartz, A. S. & Hearst, M. A. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. *Pacific Symposium on Biocomputing*., *8*, 451-462.

[48] Adar, E. (2004). SaRAD: a Simple and Robust Abbreviation Dictionary. *Bioinformatics*., *20*, 527-533.

[49] Pustejovsky, J., Castano, J., Cochran, B., Kotecki, M. & Morrell, M. (2001). Automatic extraction of acronym-meaning pairs from MEDLINE databases. *Medinfo*., *10*, 371-375.

[50] Wren, J. D. & Garner, H. R. (2002). Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. *Methods of Information in Medicine*., *41*, 426-434.

[51] Wren, J. D ., Chang, J. T., Pustejovsky, J., Adar, E., Garner, H. R. & Altman, R. B. (2005). Biomedical term mapping databases. *Nucleic Acids Research.*, *33*, D289-293.

[52] Okazaki, N. & Ananiadou, S. (2006). Building an abbreviation dictionary using a term recognition approach. *Bioinformatics*., *22*, 3089-3095.

[53] Sohn, S., Comeau, D., Kim, W. & Wilbur, W. J. (2008). Abbreviation definition identification based on automatic precision estimates. *BMC Bioinformatics*., *9*, 402.

[54] Xu, Y., Wang, Z., Lei, Y., Zhao, Y. & Xue, Y. (2009). MBA: a literature mining system for extracting biomedical abbreviations. *BMC Bioinformatics*., *10*, 14.

[55] Kuo, C. J., Ling, M. H. T., Lin, K. T. & Hsu, C. N. (2009). *BIOADI: A machine learning approach to identifying abbreviations and definitions in biological literature.* 9th International Conference on Bioinformatics., Singapore.

[56] Torii, M., Hu, Z. Z., Song, M., Wu, C. H. & Liu, H. (2007). A comparison study on algorithms of detecting long forms for short forms in biomedical text. *BMC Bioinformatics*., *8* Suppl *9*, S5.

[57] Franzen, K., Eriksson, G., Olsson, F., Asker, L., Liden, P. & Coster, J. (2002). Protein names and how to find them. *International Journal of Medical Informatics*., *67*, 49-61.

[58] Hanisch, D., Fluck, J., Mevissen, H. T. & Zimmer, R. (2003). Playing biology's name game: identifying protein names in scientific text. *Pacific Symposium on Biocomputing*., 403-414.

[59] Krauthammer, M. & Nenadic, G. (2004). Term identification in the biomedical literature. *Journal of Medical Bioinformatics*., *37*, 512-526.

[60] Proux, D., Rechenmann, F., Julliard, L., Pillet, V. V. & Jacq, B. (1998). Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction. *Genome informatics Workshop on Genome Informatics*., *9*, 72-80.

[61] Jimeno, A., Jimenez-Ruiz, E., Lee, V., Gaudan, S., Berlanga, R. & Rebholz-Schuhmann, D. (2008). Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*., 9 Suppl *3*, S3.

[62] Nenadic, G., Spasic, I. & Ananiadou, S. (2004). Mining biomedical abstracts: What is in a term? In: K.Y. Su, (ed), *Natural Language Processing - IJCNLP 2004*. Springer., Berlin, 797-806.

[63] Jacquemn, C. (2001). *Spotting and discovering terms through natural language processing,* MIT Press, Cambridge, MA.

[64] Liu, H., Hu, Z. Z., Torii, M., Wu, C. & Friedman, C. (2006). Quantitative assessment of dictionary-based protein named entity tagging. *J Am Med Inform Assoc., 13*, 497-507.

[65] Tsuruoka, Y. & Tsujii, J. (2004). Improving the performance of dictionary-based approaches in protein name recognition. *J Biomed Inform*., *37*, 461-470.

[66] Krauthammer, M., Rzhetsky, A., Morozov, P. & Friedman, C. (2000). Using BLAST for identifying gene and protein names in journal articles. *Gene*., *259*, 245-252.

[67] Tanabe, L. & Wilbur, W. J. (2002). Tagging gene and protein names in biomedical text. *Bioinformatics*., *18*, 1124-1132.

[68] Chang, J. T., Schutze, H. & Altman, R. B. (2004). GAPSCORE: finding gene and protein names one word at a time. *Bioinformatics*., *20*, 216-225.

[69] Egorov, S., Yuryev, A. & Daraselia, N. (2004). A simple and practical dictionary-based approach for identification of proteins in Medline abstracts. *J Am Med Inform Assoc*., *11*, 174-178.

[70] Hanisch, D., Fundel, K., Mevissen, H. T ., Zimmer, R. & Fluck, J. (2005). ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*., *6*, Suppl *1*, S14.

[71] Ryan, T. M., Winters, R. S., Mark, M., Yang, J., Peter, S. W. & Fernando, P. (2004). An entity tagger for recognizing acquired genomic variations in cancer literature. *Bioinformatics*., *20*, 3249.

[72] McDonald, R. & Pereira, F. (2005). Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics*., 6 Suppl *1*, S6.

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

[74] Zhou, G., Shen, D., Zhang, J., Su, J. & Tan, S. (2005). Recognition of protein/gene names from text using an ensemble of classifiers. *BMC Bioinformatics*., *6*, Suppl *1*, S7.

[75] Hatzivassiloglou, V., Duboue, P. A. & Rzhetsky, A. (2001). Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics*., *17*, Suppl *1*, S97-106.

[76] Hou, W. J. & Chen, H. H. (2004). Enhancing performance of protein and gene name recognizers with filtering and integration strategies. *Journal of Biomedical Informatics*., *37*, 448-460.

[77] Majoros, W., Subramanian, G. & Yandell, M. (2003). Identification of key concepts in biomedical literature using a modified Markov heuristic. *Bioinformatics*., *19*, 402-407.

[78] Finkel, J., Dingare, S., Manning, C. D., Nissim, M., Alex, B. & Grover, C. (2005). Exploring the boundaries: gene and protein identification in biomedical text. *BMC Bioinformatics*, *6*, Suppl *1*, S5.

[79] Li, L., Zhou, R. & Huang, D. (2009). Two-phase biomedical named entity recognition using CRFs. *Comput Biol Chem*., *33*, 334-338.

[80] Kou, Z., Cohen, W. W. & Murphy, R. F. (2005). High-recall protein entity recognition using a dictionary. Bioinformatics., 21 Suppl *1*, i266-273.

[81] Fundel, K., Guttler, D., Zimmer, R. & Apostolakis, J. (2005). A simple approach for protein name identification: prospects and limits. *BMC Bioinformatics*., 6 Suppl *1*, S15.

[82] de Bruijn B., Martin J. (2002). Getting to the (c)ore of knowledge: mining biomedical literature. *International Journal of Medical Informatics*., *67*, 7-18.

[83] Gaizauskas, R., Demetriou, G., Artymiuk, P. J. & Willett, P. (2003). Protein structures and information extraction from biological texts: the PASTA system. *Bioinformatics*., *19*, 135-143.

[84] Daniel, M. M., Hsinchun, C., Hua, S., Byron, B. M. (2004). Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser. *Bioinformatics*., *20*, 3370.

[85] Crim, J., McDonald, R. & Pereira, F. (2005). Automatically annotating documents with normalized gene lists. *BMC Bioinformatics*., *6*, Suppl *1*, S13.

[86] Hakenberg, J., Plake C., Royer L., Strobelt H., Leser U., Schroeder M. (2008). Gene mention normalization and interaction extraction with context models and sentence motifs. *Genome Biology*., *9*, Suppl *2*, S14.

[87] Huang, M., Ding, S., Wang, H. & Zhu, X. (2008). Mining physical protein-protein interactions from the literature. *Genome Biology*., *9*, Suppl *2*, S12.

[88] Krallinger, M., Valencia, A. & Hirschman, L. (2008). Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biolgy*., *9*, Suppl *2*, S8.

[89] Zhou, W., Torvik, V. I. & Smalheiser, N. R. (2006). ADAM: another database of abbreviations in MEDLINE. *Bioinformatics*., *22*, 2813-2818.

[90] Hoffmann, R., Krallinger, M., Andres, E., Tamames, J., Blaschke, C. & Valencia, A. (2005). Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci STKE.*, pe21.

[91] Skusa, A., Ruegg, A. & Kohler, J. (2005). Extraction of biological interaction networks from scientific literature. *Briefings in Bioinformatics*., *6*, 263-276.

[92] Cohen, K. B. & Hunter, L. (2008). Getting started in text mining. *PLoS Computational Biology*., *4*, e20.

[93] Stapley, B. J. & Benoit, G. (2000). Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in medline abstracts. *Pacific Symposium on Biocomputing*., *5*, 526-537.

[94] Donaldson, I., Martin, J., de Bruijn, B., Wolting, C., Lay, V. & Tuekam, B. et al. (2003). PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*., *4*, 11.

[95] Hoffmann, R. & Valencia, A. (2004). A gene network for navigating the literature. *Nature Genetics*., *36*, 664.

[96] Stephens, M., aplakal, M., Mukhopadhyay, S., Raje, R. & Mostafa, J. (2001). Detecting gene relations from MEDLINE abstracts. *Pacific Symposium on Biocomputing*., *6*, 483-496.

[97] Cooper, J. W. & Kershenbaum, A. (2005). Discovery of protein-protein interactions using a combination of linguistic, statistical and graphical information. *BMC Bioinformatics*., *6*, 143.

[98] Ray, S. & Craven, M. (2005). Learning statistical models for annotating proteins with function information using biomedical text. *BMC Bioinformatics*., *6*, Suppl *1*, S18.

[99] Jelier, R., Jenster, G., Dorssers, L. C., van der Eijk, C. C., van Mulligen, E. M. & Mons, B. et al. (2005). Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. *Bioinformatics*., *21*, 2049-2058.

[100] Jenssen, T. K., Laegreid, A., Komorowski, J. & Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*., *28*, 21-28.

[101] Alako, B. T., Veldhoven, A., van Baal, S., Jelier, R., Verhoeven, S. & Rullmann, T. et al. (2005). CoPub Mapper: mining MEDLINE based on search term co-publication. BMC *Bioinformatics*., *6*, 51.

[102] Becker, K. G., Hosack, D. A., Dennis, G., Jr.., Lempicki, R. A., Bright, T. J. & Cheadle, C., et al. (2003). PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics*., *4*, 61.

[103] Fang, Y. C., Huang, H. C. & Juan, H. F. (2008). MeInfoText: associated gene methylation and cancer information from text mining. *BMC Bioinformatics*, *9*, 22.

[104] Han, J. & Kamber, M. (2006). *Data Mining: Concepts and Techniques.,* Morgan Kaufmann.

[105] Yu, H., Hatzivassiloglou, V., Friedman, C., Rzhetsky, A. & Wilbur, W. J. (2002). *Automatic extraction of gene and protein synonyms from MEDLINE and journal articles. AMIA Symp,* 2002., 919-923.

[106] Huang, M., Zhu, X., Hao, Y., Payan, D. G., Qu, K. & Li, M. (2004). Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics.*, *20*, 3604-3612.

[107] Yu, H. & Agichtein, E. (2003). Extracting synonymous gene and protein terms from biological literature. *Bioinformatics.*, *19*, Suppl *1*, i340-349.

[108] Florence, H., Anthony, L. L. & Fred, E. C. (2004). Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. *Bioinformatics.*, *20*, 557.

[109] Divoli, A. & Attwood, T. K. (2005). BioIE: extracting informative sentences from the biomedical literature. *Bioinformatics.*, *21*, 2138-2139.

[110] Marcus, M. P., Santorini, B. & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics.*, *19*, 313-330.

[111] Dernatas, E. & Kokkinakis, G. (1995). Automatic stochastic tagging of natural language texts. *Computational Linguistics., 21*, 137-163.

[112] Kupiec, J. (1992). Robust part-of-speech tagging using a Hidden Markov Model. *Computer Speech and Language.*, *6*.

[113] Saric, J., Jensen, L. J., Ouzounova, R., Rojas, I. & Bork, P. (2005). Extraction of regulatory gene/protein networks from Medline. *Bioinformatics*.

[114] Temkin, J. M. & Gilder, M. R. (2003). Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics.*, *19*, 2046-2053.

[115] Rzhetsky, A., Iossifov, I., Koike, T., Krauthammer, M., Kra, P. & Morris, M. et al. (2004). GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics.*, *37*, 43-53.

[116] Li, Q. & Wu, Y. F. (2006). Identifying important concepts from medical documents. *J Biomed Inform.*, *39*, 668-679.

[117] Crystal, D. (1997). *The Cambridge Encyclopedia of Languages (2nd ed.)*, Cambridge University Press, Cambridge.

[118] Santos, C., Eggle, D. & States, D. J. (2005). Wnt pathway curation using automated natural language processing: combining statistical methods with partial and full parse for knowledge extraction. *Bioinformatics.*, *21*, 1653-1658.

[119] Uramoto, N., Matsuzawa, H., Nagano, T., Murakami, A., Takeuchi, H. & Takeda, K. (2004). A text-mining system for knowledge discovery from biomedical documents. *IBM Systems Journal.*, *43*, 516-533.

[120] Rindflesch, T. C., Rajan, J. & Lawrence Hunter, L. (2000). *Extracting molecular binding relationships from biomedical text.* 6th Applied Natural Language Processing Conference, 188-195.

[121] Rinaldi, F., Schneider, G., Kaljurand, K., Hess, M., Andronis, C. & Konstandi, O. et al. (2007). Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach. *Artificial Intelligence in Medicine.*, *39*, 127-136.

[122] Friedman, C. (2000). A Broad Coverage Natural Language Processing System. *American Medical Informatics Association Symposium*, 270 - 274.

[123] Novichkova, S., Egorov, S. & Daraselia, N. (2003). MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics*., *19*, 1699-1706.

[124] Allen, J. (1994). *Natural language understanding,* Benjamin-Cummings Publishing Company, New York.

[125] Sells, P. (1984). *Lectures on contemporary syntactic theories,* C S L I Publications.

[126] Chiang, J. H. & Yu, H. C. (2003). MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. *Bioinformatics*., *19*, 1417-1422.

[127] Kim, J. D., Ohta, T., Tateisi, Y. & Tsujii, J. (2003). GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, *19*, i180-i182.

[128] David, P. A. C., Bernard, F. B., William, B. L. & David, T. J. (2004). BioRAT: extracting biological information from full-length papers. *Bioinformatics*, *20*, 3206.

[129] Cunningham, H. (2000). *Software Architecture for Language Engineering.* Department of Computer Science. pp. 244. University of Sheffield.

[130] Chen, H. & Sharp, B. M. (2004). Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*., *5*, 147.

[131] Brants, T. (2000). *TnT - a statistical part-of-speech tagger.* 6th Applied Natural Language Processing Conference.

[132] Chiang, J. H., Yu, H. C., Hsu, H. J. (2004). GIS: a biomedical text-mining system for gene information discovery. *Bioinformatics*., *20*, 120.

[133] Nasukawa, T. & Nagono, T. (2001). Text analysis and knowledge mining system. *IBM Systems Journal*., *40*, 967-984.

[134] Karopka, T., Scheel, T., Bansemer, S. & Glass, A. (2004). Automatic construction of gene relation networks using text mining and gene expression data. Medical Informatics and the Internet in *Medicine*., *29*, 169-183.

[135] Neff, M. S., Byrd, R. J. & Boguraev, B. K. (2004). The Talent system: TEXTRACT architecture and data model. *Natural Language Engineering*., *10*, 307-326.

[136] Cooper, J. & Byrd R. (1998). *Lexical navigation: visually prompted query refinement.* ACM Digital Libraries Conference.

[137] Jang, H., Lim, J., Lim, J. H., Park, S. J., Lee, K. C. & Park, S. H. (2006). Finding the evidence for protein-protein interactions from PubMed abstracts. *Bioinformatics*., *22*, e220-226.

[138] Malik, R., Franke, L. & Siebes, A. (2006). Combination of text-mining algorithms increases the performance. *Bioinformatics*., *22*, 2151-2157.

[139] Cussens, J. & Nédellec, C. (eds) (2005). P*roceedings of the 4th Learning Language in Logic Workshop,* (LLL05), Bonn.

[140] Mika, S. & Rost, B. (2004). NLProt: extracting protein names and sequences from papers. *Nucleic Acids Research*., *32*, W634-637.

[141] Horn, F., Lau, A. L. & Cohen, F. E. (2004). Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. *Bioinformatics*., *20*, 557-568.

[142] Schneider, G., Rinaldi, F. & Dowdall, J. (2004). *Fast., deep-linguistic statistical dependency parsing.* 20th International Conference on Computational Linguistics. Association of Computational Linguistics, University of Geneva, Switzerland.

[143] Reynar, J., Ratnaparkhi, A. (1997). *A maximum entrophy approach to identifying sentence boundaries.* Fifth Conference on Applied Natural Language Processing, Washington, DC: University of Pennsylvania.

[144] Ratnaparkhi, A. (1996). *A Maximum Entropy Model for Part-of-Speech Tagging.* Conference on Empirical Methods in Natural Language Processing., 133-142.

[145] Minnen, G., Carroll, J. & Pearce, D. (2001). Applied morphological processing of English. *Natural Language Engineering*., *7*, 207-223.

[146] Mikheev, A. (1997). Automatics rule induction for unknown word guessing. *Computational Linguistics*., *23*, 405-423.

[147] Feng, C., Yamashita, F. & Hashida, M. (2007). Automated extraction of information from the literature on chemical-CYP3A4 interactions. *Journal of Chemical Information and Modeling*., *47*, 2449-2455.

[148] Ling, M. H., Lefevre, C., Nicholas, K. R. & Lin, F. (2007). *Re-construction of Protein-Protein Interaction Pathways by Mining Subject-Verb-Objects Intermediates.,* Second IAPR Workshop on Pattern Recognition in Bioinformatics (PRIB 2007). Springer-Verlag, Singapore.

[149] Liu, H. & Lieberman, H. (2005). *Metafor: visualizing stories as code.* 10th International Conference on Intelligent User Interfaces.

[150] Ling, M. H. (2006). An Anthological Review of Research Utilizing MontyLingua, a Python-Based End-to-End Text Processor. *The Python Papers*, *1*, 5-12.

[151] Chen, L. (2006). *Automatic construction of domain-specific concept structures.* Technischen Universitat Darmstadt.

[152] van Eck, N. J. & van den Berg, J. (2005). *A novel algorithm for visualizing concept associations.* 16th International Workshop on Database and Expert System Applications, (DEXA'05).

[153] Lee, H., Yi, G. S. & Park, J. C. (2008). E3Miner: a text mining tool for ubiquitin-protein ligases. *Nucleic Acids Research*., *36*, W416-422.

[154] Kim, S., Shin, S. Y., Lee, I. H., Kim, S. J., Sriram, R. & Zhang, B. T. (2008). PIE: an online prediction system for protein-protein interactions from text. *Nucleic Acids Research, 36*, W411-415.

[155] Plake, C., Hakenberg, J. & Leser, U. Optimizing syntax patterns for discovering protein-protein interactions. *ACM Symposium on Applied Computing*., 187-192. ACM Press (2005).

[156] Barnickel, T., Weston, J., Collobert, R., Mewes, H. W. & Stumpflen, V. (2009) Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts. *PLoS ONE*., *4*, e6393.

[157] Jiao, D. & Wild, D. J. (2009). Extraction of CYP chemical interactions from biomedical literature using natural language processing methods. *J Chem Inf Model*, *49*, 263-269.

[158] National Library of Medicine. (2003). UMLS Knowledge Sources (14th ed.).

[159] Friedman, C., Alderson, P. O., Austin, J. H., Cimino, J. J. & Johnson, S. B. (1994). A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*., *1*, 161-174.

[160] Sleator, D. & Temperley, D. (1991). *Parsing English with a Link Grammar.* Third International Workshop on Parsing Technologies.

[161] Hao, Y., Zhu, X., Huang, M. & Li, M. (2005). Discovering patterns to extract protein-protein interactions from the literature: Part II. *Bioinformatics*., *21*, 3294-3300.

[162] von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M. & Foglierini, M. et al. (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*., *33*, D433-437.

[163] Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A. & Mazo, I. (2004). Extracting human protein interactions from MEDLINE using a full-sentence parser. Bioinformatics., *20*, 604-611.

[164] Yakushiji, A., Tateisi, Y., Miyao, Y. & Tsujii, J. (2001). Event extraction from biomedical papers using a full parser. *Pacific Symposium on Biocomputing*., *6*, 408-419.

[165] Chen, E. S., Hripcsak, G., Xu, H., Markatou, M. & Friedman, C. (2008). Automated Acquisition of Disease Drug Knowledge from Biomedical and Clinical Documents: An Initial Study. *Journal of the American Medical Informatics Association.*, *15*, 87-98.

[166] Osborne, J. D., Lin, S., Zhu, L. & Kibbe, W. A. (2007). Mining biomedical data using MetaMap Transfer (MMtx) and the Unified Medical Language System (UMLS). *Methods in Molecular Biology.*, *408*, 153-169.

[167] Tiffin, N., Kelso, J. F., Powell, A. R., Pan, H., Bajic, V. B. & Hide, W. A. (2005). Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Research*., *33*, 1544-1552.

[168] Aerts, S., Haeussler, M., van Vooren, S., Griffith, O. L., Hulpiau, P. & Jones, S. J. et al. (2008). Text-mining assisted regulatory annotation. *Genome Biology*., *9*, R31.

[169] Hu, Z. Z., Narayanaswamy, M., Ravikumar, K. E., Vijay-Shanker, K. & Wu, C. H. (2005). Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics*., *21*, 2759-2765.

[170] Shah, P. K., Jensen, L. J., Boue, S. & Bork, P. (2005). Extraction of transcript diversity from scientific literature. *PLoS Computational Biology*., *1*, e10.

[171] Yang, H., Nenadic, G. & Keane, J. A. (2008). Identification of transcription factor contexts in literature using machine learning approaches. *BMC Bioinformatics*., 9 Suppl *3*, S11.

[172] Xu, H., Anderson, K., Grann, V. & Friedman, C. (2004). Facilitating cancer research using natural languge procesiong of pathological reports. 11th World Congress on *Medical Informatics*., 565-569.

[173] Yuryev, A., Mulyukov, Z., Kotelnikova, E., Maslov, S., Egorov, S. & Nikitin, A. et al. (2006). Automatic pathway building in biological association networks. *BMC Bioinformatics*., *7*, 171.

[174] Matsuzawa, H. & Fukuda, T. (2000). *Mining structured association patterns from database.* 4th Pacific and Asia International Conference on Knowledge Discovery and Data Mining (PAKDD-2000)., 233-244.

[175] Miyao, Y., Sagae, K., Saetre, R., Matsuzaki, T. & Tsujii, J. (2009). Evaluating contributions of natural language parsers to protein-protein interaction extraction. *Bioinformatics., 25*, 394-400.

[176] Rodriguez-Esteban, R., Iossifov, I. & Rzhetsky, A. (2006). Imitating Manual Curation of Text-Mined Facts in Biomedicine. *PLoS Comput Biol., 2.*

[177] Alex, B., Grover, C., Haddow, B., Kabadjov, M., Klein, E. & Matthews, M. et al. (2008). Assisted curation: does text mining really help? *Pac Symp Biocomput*, 556-567.

[178] Swanson, D. R. (1990). Somatomedin C and arginine: implicit connections between mutually isolated literatures. *Perspectives in Biology and Medicine*., *33*, 157-186.

[179] Swanson, D. R. (1990). Medical literature as a potential source of new knowledge. *Bulletin of the Medical Library Association.*, *78*, 29-37.

[180] Weeber, M., Kors, J. A. & Mons, B. (2005). Online tools to support literature-based discovery in the life sciences. *Briefings in Bioinformatics*., *6*, 277-286.

[181] Srinivasan, P., Libbus, B. & Sehgal, A. K. (2004). Mining MEDLINE: postulating a beneficial role for Curcumin Longa in retinal diseases. BioLink Linking Biological Literature, *Ontologies, and Databases*., 33-40.

[182] Bekhuis, T. (2006). Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomedical Digital Libraries*, *3*, 2.

[183] Jelier, R., Schuemie, M. J., Veldhoven, A., Dorssers, L. C., Jenster, G. & Kors, J. A. (2008). Anni 2.*0*, a multipurpose text-mining tool for the life sciences. *Genome Biology*., *9*, R96.

[184] Smalheiser, N. R., Torvik, V. I. & Zhou, W. (2009). Arrowsmith two-node search interface: A tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. *Computer Methods and Programs in Biomedicine*., *94*, 190-197.

[185] Sarkar, I. N. & Agrawal, A. (2006). Literature based discovery of gene clusters using phylogenetic methods. *AMIA Annu Symp Proc*: 689-693.

[186] Hristovski, D., Peterlin, B., Mitchell, J. A. & Humphrey, S. M. (2005). Using literature-based discovery to identify disease candidate genes. *Int J Med Inform*., *74*, 289-298.

[187] Yetisgen-Yildiz, M. & Pratt, W. (2006). Using statistical and knowledge-based approaches for literature-based discovery. *J Biomed Inform*., *39*, 600-611.

[188] Colosimo, M. E., Morgan, A. A., Yeh, A. S., Colombe, J. B. & Hirschman, L. (2005). Data preparation and interannotator agreement: BioCreAtIvE task 1B. *BMC Bioinformatics*., *6*, Suppl *1*, S12.

[189] Wilbur, W. J., Rzhetsky, A. & Shatkay, H. (2006). New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*., *7*, 356.

[190] Tanabe, L., Xie, N., Thom, L. H., Matten, W. & Wilbur, W. J. (2005). GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*., *6*, Suppl *1*, S3.

[191] Pustejovsky, J., Castaño, J., Saurí, R., Rumshisky, A., Zhang, J. & Luo, W. (2002). *Medstract: Creating Large-scale Information Servers for Biomedical Libraries.* ACL 2002 Workshop on Natural Language Processing in the Biomedical Domain.

[192] Collier, N., Park, H. S., Ogata, N., Tateishi, Y., Nobata, C. & Ohta, T. et al. (1999). *The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers.* Ninth Conference of the European Chapter of the Association for Computational Linguistics.

[193] Ohta, T., Tateisi, Y., Mima, H. & Tsujii, J. (2002). *The GENIA corpus: an annotated research abstract corpus in molecular biology domain.* Human Language Technology Conference.

[194] Cussens, J. & Dzeroski, S. (eds) (2000). *Learning Languages in Logic.* Springer, Berlin, Heidelberg, New York, Barcelona, Hong Kong, London, Milan, Paris, Singapore, Tokyo.

[195] Vincze, V., Szarvas, G., Farkas, R., Mora, G. & Csirik, J. (2008). The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, *9*, Suppl *11*, S9.

[196] Pyysalo, S., Airola, A., Heimonen, J., Bjorne, J., Ginter, F. & Salakoski, T. (2008). Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, *9*, Suppl *3*, S6.

[197] Zhang, Z., Tang, S. & Ng, S. K. (2005). Towards discovering disease-specific gene networks from online literature. *Advances in Bioinformatics and Computational Biology.* 3rd Asia-Pacific Bioinformatics Conference, 161-169.

[198] Chatr-aryamontri, A., Kerrien, S., Khadake, J., Orchard, S., Ceol, A. & Licata, L. et al. (2008). MINT and IntAct contribute to the Second BioCreative challenge: serving the text-mining community with high quality molecular interaction data. *Genome Biology., 9* Suppl *2*, S5.

[199] Lee, L. C., Horn, F. & Cohen, F. E. (2007). Automatic Extraction of Protein Point Mutations Using a Graph Bigram Association. *PLoS Comput Biol*, *3*, e16.

[200] Maguitman, A. G., Rechtsteiner, A., Verspoor, K., Strauss, C. E. & Rocha, L. M. (2006). Large-scale testing of bibliome informatics using Pfam protein families. *Pac Symp Biocomput*: 76-87.

[201] Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U. & Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, *32*, D449-451.

[202] Miotto, O., Tan, T. W. & Brusic, V. (2005). Supporting the curation of biological databases with reusable text mining. *Genome Inform*, *16*, 32-44.

[203] Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M. & Bantoft, K. et al. (2005). The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Research*., *33*, D418-424.

[204] Zhou, D. & He, Y. (2008). Extracting interactions between proteins from the literature. *Journal of Biomedical Informatics*., *41*, 393-407.

[205] Scott, M., Lu, G., Hallett, M. & Thomas, D. Y. (2004). The Hera database and its use in the characterization of endoplasmic reticulum proteins. *Bioinformatics*., *20*, 937-944.

[206] Basu, S., Bremer, E., Zhou, C. & Bogenhagen, D. F. (2006). MiGenes: a searchable interspecies database of mitochondrial proteins curated using gene ontology annotation. *Bioinformatics*., *22*, 485-492.

[207] Martinez-Bueno, M., Molina-Henares, A. J., Pareja, E., Ramos, J. L. & Tobes, R. (2004). BacTregulators: a database of transcriptional regulators in bacteria and archaea. *Bioinformatics*., *20*, 2787-2791.

[208] Percudani, R. & Peracchi, A. (2009). The B6 database: a tool for the description and classification of vitamin B6-dependent enzymatic activities and of the corresponding protein families. *BMC Bioinformatics*., *10*, 273.

[209] Blaineau, S. V. & Aouacheria, A. (2009). BCL2DB: moving 'helix-bundled' BCL-2 family members to their database. *Apoptosis*., *14*, 923-925.

[210] Holliday, G. L., Bartlett, G. J., Almonacid, D. E., O'Boyle, N. M., Murray-Rust, P. & Thornton, J. M. et al. (2005). MACiE: a database of enzyme reaction mechanisms. *Bioinformatics*., *21*, 4315-4316.

[211] Mao, C., Qiu, J., Wang, C., Charles, T. C. & Sobral, B. W. (2005). NodMutDB: a database for genes and mutants involved in symbiosis. *Bioinformatics., 21*, 2927-2929.

[212] Wood, D. L., Miljenovic, T., Cai, S., Raven, R. J., Kaas, Q. & Escoubas, P. et al. (2009). ArachnoServer: a database of protein toxins from spiders. *BMC Genomics*., *10*, 375.

[213] Testa, O. D., Moutevelis, E. & Woolfson, D. N. (2009). CC+: a relational database of coiled-coil structures. *Nucleic Acids Res*., *37*, D315-322.

[214] Li, Y. & Chen, Z. (2008). RAPD: a database of recombinantly-produced antimicrobial peptides. *FEMS Microbiol Lett.*, *289*, 126-129.

[215] Jacobs, G. H., Chen, A., Stevens, S. G., Stockwell, P. A., Black, M. A. & Tate, W. P. et al. (2009). Transterm: a database to aid the analysis of regulatory sequences in mRNAs. *Nucleic Acids Res.*, *37*, D72-76.

[216] Jayakanthan, M., Muthukumaran, J., Chandrasekar, S., Chawla, K., Punetha, A. & Sundar, D. (2009). ZifBASE: a database of zinc finger proteins and associated resources. *BMC Genomics.*, *10*, 421.

[217] Kim, C., Kwon, S., Lee, G., Lee, H., Choi, J. & Kim, Y. et al. (2009). A database for allergenic proteins and tools for allergenicity prediction. *Bioinformation.*, *3*, 344-345.

[218] Gao, J., Agrawal, G. K., Thelen, J. J. & Xu, D. (2009). P3DB: a plant protein phosphorylation database. *Nucleic Acids Res.*, *37*, D960-962.

[219] Encinar, J. A., Fernandez-Ballester, G., Sanchez, I. E., Hurtado-Gomez, E., Stricher, F. & Beltrao, P. et al. (2009). ADAN: a database for prediction of protein-protein interaction of modular domains mediated by linear motifs. *Bioinformatics.*, *25*, 2418-2424.

[220] Magkrioti, C. K., Spyropoulos, I. C., Iconomidou, V. A., Willis, J. H. & Hamodrakas, S. J. (2004). cuticleDB: a relational database of Arthropod cuticular proteins. *BMC Bioinformatics.*, *5*, 138.

[221] George, R. A., Spriggs, R. V., Thornton, J. M., Al-Lazikani, B. & Swindells, M. B. (2004). SCOPEC: a database of protein catalytic domains. *Bioinformatics., 20,* Suppl *1*, I130-I136.

[222] Li, B. & Gallin, W. J. (2004). VKCDB: voltage-gated potassium channel database. *BMC Bioinformatics.*, *5*, 3.

[223] Vucetic, S., Obradovic, Z., Vacic, V., Radivojac, P., Peng, K., Iakoucheva, L. M., et al. (2005). DisProt: a database of protein disorder. *Bioinformatics.*, *21*, 137-140.

[224] Guo, A., He, K., Liu, D., Bai, S., Gu, X. & Wei, L. et al. (2005). DATF: a database of Arabidopsis transcription factors. *Bioinformatics.*, *21*, 2568-2569.

[225] Gao, G., Zhong, Y., Guo, A., Zhu, Q., Tang, W. & Zheng, W. et al. (2006). DRTF: a database of rice transcription factors. *Bioinformatics.*, *22*, 1286-1287.

[226] Tung, M. & Gallagher, D. T. (2009). The Biomolecular Crystallization Database Version *4*, expanded content and new features. *Acta Crystallogr D Biol Crystallogr.*, *65*, 18-23.

[227] Sun, Q., Zybailov, B., Majeran, W., Friso, G., Olinares, P. D. & van Wijk, K. J. (2009). PPDB, the Plant Proteomics Database at Cornell. *Nucleic Acids Res.*, *37*, D969-974.

[228] Chandra, N. R., Kumar, N., Jeyakani, J., Singh, D. D., Gowda, S. B. , &Prathima, M. N. (2006). Lectindb: a plant lectin database. *Glycobiology.*, *16*, 938-946.

[229] Nogales-Cadenas, R., Abascal, F., Diez-Perez, J., Carazo, J. M. & Pascual-Montano, A. (2009). CentrosomeDB: a human centrosomal proteins database. *Nucleic Acids Res.*, *37*, D175-180.

[230] Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M. & Hannay, T. et al. (2009). PID: the Pathway Interaction Database. *Nucleic Acids Res.*, *37*, D674-679.

[231] McDowall, M. D., Scott, M. S. & Barton, G. J. (2009). PIPs: human protein-protein interaction prediction database. *Nucleic Acids Res.*, *37*, D651-656.

[232] Zhao, X. M., Zhang, X. W., Tang, W. H. & Chen, L. (2009). FPPI: Fusarium graminearum Protein-Protein Interaction Database. *J Proteome Res.*

[233] Chatr-aryamontri, A., Ceol, A., Peluso, D., Nardozza, A., Panni, S. & Sacco, F. et al. (2009). VirusMINT: a viral protein interaction database. *Nucleic Acids Res., 37*, D669-673.

[234] Chen, J. Y., Mamidipalli, S. & Huan, T. (2009). HAPPI: an online database of comprehensive human annotated and predicted protein interactions. *BMC Genomics*., 10 Suppl *1*, S16.

[235] Andres Leon, E., Ezkurdia, I., Garcia, B., Valencia, A. & Juan, D. (2009). EcID. A database for the inference of functional interactions in E. coli. *Nucleic Acids Res.*, *37*, D629-635.

[236] Lin, C. Y., Chen, C. L., Cho, C. S., Wang, L. M., Chang, C. M. & Chen, P. Y. et al. (2005). hp-DPI: Helicobacter pylori database of protein interactomes--embracing experimental and inferred interactions. *Bioinformatics*., *21*, 1288-1290.

[237] Goll, J., Rajagopala, S. V., Shiau, S. C., Wu, H., Lamb, B. T. & Uetz, P. (2008). MPIDB: the microbial protein interaction database. *Bioinformatics*., *24*, 1743-1744.

[238] Pawlicki, S., Le Bechec, A. & Delamarche, C. (2008). AMYPdb: a database dedicated to amyloid precursor proteins. *BMC Bioinformatics*., *9*, 273.

[239] Theodoropoulou, M. C., Bagos, P. G., Spyropoulos, I. C., Hamodrakas, S. J. (2008). gpDB: a database of GPCRs., G-proteins, effectors and their interactions. *Bioinformatics*., *24*, 1471-1472.

[240] Chuan Tong, J., Meng Song, C., Thiam Joo Tan, P. & Chee Ren, E. A AS. (2008). BEID: Database for sequence-structure-function information on antigen-antibody interactions. *Bioinformation*., *3*, 58-60.

[241] Yimeng, D., Pierre-FranÁois, B., Gianluca, P., Yann, P., James, N. & Pierre, B. (2004). ICBS: a database of interactions between protein chains mediated by ≤-sheet formation. *Bioinformatics*., *20*, 2767.

[242] Beuming, T., Skrabanek, L., Niv, M. Y., Mukherjee, P. & Weinstein, H. (2005). PDZBase: a protein-protein interaction database for PDZ-domains. *Bioinformatics*., *21*, 827-828.

[243] Dou, Y., Baisnee, P. F., Pollastri, G., Pecout, Y., Nowick, J. & Baldi, P. (2004). ICBS: a database of interactions between protein chains mediated by beta-sheet formation. *Bioinformatics*., *20*, 2767-2777.

[244] Yang, C. Y., Chang, C. H., Yu, Y. L., Lin, T. C., Lee, S. A. & Yen, C. C., et al. (2008). PhosphoPOINT: a comprehensive human kinase interactome and phospho-protein database. *Bioinformatics*., *24*, i14-20.

[245] Song, S., Huang, Y., Wang, X., Wei, G., Qu, H., Wang, W. & et al. (2009). HRGD: a database for mining potential heterosis-related genes in plants. *Plant Mol Biol., 69*, 255-260.

[246] Richardson, C. J., Gao, Q., Mitsopoulous, C., Zvelebil, M., Pearl, L. H. & Pearl, F. M. (2009). MoKCa database--mutations of kinases in cancer. *Nucleic Acids Res*., *37*, D824-831.

[247] Sagar, S., Kaur, M., Dawe, A., Seshadri, S. V., Christoffels, A. & Schaefer, U. et al. (2008). DDESC: Dragon database for exploration of sodium channels in human. *BMC Genomics*., *9*, 622.

[248] Miranda-Saavedra, D., De, S., Trotter, M. W., Teichmann, S. A. & Gottgens, B. (2009). BloodExpress: a database of gene expression in mouse haematopoiesis. *Nucleic Acids Res*., *37*, D873-879.

[249] Mao, F., Dam, P., Chou, J., Olman, V. & Xu, Y. (2009). DOOR: a database for prokaryotic operons. *Nucleic Acids Res.*, *37*, D459-463.

[250] Liu, B. & Pop, M. (2009). ARDB--Antibiotic Resistance Genes Database. *Nucleic Acids Res.*, *37*, D443-447.

[251] Dinger, M. E., Pang, K. C., Mercer, T. R., Crowe, M. L., Grimmond, S. M. & Mattick, J. S. (2009). NRED: a database of long noncoding RNA expression. *Nucleic Acids Res., 37*, D122-126.

[252] Essack, M., Radovanovic, A., Schaefer, U., Schmeier, S., Seshadri, S. V. & Christoffels, A. et al. (2009). DDEC: Dragon database of genes implicated in esophageal cancer. *BMC Cancer.*, *9*, 219.

[253] Kim, C. K., Kim, J. S., Lee, G. S., Park, B. S. & Hahn, J. H. (2008). PlantGM: a database for genetic markers in rice (Oryza sativa) and Chinese cabbage (Brassica rapa). *Bioinformation.*, *3*, 61-62.

[254] Ding, G., Lorenz, P., Kreutzer, M., Li, Y. & Thiesen, H. J. (2009). SysZNF: the C2H2 zinc finger gene database. *Nucleic Acids Res.*, *37*, D267-273.

[255] Boby, T., Patch, A. M. & Aves, S. J. (2005). TRbase: a database relating tandem repeats to disease genes for the human genome. *Bioinformatics.*, *21*, 811-816.

[256] Sakharkar, M. K. & Kangueane, P. (2004). Genome SEGE: a database for 'intronless' genes in eukaryotic genomes. *BMC Bioinformatics.*, *5*, 67.

[257] Brockman, J. M., Singh, P., Liu, D., Quinlan, S., Salisbury, J. & Graber, J. H. (2005). PACdb: PolyA Cleavage Site and 3'-UTR Database. *Bioinformatics.*, *21*, 3691-3693.

[258] Shimada, M. K., Matsumoto, R., Hayakawa, Y., Sanbonmatsu, R., Gough, C. & Yamaguchi-Kabata, Y. et al. (2009). VarySysDB: a human genetic polymorphism database based on all H-InvDB transcripts. *Nucleic Acids Res.*, *37*, D810-815.

[259] Shionyu, M., Yamaguchi, A., Shinoda, K., Takahashi, K. & Go, M. (2009). AS-ALPS: a database for analyzing the effects of alternative splicing on protein structure, interaction and network in human and mouse. *Nucleic Acids Res.*, *37*, D305-309.

[260] Koscielny, G., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Riethoven, J. J. & Nardone, F. et al. (2009). ASTD: The Alternative Splicing and Transcript Diversity database. *Genomics., 93*, 213-220.

[261] Duan, S., Zhang, W., Cox, N. J. & Dolan, M. E. (2008). FstSNP-HapMap*3*, a database of SNPs with high population differentiation for HapMap3. *Bioinformation.*, *3*, 139-141.

[262] Ackermann, A. A., Carmona, S. J. & Aguero, F. (2009). TcSNP: a database of genetic variation in Trypanosoma cruzi. *Nucleic Acids Res.*, *37*, D544-549.

[263] Palaniswamy, S. K., Jin, V. X., Sun, H. & Davuluri, R. V. (2005). OMGProm: a database of orthologous mammalian gene promoters. *Bioinformatics.*, *21*, 835-836.

[264] Kim, J., Seo, J., Lee, Y. S. & Kim, S. (2005). TFExplorer: integrated analysis database for predicted transcription regulatory elements. *Bioinformatics.*, *21*, 548-550.

[265] Gallo, S. M., Li, L., Hu, Z. & Halfon, M. S. (2006). REDfly: a Regulatory Element Database for Drosophila. *Bioinformatics.*, *22*, 381-383.

[266] Morris, R. T., O'Connor, T. R. & Wyrick, J. J. (2008). Osiris: an integrated promoter database for Oryza sativa L. *Bioinformatics., 24*, 2915-2917.

[267] Rushton, P. J., Bokowiec, M. T., Laudeman, T. W., Brannock, J. F., Chen, X., Timko, M. P. (2008). TOBFAC: the database of tobacco transcription factors. *BMC Bioinformatics.*, *9*, 53.

[268] Kaas, Q., Westermann, J. C., Halai, R., Wang, C. K., Craik, D. J. (2008). ConoServer, a database for conopeptide sequences and structures. *Bioinformatics*., *24*, 445-446.

[269] O'Brien, E. A., Zhang, Y., Wang, E., Marie, V., Badejoko, W., Lang, B. F. et al. (2009). GOBASE: an organelle genome database. *Nucleic Acids Res*., *37*, D946-950.

[270] Maselli, V., Di Bernardo, D. & Banfi, S. (2008). CoGemiR: a comparative genomics microRNA database. *BMC Genomics*., *9*, 457.

[271] Lu, T., Huang, X., Zhu, C., Huang, T., Zhao, Q. & Xie, K. et al. (2008). RICD: a rice indica cDNA database resource for rice functional genomics. *BMC Plant Biol*., *8*, 118.

[272] Lim, D., Cho, Y.M ., Lee, K. T., Kang, Y., Sung, S. & Nam, J. et al. (2009). The Pig Genome Database (PiGenome): an integrated database for pig genome research. *Mamm Genome*., *20*, 60-66.

[273] Gauthier, J. P., Legeai, F., Zasadzinski, A., Rispe, C. & Tagu, D. (2007). AphidBase: a database for aphid genomic resources. *Bioinformatics*., *23*, 783-784.

[274] Cameron, R. A., Samanta, M., Yuan, A., He, D. & Davidson, E. (2009). SpBase: the sea urchin genome database and web site. *Nucleic Acids Res*., *37*, D750-754.

[275] Nystrom, J., Fierlbeck, W., Granqvist, A., Kulak, S. C. & Ballermann, B. J. (2009). A human glomerular SAGE transcriptome database. *BMC Nephrol*., *10*, 13.

[276] Lee, B. & Shin, G. (2009). CleanEST: a database of cleansed EST libraries. *Nucleic Acids Res., 37*, D686-689.

[277] Beldade, P., Rudd, S., Gruber, J. D. & Long, A. D. (2006). A wing expressed sequence tag resource for Bicyclus anynana butterflies., an evo-devo model. *BMC Genomics*., *7*, 130.

[278] Schlamp, K., Weinmann, A., Krupp, M., Maass, T., Galle, P. & Teufel, A. (2008). BlotBase: a northern blot database. *Gene.*, *427*, 47-50.

[279] Sherlock, G., Hernanadez-Boussard, T., Kasarskis, A., Binkley, G., Matese, J. C. & Dwight, S. S. et al. (2001). The Stanford microarray database. *Nucleic Acid Research*., *29*, 152-155.

[280] Markus, R., Srinivas, V., Samuel, A., Johan, S. & Jari, H. k. (2004). ACID: a database for microarray clone information. *Bioinformatics*., *20*, 2305.

[281] Singh, M. K., Srivastava, S., Raghava, G. P. & Varshney, G. C. (2006). HaptenDB: a comprehensive database of haptens, carrier proteins and anti-hapten antibodies. *Bioinformatics., 22*, 253-255.

[282] Lomize, M. A., Lomize, A. L., Pogozheva, I. D. & Mosberg, H. I. (2006). OPM: orientations of proteins in membranes database. *Bioinformatics*., *22*, 623-625.

[283] Zhang, S., Xia, X., Shen, J., Zhou, Y. & Sun, Z. (2008). DBMLoc: a Database of proteins with multiple subcellular localizations. *BMC Bioinformatics*., *9*, 127.

[284] Zheng, C. J., Zhou, H., Xie, B., Han, L. Y., Yap, C. W. & Chen, Y. Z. (2004). TRMP: a database of therapeutically relevant multiple pathways. *Bioinformatics., 20*, 2236-2241.

[285] Barrett, T. & Edgar, R. (2006). Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol*., *411*, 352-369.

[286] Lakshmanan, L. V. S., Sadri, F. & Subramanian, S. N. (2001). SchemaSQL - An extention to SQL for multidatabase interoperability. *ACM Transactions on Database Systems, 26*, 476-519.

[287] Wyss, C. M. & Robertson, E. L. (2005). Relational languages for metadata integration. *ACM Transactions on Database Systems*, *30*, 624-660.

[288] Fristensky, B. (2007). BIRCH: a user-oriented, locally-customizable, bioinformatics system. *BMC Bioinformatics*, *8*, 54.

[289] Garcia Castro, A., Chen, Y. P. & Ragan, M. A. (2005). Information integration in molecular bioscience. *Appl Bioinformatics*, *4*, 157-173.

[290] Zhou, W., Smalheiser, N. R. & Yu, C. (2006). A tutorial on information retrieval: basic terms and concepts. *J Biomed Discov Collab.*, *1*, 2.

[291] Hirschman, L., Park, J. C., Tsujii, J., Wong, L. & Wu, C. H. (2002). Accomplishments and challenges in literature data mining for biology. *Bioinformatics.*, *18*, 1553-1561.

[292] Biagini, R. E., Krieg, E. F., Pinkerton, L. E. & Hamilton, R. G. (2001). Receiver operating characteristics analyses of Food and Drug Administration-cleared serological assays for natural rubber latex-specific immunoglobulin E antibody. *Clinical and Diagnostic Laboratory Immunology.*, *8*, 1145-1149.

[293] Gjengsto, P., Paus, E., Halvorsen, O. J., Eide, J., Akslen, L. A. & Wentzel-Larsen, T. et al. (2005). Predictors of prostate cancer evaluated by receiver operating characteristics partial area index: a prospective institutional study. *Journal of Urology.*, *173*, 425-428.

[294] Margolis, D. J., Bilker, W., Boston, R., Localio, R., Berlin, J. A. (2002). Statistical characteristics of area under the receiver operating characteristic curve for a simple prognostic model using traditional and bootstrapped approaches. *Journal of Clinical Epidemiology.*, *55*, 518-524.

[295] Rosman, A. S. & Korsten, M. A. (2007). Application of summary receiver operating characteristics (sROC) analysis to diagnostic clinical testing. *Advances in Medical Science., 52*, 76-82.

[296] Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science.*, *240*, 1285-1293.

[297] Hersh, W. (2005). Evaluation of biomedical text-mining systems: lessons learned from information retrieval. *Briefings in Bioinformatics.*, *6*, 344-356.

[298] Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K. & Chetvernin, V. et al. (2006). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research.*, *34*, D173-180.

[299] Stolovitzky, G. A., Kundaje, A., Held, G. A., Duggar, K. H., Haudenschild, C. D. & Zhou, D. et al. (2005). Statistical analysis of MPSS measurements: application to the study of LPS-activated macrophage gene expression. *Proceedings of the National Academy of Science*, U S A., *102*, 1402-1407.

[300] Shi, L. & Campagne, F. (2005). Building a protein name dictionary from full text: a machine learning term extraction approach. *BMC Bioinformatics.*, *6*, 88.

[301] Rzhetsky, A., Zheng, T. & Weinreb, C. (2006). Self-correcting maps of molecular pathways. *PLoS ONE.*, *1*, e61.

[302] Wren, J. D. & Garner, H. R. (2004). Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. *Bioinformatics.*, *20*, 191-198.

[303] Ling, M. H. T., Leferve, C. & Nicholas, K. R. (2008). Filtering microarray correlations by statistical literature analysis yields potential hypotheses for lactation research. *The Python Papers.*, *3*, 4.

[304] Grover, C., Klein, E., Lascarides, A. & Lapata, M. (2002). *XML-based NLP Tools for Analysing and Annotating Medical Language.* Second International Workshop on NLP and XML (NLPXML-2002).

[305] Grover, C., Lapata, M. & Lascarides, A. (2003). A comparison of parsing technologies for the biomedical domain. *Natural Language Engineering*., *1*, 1-38.

[306] Ling, M. H. T., Leferve, C. & Nicholas, K. R. (2008). A Case Study where Parts-of-Speech Tagging Error Does Not Adversely Affect Extraction of Protein-Protein Interactions from Text. *The Python Papers*., *3*, 65-80.

[307] Voorhees, E. & Buckland, L. P. (eds) (2005). *The Fourteen Text REtrieval Conference Proceedings. National Institute of Standards and Technology (NIST).,* the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA), Gaithersburg, Maryland.

[308] Cano, C., Monaghan, T., Blanco, A., Wall, D. P. & Peshkin, L. (2009). Collaborative text-annotation resource for disease-centered relation extraction from biomedical text. *Journal of Biomedical Informatics*.

[309] Newman, D., Hettich, S., Blake, C., Merz, C. (1998). *UCI Repository of machine learning databases.* University of California, Department of Information and Computer Science, Irvine., CA.

[310] Kano, Y., Nguyen, N., Saetre, R., Yoshida, K., Miyao, Y. & Tsuruoka, Y. et al. (2008). Filling the gaps between tools and users: a tool comparator, using protein-protein interaction as an example. *Pacific Symposium on Biocomputing*: 616-627.

[311] Lourenco, A., Carreira, R., Carneiro, S., Maia, P., Glez-Pena, D. & Fdez-Riverola, F. et al. (2009). @Note: a workbench for biomedical text mining. *J Biomed Inform*., *42*, 710-720.

[312] Altman, R. B., Bergman, C. M., Blake, J., Blaschke, C., Cohen, A. & Gannon, F. et al. (2008). Text mining for biology--the way forward: opinions from leading scientists. *Genome Biology*., *9*, Suppl *2*, S7.

[313] Muller, M., Marko, K., Daumke, P., Paetzold, J., Roesner, A. & Klar, R. (2007). Biomedical data mining in clinical routine: expanding the impact of hospital information systems. *Medinfo*., *12*, 340-344.

[314] Caporaso, J. G., Deshpande, N., Fink, J. L., Bourne, P. E., Cohen, K. B. & Hunter, L. (2008). Intrinsic evaluation of text mining tools may not predict performance on realistic tasks. *Pacific Symposium on Biocomputing*, 640-651.

[315] Roberts, P. M. & Hayes, W. S. (2008). Information needs and the role of text mining in drug development. *Pacific Symposium on Biocomputing,* 592-603.

[316] Kabiljo, R., Clegg, A. B. & Shepherd, A. J. (2009). A realistic assessment of methods for extracting gene/protein interactions from free text. *BMC Bioinformatics*., *10*, 233.

[317] Roberts, P. M. (2006). Mining literature for systems biology. *Briefings in Bioinformatics*., *7*, 399-406.

[318] Couto, F. M., Silva, M. J., Lee, V., Dimmer, E., Camon, E. & Apweiler, R. et al. (2006). GOAnnotator: linking protein GO annotations to evidence text. *J Biomed Discov Collab.*, *1*, 19.

[319] Lussier, Y., Borlawsky, T., Rappaport, D., Liu, Y. & Friedman, C. (2006). PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing. *Pac Symp Biocomput,* 64-75.

[320] Natarajan, J. & Ganapathy, J. (2007). Functional gene clustering via gene annotation sentences., MeSH and GO keywords from biomedical literature. *Bioinformation.*, *2*, 185-193.

[321] Cakmak, A. & Ozsoyoglu, G. (2008). Discovering gene annotations in biomedical text databases. *BMC Bioinformatics*., *9*, 143.

[322] Jin, B., Muller, B., Zhai, C. & Lu, X. (2008). Multi-label literature classification based on the Gene Ontology graph. *BMC Bioinformatics*., *9*, 525.

[323] Abulaish, M. & Dey, L. (2007). Biological relation extraction and query answering from MEDLINE abstracts using ontology-based text mining. *Data & Knowledge Engineering*., *61*, 228.

[324] Baumgartner WA., Jr.., Lu Z., Johnson HL., Caporaso JG., Paquette J., Lindemann A., et al. (2008). Concept recognition for extracting protein interaction relations from biomedical text. *Genome Biology., 9,* Suppl *2*, S9.

[325] Heinrich, K. E., Berry, M. W. & Homayouni, R. (2008). Gene tree labeling using nonnegative matrix factorization on biomedical literature. *Computational Intelligence and Neuroscience*., Article ID: 276535.

[326] Spasic, I., Schober, D., Sansone, S. A., Rebholz-Schuhmann, D., Kell, D. B. & Paton, N. W. (2008).. Facilitating the development of controlled vocabularies for metabolomics technologies with text mining. *BMC Bioinformatics*,*a*, Suppl, *5*, S5.