From Biomedical Literature to Knowledge: Mining Protein-Protein Interactions

Deyu Zhou[1], Yulan He[1], and Chee Keong Kwoh[2]

[1] Informatics Research Centre, The University of Reading, Reading, UK, RG6 6BXd.zhou@reading.ac.uk, y.he@reading.ac.uk
[2] School of Computer Engineering, Nanyang Technological University, Singapore 639798 asckkwoh@ntu.edu.sg

Summary. To date, more than 16 million citations of published articles in biomedical domain are available in the MEDLINE database. These articles describe the new discoveries which accompany a tremendous development in biomedicine during the last decade. It is crucial for biomedical researchers to retrieve and mine some specific knowledge from the huge quantity of published articles with high efficiency. Researchers have been engaged in the development of text mining tools to find knowledge such as protein-protein interactions, which are most relevant and useful for specific analysis tasks. This chapter provides a road map to the various information extraction methods in biomedical domain, such as protein name recognition and discovery of protein-protein interactions. Disciplines involved in analyzing and processing unstructured-text are summarized. Current work in biomedical information extracting is categorized. Current challenges in the field are also presented and possible solutions are discussed.

1 Introduction

In post genomic science, proteins are recognized as elements in complex protein interaction networks. Hence protein-protein interactions play a key role in various aspects of the structural and functional organization of the cell. Knowledge about them unveils the molecular mechanisms of biological processes. However, most of this knowledge hides in published articles, scientific journals, books and technical reports. To date, more than 16 million citations of such articles are available in the MEDLINE [1] database. In parallel with these plain text information sources, many databases, such as DIP [2], BIND [3], IntAct [4] and STRING [5], have been built to store various types of information about protein-protein interactions. Nevertheless, data in these databases were mainly hand-curated to ensure their correctness and thus limited the speed in transferring textual information into searchable structure

data. Retrieving and mining such information from the literature is very complex due to the lack of formal structure in the natural-language narrative in these documents. Thus, automatically extracting information from biomedical text holds the promise of easily discovering large amounts of biological knowledge in computer-accessible forms.

Many systems [6–10], such as EDGAR [11], BioRAT [12], GeneWays [13] and so on, have been developed to accomplish this goal, but with limited success. Table 1 lists some popular online databases, systems, and tools relating to extraction of protein-protein interactions.

Table 1. Online databases, systems, tools relating to the extraction of protein-protein interactions.

	Description	*URL*
Online databases storing protein-protein interactions		
BIND	Biomolecular Interaction Network Database contains over 200,000 human-curated interactions.	www.bind.ca/
DIP	Database of Interacting Proteins catalogs experimentally determined interactions between proteins. Until now, it contains 55,732 interactions, combining information from various sources to create a single, stable set of protein-protein interactions.	dip.doe-mbi.ucla.edu/
HPRD	The Human Protein Reference Database [14] contains interaction networks for each protein in the human proteome. All the information in HPRD has been manually extracted from the literature by expert biologists who read, interpret and analyze the published data.	www.hprd.org/
HPID	Human Protein Interaction Database integrates the protein interactions in BIND, DIP and HPRD.	www.hpid.org/
IntAct	IntAct consists of a open source database system and analysis tools for protein interaction data. It now contains more than 100,000 curated binary molecular interactions.	www.ebi.ac.uk/intact/
MINT	Molecular INTeraction database [15] is a database storing interactions between biological molecules. It focuses on experimentally verified protein interactions with special emphasis on proteomes from mammalian organisms.	mint.bio.uniroma2.it/mint/
STRING	STRING, a database consisting of known and predicted protein-protein interactions, quantitatively integrates interaction data from several sources for a large number of organisms. It currently contains 736,429 proteins in 179 species.	string.embl.de/
Online protein-protein interaction information extraction systems		
BioRAT	BioRAT is a search engine and information extraction tool for biological research.	bioinf.cs.ucl.ac.uk/biorat/
GeneWays	GeneWays is a system for automatically extracting, analyzing, visualizing and integrating molecular pathway data from the literature. It focuses on interactions between molecular substances and actions, providing a graphical consensus view on these collected information.	geneways.genomecenter.columbia.edu/
MedScan	MedScan is a commercial system based on natural language processing technology for automatic extraction of biological facts from scientific literature such as MEDLINE abstracts, and internal text documents.	www.ariadnegenomics.com/products/medscan.html
Online tools for biomedical literature mining		
iProLINK	iProLINK is a resource to facilitate text mining in the area of literature-based database curation, named entity recognition, and protein ontology development. It can be utilized by computational and biomedical researchers to explore literature information on proteins and their features or properties.	pir.georgetown.edu/iprolink/
PreBIND	PreBIND is a tool helping researchers locate biomolecular interaction information in the scientific literature. It identifies papers describing interactions using a support vector machine (SVM).	prebind.bind.ca/
PubGene	PubGene is constructed to identify the relationships between genes and proteins, diseases, cell processes, and so on based on their co-occurrences in the abstracts of scientific papers, their sequence homology, and statistical probability of their co-occurrences.	www.pubgene.org/
Chilibot	Chilibot [16] is a search software for the MEDLINE literature database to rapidly identify relationships between genes, proteins, or any keywords that the user might be interested.	www.chilibot.net/
iHOP	Information Hyperlinked over Proteins [17] constructs a gene network by converting the information in MEDLINE into one navigable resource using genes and proteins as hyperlinks between sentences and abstracts.	www.ihop-net.org/UniPub/iHOP/

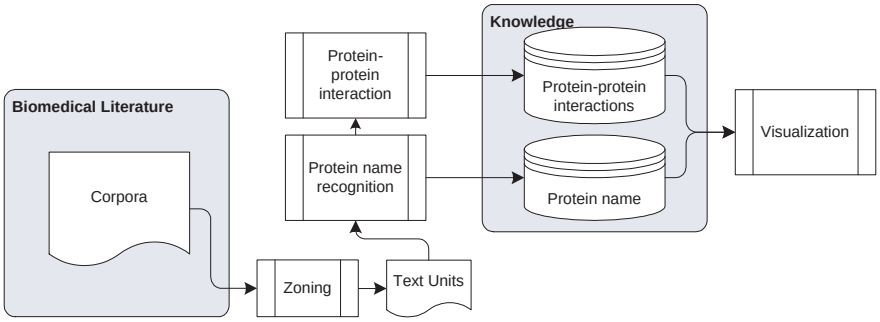

Fig. 1. A general architecture of an information extraction system for protein-protein interactions.

In general, to automatically extract protein-protein interactions, a system needs to consist of three to four major modules [13, 18], which is illustrated in Figure 1.

- *Zoning module.* It splits documents into basic building blocks for later analysis. Typical building blocks are phrases, sentences, and paragraphs. In special cases, higher-level building blocks such as sections or chapters may be chosen. Ding [19] compared the results of employing different text units such as phrases, sentences, and abstracts from MEDLINE to mine interactions between biochemical entities based on co-occurrences. Experimental results showed that abstracts, sentences, and phases all can produce comparative extraction results. However, with respect to effectiveness, sentences are significantly better than phrases and are about the same as abstracts.
- *Protein name recognition module.* Before the extraction of protein-protein interactions, it is crucial to facilitate the identification of protein names, which still remains a challenging problem [20]. Although experimental results of high recall and precision rates have been reported, several obstacles to further development are encountered while tagging protein names for the conjunctive natural of the names [21]. Chen [22] and Leser [23] provided a quantitative overview of the cause of gene-name ambiguity, and suggested what researchers can do to minimize this problem.
- *Protein-protein interaction extraction module.* As the retrieval of protein-protein interactions has attracted much attention in the field of biomedical information extraction, plenty of approaches have been proposed. The solutions range from simple statistical methods relying on co-occurrences of genes or proteins to methods employing a deep syntactical or semantical analysis.
- *Visualization module.* This module is not as crucial as the aforementioned three modules, but it provides a friendly interface for users to delve into the generated knowledge [24]. Moreover, it allows users to interact with the

system for ease of updating the system's knowledge base and eventually improve its performance.

To evaluate the performance of an information extraction system, normally recall and precision values are measured. Suppose a test dataset has T positive information (for example, protein-protein interactions), and an information extraction system can extract I "positive" information. In I, only some information is really positive which we denote as B and the remaining information is negative, however the system falsely extracts as positive which we denote as C. In T, some information is not extracted by the system which we denote as A. The relationships of A, B, and C are illustrated in Figure 2.

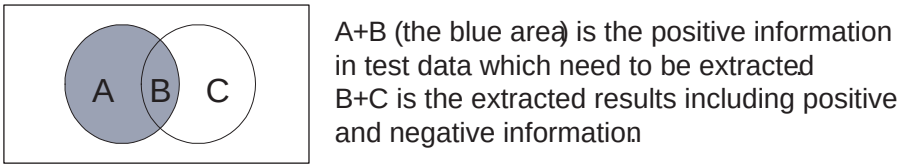

Fig. 2. Venn Diagram of information extraction results.

Based on above definitions, recall and precision can be defined as:

$$\text{Precision} = \frac{\|B\|}{\|B\| + \|C\|} \tag{1}$$

$$\text{Recall} = \frac{\|B\|}{\|A\| + \|B\|} \tag{2}$$

For example, a test dataset has 10 protein-protein interactions T. An information extracting system totally extract 11 protein-protein interactions I. In I, only 6 protein-protein interactions (B) can be found in T, which we consider as true positive (TP). The remaining 5 protein-protein interaction (C) can not be found in T, which we consider as false positive (FP). In T, 4 protein-protein interactions (A) are not extracted by the system, which we consider as false negative (FN). Thus, the recall of the system is $6/(6+4) = 60\%$ and the precision is $6/(6+5) = 54.5\%$.

Obviously, an ideal information extracting system should fulfil $\|A\| \longrightarrow 0, \|C\| \longrightarrow 0$. To reflect these two conditions, F-measure is defined by the harmonic (weighted) average of precision and recall [25] as :

$$\begin{aligned} F_\beta &= \frac{(1+\beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \\ &= \frac{(1+\beta^2)\|B\|}{(1+\beta^2)\|B\| + \beta^2\|A\| + \|C\|} \end{aligned} \tag{3}$$

where β indicates the relative value of precision. For further details of the state of the science in text mining evaluation, please refer to Hersh [26].

In this chapter, we focus on the protein name recognition and the protein-protein interaction extraction module. A brief survey and classification on the developed methodologies is provided. In general, the methods proposed so far rely on the techniques from one or more areas [27–30] including Information Retrieval (IR) [25,31], Machine Learning (ML) [32,33], Natural Language Processing (NLP, also known as computational linguistics) [34–36], Information Extraction (IE) [37–40], Text Mining [41–47], and so on. The surveyed work illustrates the progress of the field and shows the increasing complexity of the proposed methodologies.

The rest of the chapter is organized as follows. Firstly, systems and methods implemented for protein name recognition are discussed in Section 2. Then, the later sections discuss the methods and systems for extracting protein-protein interactions. Section 3 presents a survey of various methods applied in automatical extraction of protein-protein interactions from literature. In succession, challenges are identified and possible solutions are suggested.

2 Protein Name Recognition

As mentioned in section 1, recognizing protein names in the biomedical literature is crucial for the latter processing. An example of a sentence with its protein names in italic is given as follows:

Interleukin-2 (*IL-2*) rapidly activated *Stat5* in fresh PBL, and *Stat3* and *Stat5* in preactivated PBL. [PMID: 7719938]

There are various methods for recognizing protein name. Traditionally, these methods can be divided into four categories namely the dictionary based approaches, rule-based approaches, machine learning approaches, and hybrid approaches.

2.1 Dictionary Based Approaches

In dictionary based approaches, protein names are identified from text by using a provided list of protein names. These names can be identified using substring matching techniques such as exact matching and approximate string matching.

Egorov et al. [48] implemented a protein name identification system, ProtScan, using a carefully constructed dictionary. This dictionary was built based on the LocusLink database and enriched by the GenBank, GoldenPath and HUGO database entries. The system was evaluated on a gold standard, which consists of 1,000 randomly selected MEDLINE abstracts and achieved 88.6% recall and 98% precision. When evaluated on a more general set of biomedical documents other than MEDLINE abstracts, 98.5% recall and 84%

precision were reported. Krauthammer et al. [49] proposed a dictionary-system based on BLAST [50], a tool for DNA and protein sequence comparison. An exhaustive list of gene and protein names was extracted from GenBank and translated into DNA sequences to form a dictionary. Names in the dictionary and input texts were converted into nucleotide sequences and then BLAST was implemented. The system achieved a recall of 78.8% and precision of 71.7%, of which 4.4% of names not included in the dictionary are fully recognized when evaluated on a gold standard review article marked by 2 experts.

Dictionary-based approaches in general can not identify protein names that are not listed in the pre-constructed dictionary. Their performance is highly dependent on the quality of their base dictionaries.

2.2 Rule-Based Approaches

Rule-based approaches identify protein names based on a set of manually defined rules. These rules usually employ surface clues and the syntactic and semantic properties of the gene and protein names.

Fukuda et al. [51] proposed a rule based system, PROPER, for identifying protein names using surface clues on character strings. These clues include capital letters, numerical figures and non-alphabetical letters. Evaluation was conducted based on 30 abstracts from MEDLINE in the SH3 protein domain and a recall of 98.84% and a precision of 94.70% was achieved. The Yapex system, based on hand-written rules was implemented by Franzen et al. [52]. Lexical analysis of single word tokens, syntactic analysis of noun phrases was performed to identify new protein names. 99 abstracts were randomly selected from MEDLINE to form the training corpus and 101 MEDLINE abstracts formed the test corpus. Yapex achieved a recall of 66.4% and a precision of 67.8%. In GPmarkup [53], abbreviations were first mapped to full names using a set of guidelines and protein symbols were mapped to the names by a set of pattern-matching rules. The mappings were performed on 11 million MEDLINE records and the abbreviation-name or symbol-name pairs were stored in a knowledge database. Non-protein abbreviation-name pairs in the database were then filtered out based on a set of heuristic rules. 50 abstracts from MEDLINE were randomly selected to form the test set and it achieved a recall of 73% and a precision of 93%.

Rule-based systems have the advantage that rules are able to be defined and extended when needed. However, the construction of rules has to be done manually and can be very time-consuming.

2.3 Machine Learning Approaches

Machine Learning approaches use various algorithms to automatically identify protein names. There are three commonly used approaches, Naive Bayes (NB), Support Vector Machine (SVM), and Hidden Markov Model (HMM).

Naive Bayes

The NB Classifier is the most commonly used approach to identify protein names. It is a simple probabilistic model based on the Bayes' rule. It assumes that the effect of one feature on a given class is independent from that of another feature.

Nobata et al. [54] developed a system by calculating the similarity between a string and a class. NB was used to estimate the probability of a word occurring in a particular class. 100 abstracts were tagged by a human expert using Genia Ontology. Out of the 100 abstracts, 20 were used for testing and the remaining 80 were used for training. The system achieved an F-measure of 65.8%. Wilbur [55] considered several approaches based on NB. As the NB algorithm assumes that values are independent of each other and each term can be weighted separately based on its distribution in the training set. Documents in the test set are then scored by summing the weights of the terms they contain. The test set consisted of 100 documents and the training set consisted of 3,021 documents. The precision obtained was 71.4%. The staged NB algorithm was also be implemented. NB was first trained on the entire training set, then tested using both the training and test sets. A second training involved the positive examples and the negative ones that were unable to be separated in the first training. The precision achieved for this algorithm was 78.9%.

The main advantages of using the Naive Bayes approach is that it is fast to train and evaluate.

Support Vector Machine

A support vector machine (SVM) is a supervised learning technique for classification and regression. Mika and Rost [56] proposed a system, NLProt, that combines dictionary and rule-based filtering together with SVMs to tag protein names. The system used two dictionaries to perform pre-filtering. The first dictionary is a protein name dictionary with names generated from SWISS-PROT and TrEMBL [57] and the second is a common dictionary containing non-protein names. Input text is then tagged and run on four trained SVMs. When tested on the Yapex corpus, the system achieved an F-measure of 75% compared to the 67.1% on the Yapex system. GAPSCORE [58] identifies protein names based on their syntax, appearance, morphology, context and abbreviations. Features in it were developed on a Yapex independent corpus in order to obtain an accurate evaluation of the performance. A training set of 735 abstracts from MEDLINE was used for training the NB, Maximum Entropy (ME), and SVM classifiers. When evaluated on the Yapex training set, SVM outperformed the other two classifiers with a recall, precision and F-measure of 79.3%, 77%, and 78.1% respectively. When evaluated on the Yapex test set, SVM achieved a recall of 70.3%, a precision of 81.4%, and an F-measure of 75.4%. Hakenberg et al. [59] developed a system to solve the

Name Entity Recognition (NER) problem posed by BioCreAtIve task 1A[3]. In the system, words are separated into tokens and an SVM is used to identify features that describe gene and non-gene names. Then, post-processing is performed by passing the tokens through a POS (Part-of-Speech)-tagger to find complete gene phrases. On a given test corpus of 5,000 previously unseen and untagged sentences, the system attained a recall of 72.8%, precision of 71.4% and an F-measure of 72.1% on the closed-division. The system is later enhanced by performing Recursive Feature Elimination (RFE) where features with the lowest weights are removed until 150 features remain. Post-processing is then done. Recall and precision of 82.8% and 83.4% respectively were achieved with this enhanced system.

SVM is an extremely powerful binary non-linear classifier. However, it is computationally demanding to train and run when the dataset is very large. It is also sensitive to noisy data and prone to overfitting, which in turn leads to generalization failures.

Hidden Markov Model

A Hidden Markov Model (HMM) is a variant of a finite state machine where the modelled system is assumed to be a Markov process with unknown parameters. It consists of a finite set of states, in which each state is associated with a probability value. However, the states in an HMM are not directly visible to an external observer, only the observations are visible.

Collier et al. [61] proposed a model to find the most probable class which a word belongs to. A first-order HMM is used to implement the model. 1,000 MEDLINE abstracts related to molecular biology were selected and marked up by an expert. Out of these 1,000 abstracts, 800 were used for training and 200 were used for testing. The system reported an F-measure of 72.8%. PowerBioNE [62] is a system that is implemented using HMM and an HMM-based name entity recognizer. A pattern-based post processing was also done to extract rules from the training data so as to deal with the cascaded entity name phenomenon. The GENIA Corpus v3.0 which contains 2,000 MEDLINE abstracts of 360 thousand words was used. 200 abstracts were selected as the testing data and the remaining 1800 formed the training data. On twenty-three classes of the GENIA corpus, the system achieved an F-measure of 66.6%. On the "protein" class, the system achieved an F-measure of 75.8%.

In general, machine learning methods are useful when the annotated training set is available, as it requires experts to determine the protein names and this can be very time-consuming. In addition, the training set must also contain enough amounts of data in order to prevent the data sparseness problem.

[3] BioCreAtIvE task 1A [60] focuses on extracting gene names and participants involved were given 10,000 MEDLINE sentences with tagged protein and gene names.

2.4 Hybrid Approaches

Hybrid approaches use a combination of the above mentioned methods in the extraction of protein and gene names.

Tanabe and Wilbur [63] presented a method that uses a combination of rules and machine learning strategies. Rules are automatically generated by the Brill POS Tagger and then augmented with hand-crafted rules. The Brill tagger is trained on a corpus of 7,000 MEDLINE sentences to produce rules for tagging the texts. Next, rule-based post-processing rules are applied to identify potential gene names and then NB learning is used to rank documents based on their likelihood to contain a gene name. A test corpus of 56,469 MEDLINE abstracts was used to identify gene and protein names and results showed that higher performance can be achieved on documents with a higher Bayesian score. PROTEX [64] is a system employing a set of heuristic rules, a probabilistic model and a protein name dictionary to identify protein names. In the approach, heuristic rules reported in [51, 52] were first used to detect protein names, and then a probabilistic model was used to identify complete protein names. Finally, a dictionary compiled from the SWISS-PROT and TrEMBL protein databases were used to detect protein names that were not identified earlier. The Yapex gold standard was used for training and testing respectively. The system was then compared to the system by Franzen et al. [52]. Based on the exact matching evaluation criteria, PROTEX reported a recall, precision, and F-measure of 67.7%, 60.2%, and 63.7% respectively, outperforming Yapex.

A table summarizing the various algorithms and their performance is tabulated in Table 2.

Table 2. Performance of existing protein name recognition methods and the data corpora used.

Category	**Result (%)**		**Corpus**	**Ref**
	Recall	Precision		
Dictionary-based	88.6	98	1000 randomly selected abstracts from MEDLINE	[48]
	78.8	71.7	Gold standard review articles marked by 2 experts	[49]
Heuristic rule-based	98.8	94.7	Test Set: 30 abstracts on the SH3 protein domain from MEDLINE	[51]
	66.4	67.8	Training Set: 99 random abstracts from MEDLINE; Test Set: 101 abstracts from MEDLINE.	[52]
	73	93	Test Set: 50 abstracts from MEDLINE	[65]
Naive Bayes	-	-	Training Set: 80 abstracts; Test Set: 20 abstracts	[54]
	71.4	-	Training Set: 3,021 documents; Test set: 100 documents	[55]
	78.9	-	Training Set: 3,021 documents; Test set: 3,121 documents	[55]
SVM	-	-	Training and Test Set: Yapex Corpus	[56]
	58.5	56.7	Training set: 735 abstracts from MEDLINE; Test Set: Yapex Test Set	[58]
	83.4	82.8	Training set: 10,000 sentences from MEDLINE. Test set: 5000 sentences	[59]
	74.2	75.7	Training Set: 1,600 abstracts from Genia Corpus; Test Set: 400 abstracts from Genia Corpus	[66]
HMM	-	-	Training Set: 800 abstracts from MEDLINE; Test Set: 200 abstracts from MEDLINE	[61]
	-	-	Training Set: 1,800 abstracts from GENIA Corpus; Test Set: 200 abstracts from GENIA Corpus	[62]
Hybrid Systems	67.7	60.2	Training and Test Set: Yapex Corpus	[64]

2.5 Challenges

Despite the availability of many well-known nomenclatures for biomedical entities, there is no community-wide agreement on how a particular gene should be named. One name can stand for a particular gene, may include homologue of this gene in other organisms, may also encompass the protein the gene encodes. As a consequence, recognition of protein names automatically in the biomedical literature is not straightforward. In this section, we list the several open issues.

- *Ambiguous Names.* An ambiguous name denotes different entities. Some protein names are not distinguished from common English words, such as "white", "shaggy" and son on. Some names may denote biomedical entities of different classes. Other names may refer to certain entities before, but refer to another entities now.
- *Multi-word names.* Multi-word names are names consisting of more than one word (or token). For gene and protein names, multi-word names are rather than an exception. Multi-word names are not only harder to find, but in many cases there is no agreement on the exact borders of such names.
- *Synonyms and acronyms.* In synonymy relation, a protein name can be denoted by multiple names. Acronyms are abbreviation of names and are very popular in scientific writing because they allow for shorter texts. However, acronyms are difficult to resolve to their true names because they are often homonyms.
- *Names of newly discovered genes and proteins.* The overwhelming growth rate and the constant discovery of novel genes and proteins make protein name recognition more complex. Methods based on dictionary can not figure out these new names because registering the new names of genes and proteins is time-consuming and occurs much later.

3 Methodologies

This section presents a brief discussion on the existing techniques and methods for extracting protein-protein interactions. In general, current approaches can be divided into three categories:

- *Computational linguistics-based methods.* To discover knowledge from unstructured text, it is natural to employ computational linguistics and philosophy, such as syntactic parsing or semantic parsing to analyze sentence structures. Methods of this category define grammars to describe sentence structures and use parsers to extract syntactic information and internal dependencies within individual sentences. Approaches in this category can be applied to different knowledge domains after being carefully tuned to the specific problems. But, there is still no guarantee that the performance

in the field of biomedicine can achieve comparable performance after tuning. Until recently, methods based on computational linguistics still could not generate satisfactory results.

- *Rule-based methods.* Rule-based approaches define a set of rules for possible textual relationships, called patterns, which encode similar structures in expressing relationships. When combined with statistical methods, scoring schemes depending on the occurrences of patterns to describe the confidence of the relationship are normally used. Similar to computational linguistics methods, rule-based approaches can make use of syntactic information to achieve better performance, although it can also work without prior parsing and tagging of the text.
- *Machine learning and statistical methods.*
  Machine learning refers to the ability of a machine to learn from experience to extract knowledge from data corpora. As opposed to the aforementioned two categories that need laborious effort to define a set of rules or grammars, machine learning techniques are able to extract protein-protein interaction patterns without human intervention.
  Statistical approaches are based on word occurrences in a large text corpus. Significant features or patterns are detected and used to classify the abstracts or sentences containing protein-protein interactions, and characterize the corresponding relations among genes or proteins.

It has to be mentioned that many existing systems in fact adopt a hybrid approach for better performance by combining methods from two or more of the aforementioned categories.

Figure 3 illustrates the process of information extraction on an example sentence by employing the typical methods in the above three categories.

3.1 Computational Linguistics-Based Methods

In general, computational linguistics-based methods employ linguistic technology to grasp syntactic structures or semantic meanings from sentences.

Techniques for analyzing a sentence and determining its structure in computational linguistics are called parsing techniques. Parsing the corpus firstly to obtain the morphological and syntactic information for each sentence is extremely important, and probably only after that, it would be possible to fulfill sophisticated tasks such as identifying the relationship between proteins and gene products in a fully automatic way. However, it is well known that parsing unrestricted texts, such as those in the biomedical domain, is extremely difficult.

The methods in this category can be further divided into two types, based on the complexity of the linguistics methods, as shallow (or partial) parsing or deep (or full) parsing. Shallow parsing techniques aim to recover syntactic information efficiently and reliably from unrestricted text, by sacrificing completeness and depth of analysis, while deep parsing techniques analyze

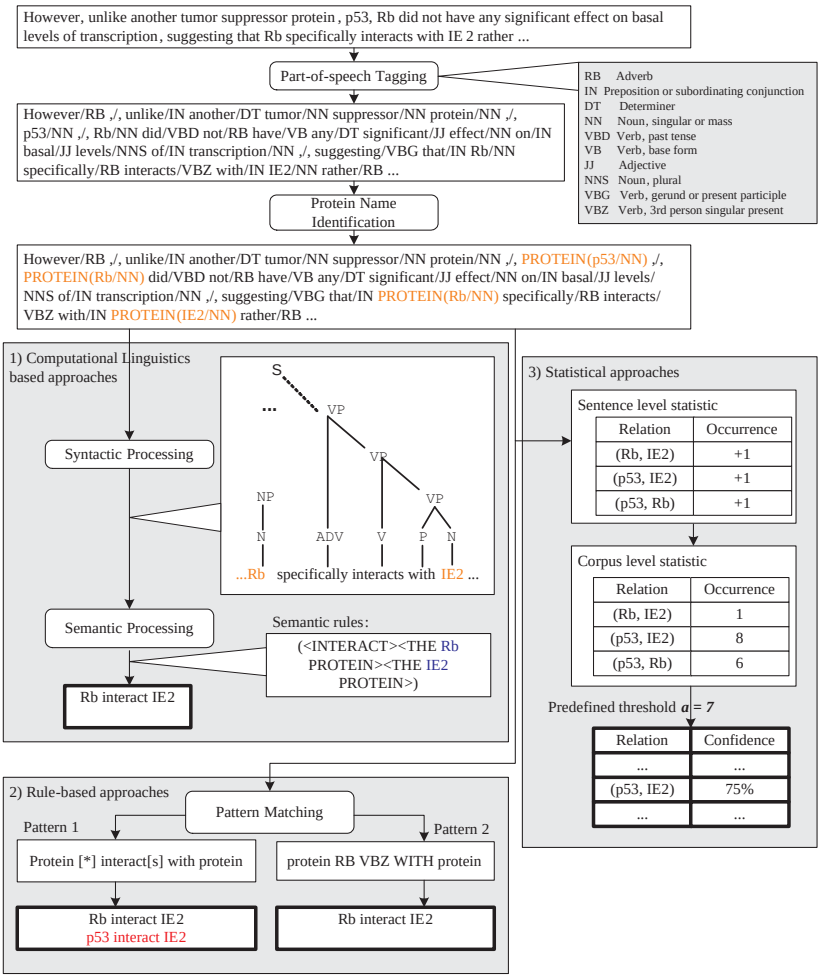

Fig. 3. General dataflow of information extraction system employing different methodologies.

the entire sentence structure, which normally achieve better performance but with increased computational complexity.

Shallow Parsing Approaches

Shallow parsers [67–71] perform partial decomposition of a sentence structure. They first break sentences into none-overlapping chunks, then extract local dependencies among chunks without reconstructing the structure of an entire sentence. Sekimizu used shallow parser, EngCG, to generate three kinds of tags, such as syntactic, morphological, and boundary tags [67]. Based on the

tagging results, subjects and objects were recognized for the most frequently used verbs in a collection of abstracts which were believed to express the interactions between proteins, genes. Thomas [69] modified a preexisting parser based on the cascaded finite state automata (FSA). Predefined templates were then filled with information about protein interactions based on the parsing results for three verbs: *interact with*, *associate with*, *bind to*. Pustejovsky [70] targeted "inhibit" relations in the text and also built an FSA to recognize these relations. Leroy [71] used a shallow parser to automatically capture the relationships between noun phrases in free text. The shallow parser is based on four FSAs to structure the relations between individual entities and model generic relations not limited to specific words. By elaborate design, the parser can also recognize coordinating conjunctions and capture negation in text, a feature usually ignored by others. The precision and recall rates reported for shallow parsing approaches are estimated at 50-80% and 30-70%, respectively.

Deep Parsing Approaches

Systems based on deep parsing deal with the structure of an entire sentence and therefore are potentially more accurate. Variations of the deep parsing-based approach have been proposed [10, 72–81]. Based on the way of constructing grammars, deep parsing-based approaches can be divided into two types: rationalist methods and empiricist methods. Rational methods define grammars by manual efforts, while empiricist methods automatically generate the grammar by some observations.

Rationalist Methods

Yakushiji [75] used a general full parser with grammars for biomedical domain to extract interaction events by filling sentences into slots of semantic frames. Information extraction itself is done using pattern matching on the canonical structure. Park [74] proposed bidirectional incremental parsing with combinatory categorial grammar (CCG). This method first localized target verbs, and then scanned the left and right neighborhood of the verb respectively. The lexical and grammatical rules of CCG are more complicated than those of a general context-free grammar (CFG). The recall and precision rate of the system were reported to be 48% and 80%. Temkin [78] introduced a lexical analyzer and a CFG to extract protein, gene and small molecule interactions with a recall rate of 63.9% and precision rate of 70.2%. Ding [79] investigated link grammar parsing for extracting biochemical interactions. It can handle many syntactic structures and is computationally relatively efficient. A better overall performance was achieved compared to those biomedical term co-occurrence based methods. Ahmed [10] split complex sentences into simple clausal structures made up of syntactic roles based on a link grammar. Complete interactions were then extracted by analyzing the matching contents of syntactic roles and their linguistically significant combinations. In

GENIES [76], a parser and a semantic grammar consisting of a large set of nested semantic patterns (incorporating some syntactic knowledge) are used. Unlike other systems, GENIES is capable of extracting a wide variety of different relations between biological molecules as well as nested chains of relations. However, the downside of the semantic grammar-based systems such as GENIES is that they may require complete redesign of the grammar in order to be tuned for used in different domain.

Empiricist Methods

Many empiricist methods [77, 80] have been proposed to automatically generate the language model to mimic the features of unstructured sentences. For example, Seymore [72] used Hidden Markov Model (HMM) for extracting important fields from the headers of computer science research papers. Following the trend, Souyma [73] applied HMM to the biomedical domain to describe the structure of sentences. More recently, Skounakis [82] proposed an approach that is based on hierarchical HMMs to represent the grammatical structure of the sentences being processed. Firstly, shallow parser to construct a multi-level representation of each sentence being processed was used. Then hierarchical HMMs to capture the regularities of the parses for both positive and negative sentences were trained. In [83], a broad-coverage probabilistic dependency parser was used to identify sentence level syntactic relations between the heads of the chunks. The parser used a hand-written grammar combined with a statistical language model that calculates lexicalized attachment probabilities.

3.2 Rule-Based Approaches

In rule-based approaches [6, 7, 9, 12, 84–92], a set of rules need to be defined which may be expressed in forms of regular expressions over words or POS tags. Based on the rules, relations between entities that are relevant to tasks such as proteins, can be recognized.

Ng [84] defined five rules based on the word form, such as <A> ... <fn> ... <B> in which the symbols A, B refer to protein names while the symbol fn refers to the verb which describes the interaction relationship. Obviously, such rules are too simple to produce satisfactory results. Ono [87] manually defined a set of rules based on syntactic features to preprocess complex sentences, with negation structures considered as well. It achieves good performance with a recall rate of 85% and precision rate of 84% for *Saccharomyces cerevisiae (yeast)* and *Escherichia coli*. Blaschke [7] induced a probability score to each predefined rule depending on its reliability and used it as a clue to score the interaction events. Sentence negations and the distance between two protein names were also considered. In [89], gene-gene interactions were extracted by scenarios of patterns which were constructed manually. For example, "gene product acts as a modifier of gene" is a scenario of the predicate act, which can

cover a sentence such as: "Egl protein acts as a repressor of BicD". Egl and BicD can be extracted as an argument of an event for the predicate acts. Shatkay and Leroy [88] employed preposition-based parsing to generate templates. It achieved a template precision of 70% when processing literature abstracts.

Using predefined rules can generate nice results. It is however not feasible in practical applications as it requires heavy manual processing to define patterns when shifting to another domain.

Huang [90] tried to automatically construct the protein-protein interaction patterns. At first, part-of-speech tagging was employed. Then dynamic programming to automatically extract similar patterns from sentences based on POS tags was used. Based on the automatically constructed patterns, protein-protein interactions can be identified. Their results gave precision of 80.5% and recall of 80.0%. Phuong [93] used some sample sentences, which were parsed by a link grammar parser, to learn extraction rules automatically. By incorporating heuristic rules based on morphological clues and domain specific knowledge, the method can remove the interactions that are not between proteins.

Rule-based approaches have been found to be overall limiting in the set of interactions that can be extracted by the extent of the recognition rules that were implemented, and also by the complexity of sentences being processed. Specifically, complicated cases such as interaction descriptions that span several sentences of text are often missed by these approaches. The shortcoming of such approaches is their inability to correctly process anything other than short, straightforward statements, which are quite rare in information-saturated biomedical literature. They also ignore many important aspects of sentence construction such as mood, modality, and sometimes negation, which can significantly alter or even reverse the meaning of the sentence.

3.3 Machine-Learning and Statistical Approaches

Many machine-learning (ML) methods have been proposed ranging from simple methods such as deducing relationship between two terms based on their co-occurrences to complicated methods which employ NLP technologies. Approaches combing machine learning and NLP have been discussed in section 3.1. Here we focus on the methods without employing NLP techniques.

A variety of machine-learning and statistical techniques based on the discovery of co-occurrence of protein names have been applied for protein-protein information extraction [8, 94–106]. They can be further divided into different types based on the mining units, such as abstracts, sentences and so on.

Approaches proposed in Miguel and Marcottle [94, 100] aim to extract protein-protein interactions from a set of abstracts. Miguel [94] used a group of relevant documents against a set of random documents to extract domain specific information such as gene functions and interactions. Marcottle [100] was only interested in retrieving a large number of documents that probably contained information about protein-protein interactions.

The first machine-learning sentence-based information extraction system in molecular biology was described in Craven and Kumlien [96]. They developed a Bayesian classifier which, given a sentence containing mentions of two items of interest, returns a probability that the sentence asserts some specific relations between them. Later systems have applied other technologies, including hidden Markov models and support vector machines, to identify sentences describing protein-protein interactions.

Other approaches [8, 97–99] focus on a pair of proteins and detect the relations between them using probability scores. Stapley [97] used fixed lists of gene names and detected relations between these genes by means of co-occurrences in MEDLINE abstracts. A matrix that contains distance dissimilarity measurement of every pair of genes based on their joint and individual occurrence statistics was constructed based on a user-defined threshold. Stephens [98] furthered the method to discover relationships using more complicated computation on co-occurrences. Jenssen [99] used a similar approach to find relations between human gene clusters obtained from DNA array experiments. Donaldson [8] constructed PreBIND and Textomy - an information extraction system that uses support vector machines to evaluate the importance of protein-protein interactions.

Simple statistical methods such as those based on protein co-occurrence information can not precisely describe the relations between proteins and therefore tend to generate high false negative error rate. On the contrary, complex statistical models need a large amount of training data in order to reliably estimate model parameters, which is usually difficult to obtain in practical applications. To strike the balance, we applied the hidden vector state model (HVS) which was previously used in spoken language understanding to extract protein interactions [107]. Unlike other statistical parsers which need fully-annotated treebank data for training, the HVS model explores the embedded sentence structures using only lightly annotated corpus. The details of how this is done can be found in [108].

3.4 Discussion

The performance of the existing protein-protein interaction extraction methods along with the data corpora they used are listed in Table 3.

As in the area of extracting information about protein-protein interactions, competitive evaluations have played important roles in pushing the field of IE and NLP. Several evaluations have been held in recent years. Procreative challenge (Critical Assessment of Information Extraction in Biology) [109] began in 2004 and provided two common evaluation tasks to assess the state of the art methods for text mining applied to biological problems. The first task dealt with extraction of gene or protein names from text, and their mappings into standardized gene identifiers for three model organism databases (fly, mouse, yeast). The second task [110] addressed issues of functional annotation, requiring systems to identify specific text passages that supported Gene

Table 3. Performance of existing protein-protein interaction extraction methods and the data corpora used.

Category	Result (%)		Corpus	Ref
	Recall	Precision		
Shallow Parsing	-	73	34343 sentences from abstracts retrieved from MEDLINE using keywords "leucine zipper", "zinc finger", "helix loop helix motif"	[67]
	29	69	2565 unseen abstracts extracted from MEDLINE with the keywords molecular, interaction and protein for year 1,998 (560k words)	[69]
	57	90	Training set consists of 500 abstracts from MEDLINE. Evaluation set consists of 56 abstracts collected using search strings "protein" and "inhibit"	[70]
	62	89	26 abstracts	[71]
Deep Parsing	48	80	492 sentences out of 250,000 abstracts on cytosine in MEDLINE	[74]
	63.9	70.2	The test corpus consists of 100 randomly selected scientific abstracts from MEDLINE.	[78]
	-	96	Articles from cell containing 7,790 words revealing 51 binary relations	[76]
	21	91	3.4 million sentences from approximately 3.5 million MEDLINE abstracts dated after 1,988 containing at least one notation of a human protein	[80]
	26.94	65.66	229 abstracts from MEDLINE correspond to 389 interactions from the DIP database.	[10]
Rule Based	47	70	474 sentences from 50 abstracts retrieved using "E2F1"	[88]
	86.8 Yeast 82.5 Escherichia	94.3 Yeast 93.5 Escherichia	834 and 752 sentences containing at least two protein names and one relation keyword for yeast and E.coli obtained by a MEDLINE search using the following keywords, "protein binding" as a MESH term and "yeast", "E coli", "protein", and "interaction"	[87]
	39.7	44.9	Five different sets of abstracts were used: 1. 1,435 MEDLINE abstracts directly referenced from each of the Drosophila Swiss-prot entries. 2. 4,109 MEDLINE abstracts referenced directly from Fly Base. 3. 111,747 abstracts retrieved by extending the set (2) with the Neighbors utility. 4. 518 MEDLINE abstracts containing any of the protein names (related with cell cycle control) and Drosophila in the MESH list of terms. 5. 6,278 MEDLINE abstracts by expanding set (4) using Neighbors to identify all related abstracts.	[7, 85]
	60	87	3,343 abstracts were obtained by querying MEDLINE with the following keywords: "Saccharomyces cerevisiae", "protein", and "interaction". The abstracts were filtered and 550 sentences were retained containing at least one of four keywords "interact", "bind", "associate", "complex" or one of their inflections.	[93]
	80.0	80.5	The top 50 biomedical papers were retrieved from the Internet by querying using the keyword "protein-protein interaction". Full texts were segmented into 65,536 sentences and the sentences with fewer than two protein names were discarded. The final corpus consists of about 1,200 sentences.	[90]

Ontology annotations for specific proteins, given full text articles. Genic Interaction Extraction Challenge [111] was associated with Learning Language in Logic Workshop (LLL05). The challenge focuses on information extraction of gene interactions in *Bacillus subtilin*, a model bacterium. It was reported that the best F-measure achieved with the balanced recall and precision is around 50%.

As annotated corpora are important to the development as well as the evaluation of protein-protein extraction systems, some online available annotated corpora are listed in Table 4.

4 Challenges and Possible Solutions

The continuing growth and diversification of the scientific literature, a prime resource for accessing worldwide scientific knowledge, will require tremendous

Table 4. Online annotated corpora for the extraction of protein-protein interactions.

Corpus Name	*Description*	*URL*
GENA	GENA corpus version 3.0 consists of 2,000 MEDLINE abstracts with more than 400,000 words and almost 100,000 annotations for biological terms.	www-tsujii.is.s.u-tokyo.ac.jp/GENIA/
Apex	It consists of two collections, training collection consisting of 99 abstracts with 1,745 protein names, test collection consisting of 101 abstracts with 1,966 protein names. The protein names in all the abstracts were annotated manually.	www.sics.se/humle /projects/prothalt/
Penninite	The corpus consists of 2,258 MEDLINE abstracts in two domains: 1) the molecular genetics of oncology (1,158 abstracts); 2) the inhibition of enzymes of the CYP450 class (1,100 abstracts).	bioie.ldc.upenn.edu/
LLL05 challenge Corpus	There are 80 sentences in the training set, including 106 examples of genic interactions without coreferences and 165 examples of interactions with coreferences.	genome.jouy.inra.fr /texte/LLLchallenge/

systematic and automated efforts to utilize the underlying information. In the near future, tools for knowledge discovery will play a pivotal role in systems biology. The increasing fervor on the field of biomedical information extraction gives the evidence. IE in biomedicine has been studied for approximately ten years. Over these years, IE systems in biomedicine have grown from simple rule-based pattern matcher to sophisticated, hybrid parser employing computational linguistics technology. But, until now, there are still several severe obstacles to overcome.

Firstly, biomedical IE methods generate poorer results compared with other domains such as newswire. In general, biomedical IE methods are scored with F-measure, with the best methods scoring about 0.85 without considering the limitation of test corpus, which is still far from users' satisfaction. The main reason is that information from ontologies[4] or terminologies is not well used. Until recently, most biomedical IE systems do not make use of information from ontologies or terminologies. Hence, ontologies together with terminological lexicons are prerequisites for advanced biomedical IE. Since different ontologies are employed in different systems currently, unification seems necessary and impendent. Also, biomedical text needs to be semantically annotated and actively linked to ontologies.

[4] Ontologies, structured lists of terms, are often used by NLP technologies to establish the semantic function of a word in a document. The simplest form of ontology is a lexicon or a list of terms that belong to a particular class. A lexicon usually consists of specialized terms and (optionally) their definitions. Another form of ontology is a thesaurus, a collection of terms and their synonyms which are of immense utility for NLP. A popular ontology in biomedicine is Gene Ontology (GO) [112, 113].

Secondly, relations between biological entities, such as proteins or genes are conditional and may change when the same entities are considered in a different functional context. As a consequence, every relation between entities should be linked with the functional context in which the relation was observed. Moreover, without considering the observed context, it is meaningless and impossible to make general statements whether a relation detected by literature mining is a "yes" or a "no" relation. Obviously, to overcome this obstacle, in-depth analysis based on more elaborately constructing grammars or rules in sentence or phrase level is requisite. Hopefully, it will result in the increase of performance.

Thirdly, it seems to be crucial to the success of biomedical IE to bridge the gap between biologists and computational scientists. Currently, this field is dominated by researchers with computational background; however, the biomedical knowledge is only possessed by biologists. That is crucial for defining standards for evaluation; for identification of specific requirements, potential applications and integrated information system for querying, visualization and analysis of data on a large scale; for experimental verification to facilitate the understanding of biological interactions. Hence, to attract more biologists into the field, it is important to design simple and friendly user interfaces that make the tools accessible to non-specialists.

Fourthly, the knowledge extracted from the literature may contradict itself under different environment, conditions, or because of author's errors, experimental errors or other issues. Although the contradictory knowledge may occupy minor part of the whole interaction network, it is worth more attention. To handle this challenge, one way is to categorize the corpora and define the confidence value for each category. For contradictory knowledge, the decision can be made based on these confidence values. The solution can also be applied to handling different parts of an article, such as the abstract, introduction, references and so on, which obviously are of different confidences.

Fifthly, some problems exist not only in the field of biomedical IE, but also in the field of NLP. Two of them are: (1) Dealing with negative sentences, which constitutes a well-known problem in language understanding [114]. (2) Resolving coreferences, the recognition of implicit information in a number of sentences may contain key information, e.g. protein names, that later are used implicitly in other sentences. Results in LLL challenge 05 show that F-measure can only achieve 25% when considering coreferences.

Finally, the development of gold standard for evaluation systems is still under way, far from maturity, which requires more concerted efforts. The experience in the newswire domain shows that the construction of evaluation benchmarks in the face of common challenges contribute greatly to the rapid development of IE. Thus it is crucial to attach importance to evaluate systems development in biomedicine. Also, efforts will be required to focus on linking the knowledge in the databases with text sources available. It is believed that in the future, biomedical IE might provide new approaches for relation dis-

covery that exploit efficiently indirect relationships derived from bibliographic analysis of entities contained in biological databases.

References

1. Pubmed-overview. http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html.
2. I. Xenarios, L. Salwinski, XJ. Duan, P. Higney, SM. Kim, and D. Eisenberg. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 30(1):303–5, 2002.
3. GD. Bader, D. Betel, and CW. Hogue. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Research*, 31(1):248–250, 2003.
4. H. Hermjakob, L. Montecchi-Palazzi, and C. Lewington. IntAct: an open source molecular interaction database. *Nucleic Acids Research.*, 1(32(Database issue)):452–5, 2004.
5. C. von Mering, LJ. Jensen, B. Snel, SD. Hooper, and M. Krupp. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33(Database issue):433–7, 2005.
6. L. Wong. PIES, a protein interaction extraction system. In *Proceedings of the Pacific Symposium on Biocomputing.*, pages 520–531, Hawaii, U.S.A, 2001.
7. Christian Blaschke and Alfonso Valencia. The Frame-Based Module of the SUISEKI Information Extraction system. *IEEE Intelligent Systems*, 17(2):14–20, 2002.
8. I. Donaldson, J. Martin, B. de Bruijn, and C. Wolting. PreBIND and Textomy–mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, 4(11), 2003.
9. Jung-Hsien Chiang, Hsu-Chun Yu, and Huai-Jen Hsu. GIS: a biomedical text-mining system for gene information discovery. *Bioinformatics*, 20(1):120–121, 2004.
10. Syed Toufeeq Ahmed, Deepthi Chidambaram, Hasan Davulcu, and Chitta Baral. IntEx: A Syntactic Role Driven Protein-Protein Interaction Extractor for Bio-Medical Text. In *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Database 2005*, pages 54–61, 2005.
11. TC Rindflesch, L Tanabe, JN. Weinstein, and L. Hunter. EDGAR: extraction of drugs, genes and relations from the biomedical literature. In *Proceedings of Pacific Symposium Biocomputing*, pages 517–28, 2000.
12. David P. A. Corney, Bernard F. Buxton, William B. Langdon, and David T. Jones. BioRAT: extracting biological information from full-length papers. *Bioinformatics*, 20(17):3206–3213, 2004.
13. Rzhetsky A, Iossifov I, Koike T, Krauthammer M, Kra P, Morris M, Yu H, Dubou PA, Weng W, Wilbur WJ, Hatzivassiloglou V, and Friedman C. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatic*, 37(1):43–53, February 2004.
14. Suraj Peri, J. Daniel Navarro, and Ramars Amanchy. Development of Human Protein Reference Database as an Initial Platform for Approaching Systems Biology in Humans. *Genome Research*, 13:2363–2371, 2003.

15. A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni. MINT: a Molecular INTeraction database. *FEBS letters*, 513(1):135–40, 2002.
16. Hao Chen and Burt M Sharp. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*, 8(5):147, 2004.
17. Robert Hoffmann and Alfonso Valencia. A gene network for navigating the literature. *Nature Genetics*, 36:664, 2004.
18. Brigitte Mathiak and Silke Eckstein. Five Steps to Text Mining in Biomedical Literature. In *Data Mining and Text Mining for Bioinformatics European Workshop*, 2004.
19. J. Ding, D. Berleant, D. Nettleton, and E. Wurtele. Mining MEDLINE: Abstracts, sentences, or phrases. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 326–337, Hawaii, U.S.A, 2002.
20. Michael Krauthammer and Goran Nenadic. Term identification in the biomedical literature. *Journal of biomedical informatics*, 37(6):512–26, 2004.
21. H Pearson. Biology's name game. *Nature*, 411(6838):631–632, 2001.
22. Lifeng Chen, Hongfang Liu, and Carol Friedman. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, 21(2):248–256, 2005.
23. Ulf Leser and Jörg Hakenberg. What makes a gene name? Named entity recognition in the biomedical literature. *Briefings in Bioinformatics*, 6(4):257–269, 2005.
24. Harold J Drabkin, Christopher Hollenbeck, David P Hill, and Judith A Blake. Ontological visualization of protein-protein interactions. *BMC Bioinformatics. 2005*, 6(29), 2005.
25. C.J. van Rijsbergen. *Information Retrieval*. 1999.
26. William Hersh. Evaluation of biomedical text-mining systems: Lessons learned from information retrieval. *Briefings in Bioinformatics*, 6(4):344–356, 2005.
27. Andrade MA and Bork P. Automated extraction of information in molecular biology. *FEBS Lett.*, 476(1-2):12–7, June 2000.
28. Lynette Hirschman, Jong C. Park, Junichi Tsujii, Limsoon Wong, and Cathy H. Wu. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12):1553–1561, 2002.
29. H. Shatkay and R. Feldman. Mining the biomedical literature in the genomic era: an overview. *Journal of Computational Biology*, 10(6):821–855, 2003.
30. Lars Juhl Jensen, Jasmin Saric, and Peer Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics*, 7:119–129, February 2006.
31. William R. Hersh. *Information Retrieval: A Health and Biomedical Perspective*. 2003.
32. Jeffrey T. Chang. *Using Machine Learning to Extract Drug and Gene Relationships from Text*. PhD thesis, Stanford University, September 2003.
33. Lluís Màrquez. Machine learning and natural language processing. Technical Report LSI-00-45-R, Departament de Llenguatges i Sistemes Informàtics (LSI), Universitat Politecnica de Catalunya (UPC), Barcelona, Spain, 2000.
34. K. Bretonnel Cohen and Lawrence Hunter. *Natural language processing and systems biology*. Series: Computational Biology, Vol. 5 Dubitzky, Werner; Azuaje, Francisco (Eds.), 2004.
35. Yandell MD and Majoros WH. Genomics and natural language processing. *Nature Reviews Genetics*, 3(8):601–10, August 2002.

36. Lawrence Hunter and K. Bretonnel Cohen. Biomedical Language Processing: What's Beyond PubMed? *Molecular Cell*, 21(5):589–594, 2006.
37. Claire Cardie. Empirical Methods in Information Extraction. *AI Magazine*, 18(4):65–80, 1997.
38. C. Blaschke, R. Hoffmann, J. Oliveros, and A. Valencia. Extracting information automatically from biological literature. *Comparative and Functional Genomics*, 2(5):2:310–313, 2001.
39. H. Cunningham. Information Extraction, Automatic. *Encyclopedia of Language and Linguistics, 2nd Edition*, 2005.
40. Andre Skusa, Alexander Rüegg, and Jacob Köhler. Extraction of biological interaction networks from scientific literature. *Briefings in Bioinformatics*, 6(3):263–276, 2005.
41. Text mining in the life sciences. Technical report, 2004.
42. B. De Bruijn and J. Martin. Literature Mining in Molecular Biology. In *Proceedings of the EFMI Workshop on Natural Language Processing in Biomedical Application*, pages 1–5, 2002.
43. M. Krallinger, RA. Erhardt, and A. Valencia. Text-mining approaches in molecular biology and biomedicine. *Drug Discovery Today*, 10(6):439–45, 2005.
44. Irena Spasic, Sophia Ananiadou, John McNaught, and Anand Kumar. Text mining and ontologies in biomedicine: Making sense of raw text. *Briefings in Bioinformatics*, 6(3):239–251, 2005.
45. Aaron M. Cohen and William R. Hersh. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1):57–71, 2005.
46. Sophia Ananiadou and John Mcnaught. *Text mining for biology and biomedicine*. 2006.
47. Hagit Shatkay and M. Craven. *Biomedical text mining*. 2006.
48. S. Egorov, A. Yuryev, and N. Daraselia. A simple and practical dictionary-based approach for identification of proteins in medline abstracts. *Journal of the American Medical Informatics Association*, 11:174–178, 2004.
49. Krauthammer M, Rzhetsky A, Morozov P, and Friedman C. Using blast for identifying gene and protein names in journal articles. *Gene*, 259(1):245–252, 2000.
50. Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. Basic local alignment search tool. *Journal of molecular biology*, 215:403–410, 1990.
51. K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi. Toward information extraction: Identifying protein names from biological papers. In *Proceeding of the Pacific Symposium on Biocomputing*, pages 707–718, Hawaii, USA, 1998.
52. Franzen K., Eriksson G., Olsson F., Asker L., Liden P., and Coster J. Protein names and how to find them. *International Journal of Medical Informatics,*, 67(1):49–61, 2002.
53. Hong Yu, Vasileios Hatzivassiloglou, Andrey Rzhetsky, and W. John Wilburc. Automatically identifying gene/protein terms in medline abstracts. *Journal of Biomedical Informatics*, 35:322–330, 2002.
54. C. Nobata, N. Collier, and J. Tsujii. Automatic term identification and classification in biology texts. In *the 5th Natural Language Processing Pacific Rim Symposium*, Beijing, China, 1999.
55. Wilbur WJ. Boosting naive Bayesian learning on a large subset of MEDLINE. In *Proceedings of AMIA Symposium*, pages 918–922, Beijing, China, 2000.
56. S. Mika and B. Rost. Nlprot: extracting protein names and sequences from papers. *Nucleic Acids Research*, 32:634–637, 2004.

57. B. Boeckmann, A. Bairoch, R. Apweiler, M. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. ODonovan, and I. Phan. The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, 31(1):365–370, 2003.
58. J.T. Chang, H. Schutze, and R.B. Altman. Gapscore: finding gene and protein names one word at a time. *Bioinformatics*, 20(2):216–225, 2004.
59. Jorg Hakenberg, Steffen Bickel, Conrad Plake, Ulf Brefeld, Hagen Zahn, Lukas Faulstich, Ulf Leser, and Tobias Scheffer. Systematic feature evaluation for gene name recognition. *BMC Bioinformatics*, 6:S9, 2005.
60. Alexander Yeh, Alexander Morgan, Marc Colosimo, and Lynette Hirschman. Biocreative task 1a: gene mention finding evaluation. *BMC Bioinformatics*, 6:S2, 2005.
61. Nigel Collier, Chikashi Nobata, and Jun ichi Tsujii. Extracting the names of genes and gene products with a hidden Markov model. In *Proceedings of the 18th conference on Computational linguistics*, pages 201–207, Saarbrucken, Germany, 2000.
62. Zhou G, Zhang J, Su J, Shen D, and Tan C. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20(7):1178–1190, 2004.
63. L. Tanabe and W.J. Wilbur. Tagging gene and protein names in biomedical text. *Bioinformatics*, 18(8):1124–1132, 2002.
64. K. Seki and J. Mostafa. A hybrid approach to protein name identification in biomedical texts. *Information Processing and Management*, 41:723–743, 2005.
65. Hong Yu, Vasileios Hatzivassiloglou, Andrey Rzhetsky, and W. John Wilbur. Automatically identifying gene/protein terms in MEDLINE abstracts. *Biomedical Informatics*, 35(5/6):322–330, 2002.
66. Kaoru Yamamoto, Taku Kudo, Akihiko Konagaya, and Yuji Matsumoto. Protein name tagging for biomedical annotation in text. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine*, pages 65 – 72, Sapporo, Japan, 2003.
67. T. Sekimizu, H. Park, and J. Tsujii. Identifying the interaction between genes and gene products based on frequently seen verbs in MEDLINE abstracts. In *Workshop on Genome Informatics*, volume 9, pages 62–71, 1998.
68. TC. Rindflesch, L Hunter, and AR. Aronson. Mining molecular binding terminology from biomedical text. In *Proceedings of AMIA Symposium*, pages 127–31, 1999.
69. J. Thomas, D. Milward, C. Ouzounis, and S. Pulman. Automatic extraction of protein interactions from scientific abstracts. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 541–552, Hawaii, U.S.A, 2000.
70. J. Pustejovsky, J. Castano, J. Zhang, M. Kotecki, and B. Cochran. Robust Relational Parsing Over Biomedical Literature: Extracting Inhibit Relations. In *Proceedings of the Pacific Symposium on Biocomputing.*, pages 362–373, Hawaii, U.S.A, 2002.
71. Gondy Leroy, Hsinchun Chen, and Jesse D. Martinez. A Shallow Parser Based on Closed-Class Words to Capture Relations in Biomedical Text. *Journal of Biomedical Informatics*, 36(3):145–158, 2003.
72. Kristie Seymore, Andrew McCallum, and Roni Rosenfeld. Learning Hidden Markov Model Structure for Information Extraction. In *AAAI 99 Workshop on Machine Learning for Information Extraction*, 1999.

73. Souyma Ray and Mark Craven. Representing Sentence Structure in Hidden Markov Models for Information Extraction. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-2001)*, pages 1273–9, 2001.
74. J. Park, H. Kim, and J. Kim. Bidirectional incremental parsing for automatic pathway identification with combinatory categorical grammar. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 6, pages 396–407, Hawaii, U.S.A, 2001.
75. A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii. Event extraction from biomedical papers using a full parser. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 6, pages 408–419, Hawaii, U.S.A, 2001.
76. Carol Friedman, Hong Yu Pauline Kra, Michael Krauthammer, and Andrey Rzhetsky. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17:S74–S82, 2001.
77. S. Novichkova, S. Egorov, and N. Daraselia. MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics*, 19(13):1699–1706, 2003.
78. Joshua M. Temkin and Mark R. Gilder. Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics*, 19(16):2046–2053, 2003.
79. Jing Ding, Daniel Berleant, Jun Xu, and Andy W. Fulmer. Extracting Biochemical Interactions from MEDLINE Using a Link Grammar Parser. In *15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'03)*, 2003.
80. Nikolai Daraselia, Anton Yuryev, Sergei Egorov, Svetalana Novichkova, Alexander Nikitin, and llya Mazo. Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics*, 20(5):604–611, 2004.
81. Shengyang Tan and Chee Keong Kwoh. Cytokine Information System and Pathway Visualization. In *International Joint Conference of InCoB, AASBi and KSBI (BIOINFO2005)*, 2005.
82. Marios Skounakis, Mark Craven, and Souyma Ray. Hierarchical Hidden Markov Models for Information Extraction. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, 2003.
83. Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, and James Dowdall. Mining relations in the GENIA corpus. In *Second European Workshop on Data Mining and Text Mining for Bioinformatics*, 2004.
84. See-Kiong Ng and Marie Wong. Toward Routine Automatic Pathway Discovery from On-line Scientific Text Abstracts. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, 1999.
85. Christian Blaschke, Miguel A. Andrade, Christos Ouzounis, and Alfonso Valencia. Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 60–67. AAAI Press, 1999.
86. Valencia A. Blaschke C. The potential use of SUISEKI as a protein interaction discovery tool. In *Workshop on Genome Informatics*, pages 12:123–34, 2001.
87. Toshihide Ono, Haretsugu Hishigaki, Akira Tanigam, and Toshihisa Takagi. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, 17(2):155–161, 2001.

88. G. Leroy and H. Chen. Filling preposition-based templates to capture information from medical abstracts. In *Pacific Symposium Biocomputing*, pages 350–361, 2002.
89. Denys Proux, Franois Rechenmann, and Laurent Julliard. A Pragmatic Information Extraction Strategy for Gathering Data on Genetic Interactions. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 279–285. AAAI Press, 2000.
90. Minlie Huang, Xiaoyan Zhu, and Yu Hao. Discovering patterns to extract protein-protein interactions from full text. *Bioinformatics*, 20(18):3604–3612, 2004.
91. Hong-Woo Chun, Young-Sook Hwang, and Hae-Chang Rim. Unsupervised Event Extraction from Biomedical Literature using Co-occurrence Information and Basic Patterns. *In the Lecture Notes in Artificial Intelligence*, pages 777–786, 2005.
92. Yu Hao, Xiaoyan Zhu, Minlie Huang, and Ming Li. Discovering patterns to extract protein-protein interactions from the literature: Part II. *Bioinformatics*, 21(15):3294–3300, 2005.
93. Tu Minh Phuong, Doheon Lee, and Kwang Hyung Lee. Learning rules to extract protein interactions from biomedical text. In *The Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD-03)*, 2003.
94. Miguel A. Andrade and Alfonso Valencia. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families . *Bioinformatic*, 14(7):600–607, 1998.
95. Mark Craven. Learning to extract relations from medline. In *Proceedings of the AAAI 1999 Workshop on Machine Learning for Information Extraction*, pages 25–30, 1999.
96. Craven Mark and Kumlien Johan. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*, pages 77–86, Heidelberg, Germany, 1999.
97. B. Stapley and G. Benoit. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 529–540, Hawaii, U.S.A, 2000.
98. M. Stephens, M. Palakal, S. Mukhopadhyay, R. Raje, and J. Mostafa. Detecting gene relations from MEDLINE abstracts. In *Proceeding of the Pacific Symposium on Biocomputing*, volume 6, pages 483–95, Hawaii, USA, 2001.
99. TK Jenssen, A Laegreid, J Komorowski, and E. Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28(1):21–28, 2001.
100. Edward M. Marcotte, Ioannis Xenarios, and David Eisenberg. Mining literature for protein-protein interactions. *Bioinformatics*, 17(4):359–363, 2001.
101. Udo Hahn and Martin Romarker. Rich knowledge capture from medical documents in the MEDSYNDIKATE system. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 338–349, Hawaii, U.S.A, 2002.
102. Jae-Hong Eom and Byoung-Tak Zhang. PubMiner: Machine Learning-Based Text Mining System for Biomedical Information Mining. In *11th International Conference, AIMSA 2004, Varna, Bulgaria, Proceedings*, 2004.
103. Barbara Rosario and Marti Hearst. Multi-way Relation Classification: Application to Protein-Protein Interaction. In *HLT-NAACL'05*, Vancouver, 2005.

104. Raymond J. Mooney and Razvan Bunescu. Mining knowledge from text using information extraction. *SIGKDD Explor. Newsl.*, 7(1):3–10, 2005.
105. Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun K. Ramani, and Yuk Wah Wong. Comparative experiments on learning information extractors for proteins and their interactions. *Journal of Artificial Intelligence in Medicine*, pages 139–155, 2005.
106. Hong woo Chun, Yoshimasa Tsuruoka, Jin-Dong Kim, Rie Shiba, Naoki Nagata, Teruyoshi Hishiki, and Jun'ichi Tsujii. Extraction of Gene-Disease Relations from MedLine using Domain Dictionaries and Machine Learning. In *The Pacific Symposium on Biocomputing (PSB)*, pages 4–15, 2006.
107. Deyu Zhou, Yulan He, and Chee Keong Kwoh. Extracting Protein-Protein Interactions from the Literature using the Hidden Vector State Model. In *International Workshop on Bioinformatics Research and Applications*, Reading, UK, 2006.
108. Yulan He and Steve Young. Semantic processing using the hidden vector state model. *Computer Speech and Language*, 19(1):85–106, 2005.
109. Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(Suppl 1), 2004.
110. Blaschke Christian, Yeh Alexander, Camon Evelyn, Colosimo Marc, Apweiler Rolf, Hirschman Lynette, and Velencia Alfonso. Do you do text? *Bioinformatics*, 21(23):4199–4200, 2005.
111. C. Nédellec. Learning Language in Logic - Genic Interaction Extraction Challenge. In *Learning Language in Logic workshop (LLL05)*, pages 31–37, 2005.
112. M. Ashburner, C. Ball, J. Blake, and D. Botstein. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–9, 2000.
113. Jane Lomax. Get ready to GO! A biologist's guide to the Gene Ontology. *Briefings in Bioinformatics*, 6(3):298–304, 2005.
114. Gerald Salton. *Automatic Text Processing*. Addison-Wesley series in Computer Science, 1989.