Available online at www.sciencedirect.com
ScienceDirect
Journal of Biomedical Informatics 41 (2008) 393–407
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Methodological Review

Extracting interactions between proteins from the literature

Deyu Zhou *, Yulan He

Informatics Research Centre, The University of Reading, Reading, RG6 6BX, UK

Received 13 March 2007
Available online 15 December 2007

Abstract

During the last decade, biomedicine has witnessed a tremendous development. Large amounts of experimental and computational biomedical data have been generated along with new discoveries, which are accompanied by an exponential increase in the number of biomedical publications describing these discoveries. In the meantime, there has been a great interest with scientific communities in text mining tools to find knowledge such as protein–protein interactions, which is most relevant and useful for specific analysis tasks. This paper provides a outline of the various information extraction methods in biomedical domain, especially for discovery of protein–protein interactions. It surveys methodologies involved in plain texts analyzing and processing, categorizes current work in biomedical information extraction, and provides examples of these methods. Challenges in the field are also presented and possible solutions are discussed.

Keywords: Information extraction; Biomedicine; Computational linguistics; Machine learning; Text mining; Protein–protein interactions

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 database [1]. 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 the extraction of protein–protein interactions.

In general, to automatically extract protein–protein interactions, a system needs to consist of three to four major modules [13,14], which is illustrated in Fig. 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 et al. [15] 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,

* Corresponding author. Fax: +44 118 975 4421.
E-mail addresses: d.zhou@reading.ac.uk (D. Zhou), y.he@reading.ac.uk (Y. He).

1532-0464/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2007.11.008

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 56,186 interactions, combining information from various sources to construct a single, stable set of protein–protein interactions	dip.doe-mbi.ucla.edu
HPRD	The Human Protein Reference Database [21] 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 articles	www.hprd.org
HPID	Human Protein Interaction Database integrates the protein interactions in BIND, DIP and HPRD	www.hpid.org
IntAct	IntAct consists of an open source database and several analysis tools for protein interaction data. It now contains more than 150,000 curated binary molecular interactions	www.ebi.ac.uk/intact
MINT	Molecular INTeraction database [22] 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 1,513,782 proteins in 373 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		
CBioC	Collaborative Bio Curation [23] uses automatic text extraction as a starting point to initialize the interaction database. After that, researchers in biomedical domain contribute to the curation process by subsequent edits	cbioc.eas.asu.edu
Chilibot	Chilibot [24] is a search software for MEDLINE literature database to rapidly identify relationships between genes, proteins, or any keywords that the user might be interested	www.chilibot.net
GoPubMed	GoPubMed [25] is a search engineer that allows users to explore PubMed search results with the Gene Ontology (GO), a hierarchically structured vocabulary for molecular biology	www.gopubmed.org
iHOP	Information Hyperlinked over Proteins [26] 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
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 the 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	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
Whatizit	Whatizit is a text processing tool that can identify molecular biology terms and linking them to publicly available databases. Identified terms are wrapped with XML tags that carry additional information, such as the primary keys to the databases where all the relevant information is kept. It is also a MEDLINE abstracts search engine	www.ebi.ac.uk/webservices/whatizit/info.jsf

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 [16]. 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 [17]. Chen et al. [18] and Leser et al. [19] 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 rely-

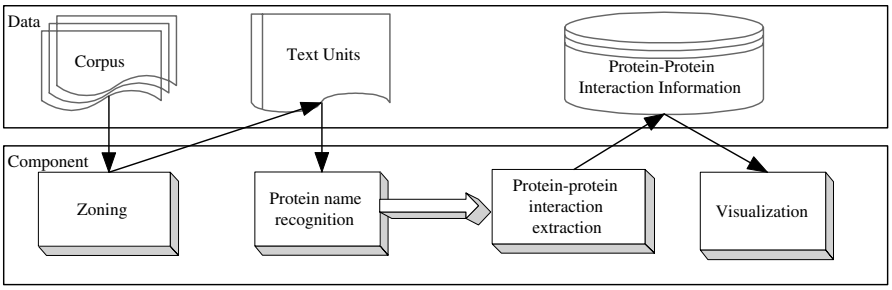


Fig. 1. A general architecture of an information extraction system for protein–protein interactions.

ing 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 [20]. Moreover, it allows users to interact with the system for ease of updating the system's knowledge base and eventually improve its performance.

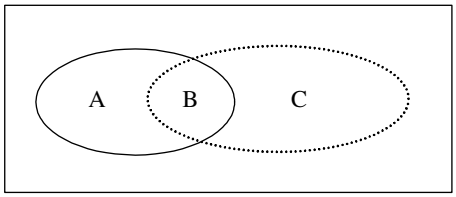
To evaluate the performance of an information extraction system, normally recall and precision values are measured. Suppose a test dataset has T positive information (for example, protein–protein interactions), and an information extraction system can extract I "positive" information. In I, only some 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 Fig. 2.

Based on the 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 ($\|T\| = 10$). An information extracting system extracts 11 protein–protein interactions ($\|I\| = 11$). In I, only 6 protein–protein interactions (B) can be found in T, which are considered as true positive (TP). The remaining 5 protein–protein interaction (C) can not be found in T, which are considered as false positive (FP). In T, 4 protein–protein interactions (A) are not extracted by the system, which are considered as false negative (FN). Thus, the recall of the system is $6/(6+4) = 60\%$ and the precision is $6/(6+5) = 54.5\%$.



T=A+B is the positive information in test data which need s to be extracted.
I=B+C is the extracted results including positive and negative information.

Fig. 2. Venn diagram of information extraction results.

Obviously, an ideal information extracting system should fulfill $\|A\| \to 0, \|C\| \to 0$. To reflect these two conditions, F-measure is defined by the harmonic (weighted) average of precision and recall [27] as:

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

where β indicates a relative weight of precision. For further details of the state of the science in text mining evaluations, please refer to Hersh [28].

In this paper, we focus on the protein–protein interaction extraction module and provide a brief survey and classification on the developed methodologies. In general, the methods proposed so far rely on the techniques from one or more areas [29–32] including Information Retrieval (IR) [27,33], Machine Learning (ML) [34,35], Natural Language Processing (NLP) [36–38], Information Extraction (IE) [39–42] and Text Mining [43–48]. Earlier work focused on limited linguistic context and relied on word co-occurrences and pattern matching. Later computational linguistic techniques that could handle relations in complex sentences were employed. The surveyed work illustrates the progress of the field and shows the increasing complexity of the proposed methodologies.

The rest of the paper is organized as follows. The next section presents a survey of various methods applied in automatical extraction of protein–protein interactions from the literature. In succession, challenges are identified and possible solutions are suggested.

2. 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.

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

2.1. 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.

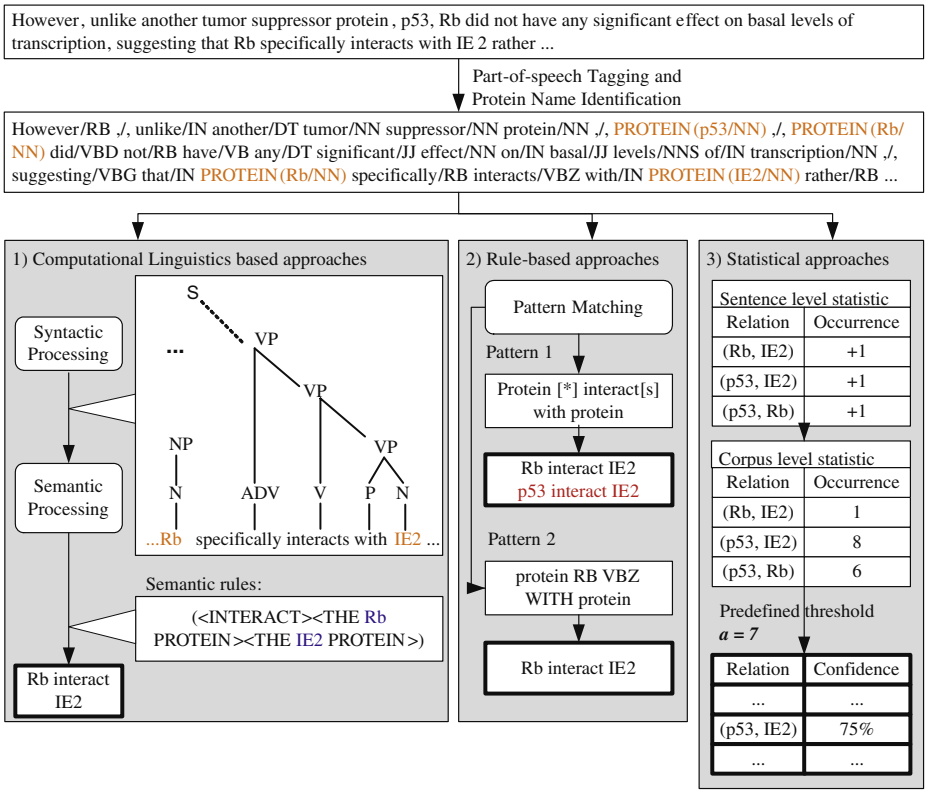


Fig. 3. General dataflow of information extraction system employing different methodologies.

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

2.1.1. Shallow parsing approaches

Shallow parsers [49–53] 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 et al. used shallow parser, EngCG, to generate three kinds of tags, such as syntactic, morphological, and boundary tags [49]. 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 et al. [51] 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 et al. [52] targeted ''inhibit'' relations in the text and also built an FSA to recognize these relations. Leroy et al. [53] 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 texts, 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.

2.1.1.1. An example. To delve into the mechanism of shallow parsing, the method reported in [53] is used to illustrate the process of detecting relations from free text.

Consider that prepositions indicate different types of relations between phases and can be distinguished based on their operative classes, the parser is based on four cascaded FSAs with one FSA for basic sentences and the other three FSAs dealing with the three top highly occurred prepositions (*by, of, in*). Fig. 4 depicts an overview of the four FSAs.

The FSA for Basic Sentences (BS-FSA) is used to model short basic sentence containing minimally two nouns or noun phrases and a verb. The structure of BS-FSA and an example is given in Fig. 4a.

The FSA for the preposition ''of'' (OF-FSA) deals with structures surrounding one or two ''of's''. It can handle two subtypes of patterns. The simple pattern involves noun phrases only, such as the example 1) shown in Fig. 4b-1. The complex pattern contains nominalizations (turning a verb or an adjective into a noun), such as the example 2) shown in Fig. 4b-2.

The FSA dealing with the preposition ''by'' (BY-FSA) can stand alone or can be cascaded with the OF-FSA. When on its own, the FSA requires the presence of a verb and two noun phrase or nominalizations, such as the example shown in Fig. 4c.

The FSA dealing with the preposition ''in'' (IN-FSA) can stand alone when there is a verb available, or it can be combined with the OF- or BY-FSA. The structure of the IN-FSA and an example is given in Fig. 4d.

When the parser reaches an end state successfully, the original relation is extracted to fill in the parser relation template which contains up to five elements, such as relation negation, left-hand side elements, connector modifier, connector, and right-hand side elements. For example, the relation extracted from the abstract title ''Regulation of E2F1 activity by acetylation'', is ''acetylation (left-hand side elements), regulates (connector), E2F1 (right-hand side elements)''.

Obviously, shallow parsers perform well for capturing relatively simple binary relationships between entities in a sentence, but fail to recognize more complex relationships expressed in various coordinating and relational clauses. For sentences containing complex relations between three or more entities, such approaches usually yield erroneous results. Approaches based on full-sentence parsing tend to be more precise.

2.1.2. 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 approaches have been proposed [10,54–63]. 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.

2.1.2.1. Rationalist methods. Yakushiji et al. [57] used a general full parser with grammar 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 et al. [56] 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)[1]. The recall and precision rate of the system were reported to be 48%

[1] In linguistics and computer science, a CFG is a formal grammar in which every production rule is of the form $V \rightarrow w$ where V is a non-terminal symbol and w is a string consisting of terminals and/or non-terminals. The term ''context-free'' comes from the fact that the non-terminal V can always be replaced by w, regardless of the context in which it occurs. Context-free grammars are powerful to describe the structure of sentences, and also simple enough to allow the construction of efficient parsing.

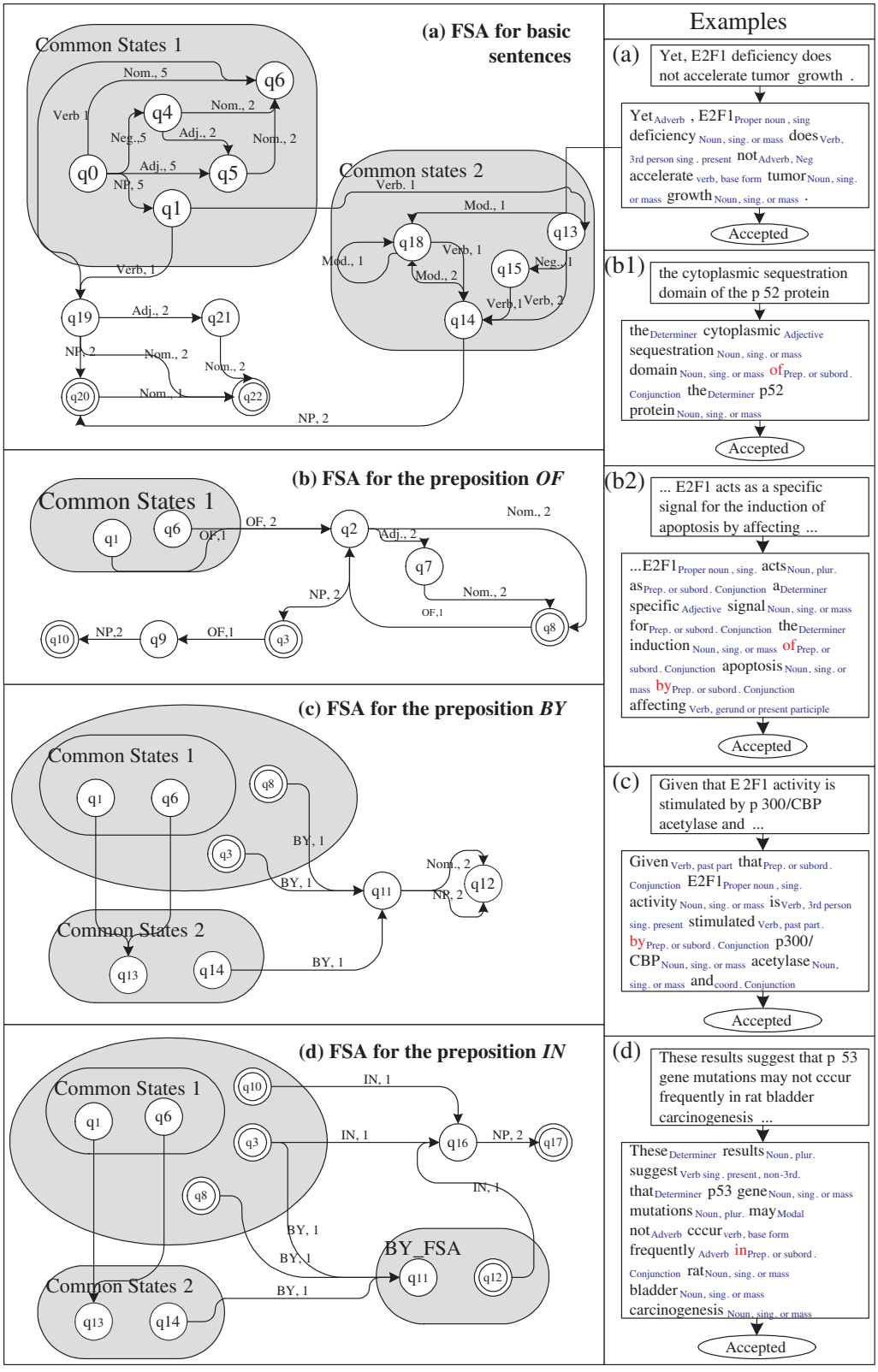


Fig. 4. Four finite-state automata describing sentence structures (Nom., nominalization; Mod, modifier; Neg., negation; NP, noun phraser or noun; Adj., adjective).

and 80%. Temkin and Gilder [60] 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 et al. [61] investigated link grammar parsing for extracting biochemical interactions. It can handle many syntactic structures and is computationally rela-

tively efficient. A better overall performance was achieved compared to those biomedical term co-occurrence based methods. Ahmed et al. [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 [58], 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.

2.1.2.2. An example. The process of using deep parsing based on rationalist methods to detect protein–protein interactions can be illustrated by the method proposed in [60], which employs a predefined context-free grammar (CFG).

To develop a concise set of grammar production of rules allowing for the detection of protein, gene, and small molecule (PGSM) interactions, a large corpus of 500 non-topic specific scientific abstracts pulled from PubMed [1] containing various representations of interaction data in unstructured text is manually analyzed. Biochemists read and highlighted the abstracts for relevant sentences describing interactions that were then used to derive the production rules. Fig. 5 shows the parsing process using the defined CFG.

2.1.2.3. Empiricist methods. Many empiricist methods [59,62] have been proposed to automatically generate the language model to mimic the features of unstructured sentences. For example, Seymore et al. [54] used Hidden Markov Model (HMM) for extracting important fields from the headers of computer science research papers. Following the trend, Ray and Craven [55] applied HMM to the biomedical domain to describe the structure of sentences. Skounakis et al. [64] proposed an approach that is based on hierarchical HMMs to represent the grammatical structure

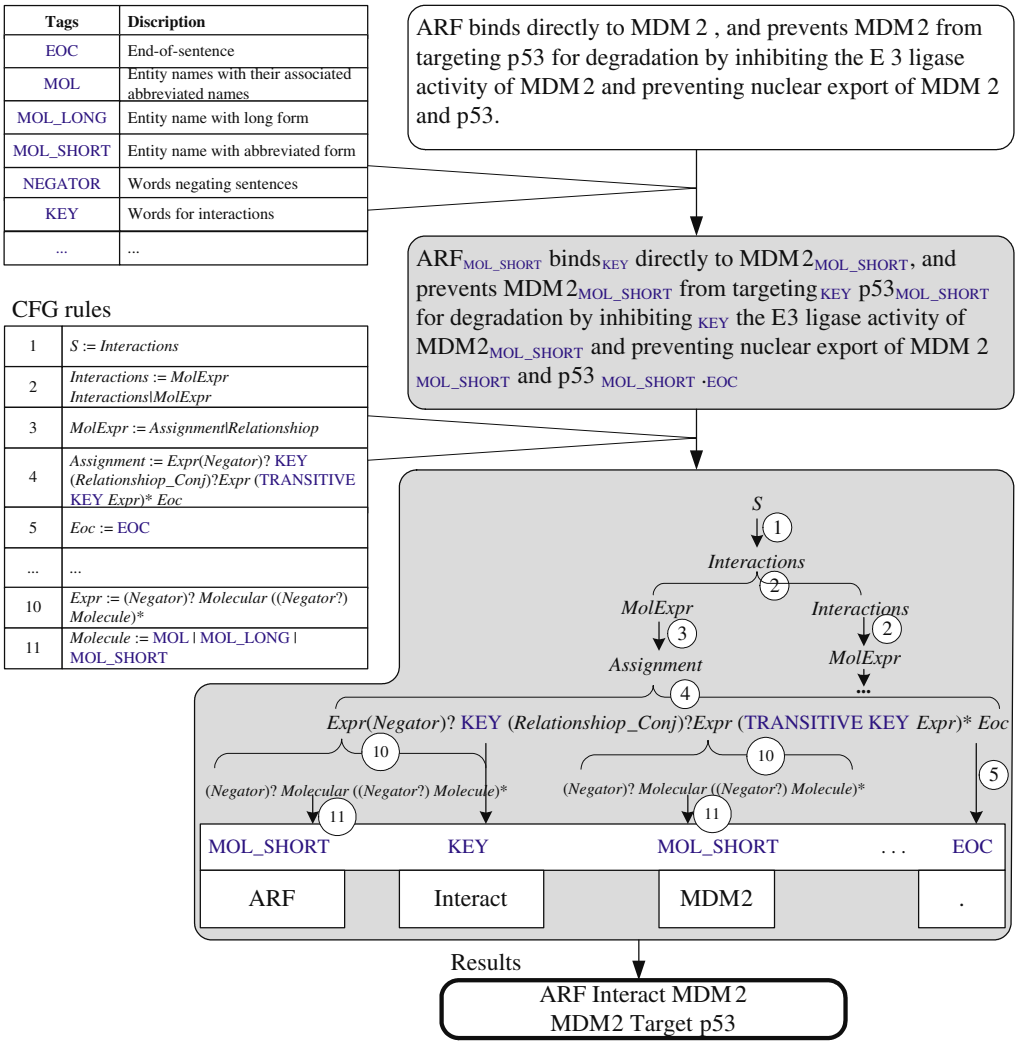


Fig. 5. A parsing example using CFG.

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 [65], 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. Recently, Katrin et al. [66] proposed RelEx based on the dependency parse trees to extract relations in biomedical texts. It was applied on one million MEDLINE abstracts to extract gene and protein relations. About 150,000 relations were extracted with an estimated performance of both 80% precision and 80% recall. Rinaldi et al. [67] also employed a probabilistic dependency parser, Pro3Gres, to output functional dependency structures. Based on these structures, functional relations (e.g. interactions between proteins and genes) were extracted. Experiments were conducted on tow different corpora, the GENIA corpus and the ATCR corpus. Precision values range from 52% to 90% and recall values range from 40% to 60% based on different evaluation methods.

2.1.2.4. An example. To show the way of using empiricist deep parsing to extract protein–protein interactions, the method proposed in [66] is used, which employs the Standford Lexicalized Parser[2] to generate dependency parse trees. The parser is based on the unlexicalized probabilistic context-free grammars (PCFGs) [68]. Usually, two data sets are employed to train the parser, one is the standard LDC Penn Treebank WSJ secs 2–21 and the other is an augmented one, better for questions, commands, and text from different genres.

The whole process can be divided into three steps, preprocessing, extracting and postprocessing. In the preprocessing step, a dependency parse tree is generated for each sentence by the Standford lexicalized parser. Also, gene and protein names are recognized based a synonym dictionary. Moreover, noun-phase chunks are identified and combined with the dependency parse trees to generate chunk dependency parse trees. Based on the chunk dependency parse trees, pathes connecting pairs of proteins are identified in the extracting step based on the three predefined rules. These rules describe the most frequently used constructs for depicting relations, such as effector-relation-effectee (e.g. ''IL-4 suppressed IL-2 and IFN-gamma mRNA levels in primary human T cells, and addition of anti-CD28 antibodies relieved this suppression''), relation-of-effectee-by-effector (e.g. ''Taken together, these results indicate that IL-6 and IL-8 release by protein I/II-activated FLSs is regulated by FAK independently of Tyr-397 phosphorylation''), and relation-between-effector-and-effectee (e.g. ''In human AM, Pc promoted direct interaction of MR and TLR2, IL-8 release was reduced markedly upon...''). Candidate relations are created for each sentence base on these extracted paths. These candidate relations are filtered in the postprocessing step. The filtration consists of negation check (excluding negated relations) and restricting to focus domain (excluding the relations which do not contain any word in a set of predefined relation restriction terms). After filtration, effector and effectee detection and enumeration resolution are performed. For a given sentence, Fig. 6 shows the internal results in each step. It can be observed that this method depends highly on the precision of dependency parse tree generated by the Stanford lexicalized parser.

Full-parsing methods analyze the structure of an entire sentence in order to achieve higher accuracy. However, they still cannot handle all kinds of sentences, especially those with complex structures. Moreover, analyzing the whole sentence structure incurs higher computational and time complexity.

2.2. Rule-based approaches

In rule-based approaches [6,7,9,12,69–77], a set of rules need to be defined which may be expressed in forms of regular expressions over words or part-of-speech (POS) tags. Based on the rules, relations between entities that are relevant to tasks such as proteins, can be recognized.

Ng and Wong [69] 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 et al. [72] 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 and Valencia [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 [74], 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. Leroy and Chen [73] employed preposition-based parsing to generate templates. It achieved a template precision of 70% when processing the 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 et al. [75] tried to automatically construct the protein–protein interaction patterns. At first, part-of-

[2] http://nlp.stanford.edu/downloads/lex-parser.shtml

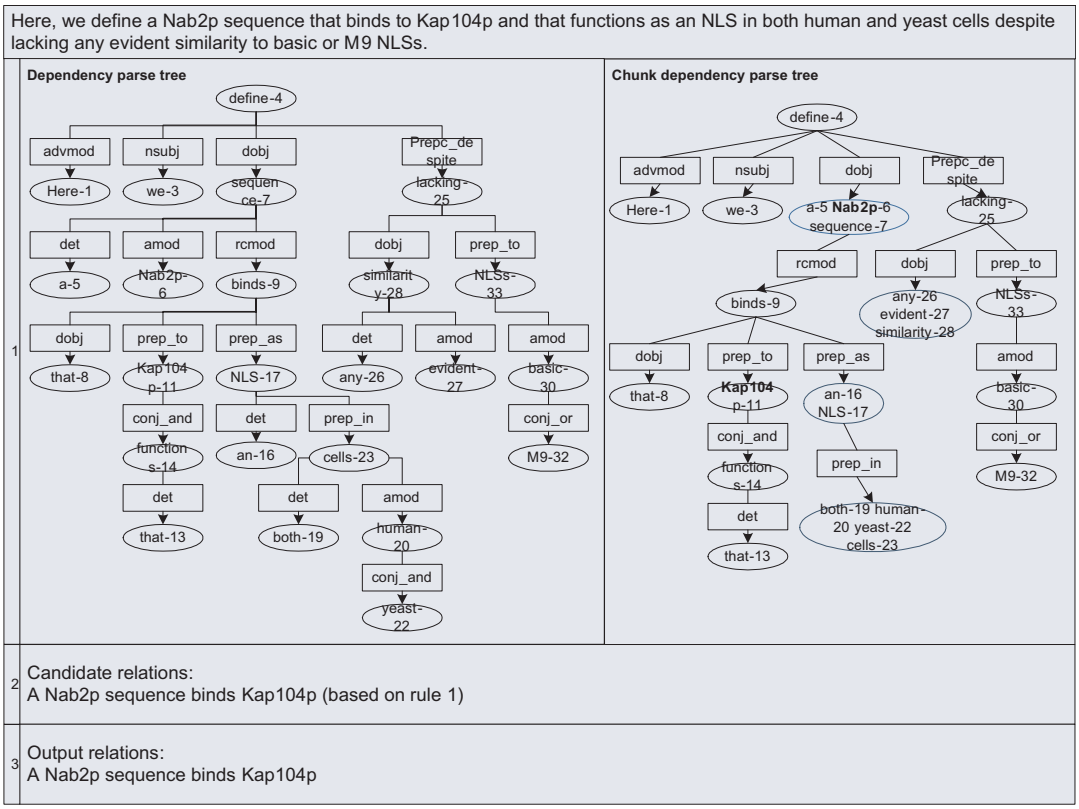


Fig. 6. An example employing the Stanford Lexicalized Parser to generate chunk dependency parse tree.

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 et al. [78] 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.

2.2.1. An example

In this section, we illustrate the process of employing the rule-based method proposed in [72] to detect protein–protein interactions. The whole process can be divided into three steps:

2.2.1.1. Identification of protein names. Protein names were first identified from sentences based on a predefined biomedical entity dictionary.

2.2.1.2. Preprocessing compound or complex sentences. Sentences were firstly parsed by employing POS tagging. Then predefined rules based on the generated POS tags were applied to split those complex sentences. For example, the sentence "The gap1 mutant blocked stable association of Ste4p with the plasma membrane, *and* the ste18 mutant blocked stable association of Ste4p with both plasma membranes and internal membranes" is split into two parts, one is "The gap1 mutant blocked stable association of Ste4p with the plasma membrane", the other is "the ste18 mutant blocked stable association of Ste4p with both plasma membranes and internal membranes", when applying the rule below:

If a sentence matches the pattern P1 [(,CCDT)|(,IN)|: |;] P2, where CC denotes coordinating conjunction and DT denotes determiner, **then** the sentence can be split into P1 and P2.

2.2.1.3. Recognition of the protein–protein interaction. A set of word patterns was defined for the recognition of protein–protein interactions. For example, the defined word patterns could be "A interact with B", "interaction of A (with—and) B", "interaction (between|among) A and B" and so on. A and B here indicate protein names. For the sentence "We define a *Nab2p* sequence that binds to Kap104p", the interaction "bind: Nab2p, Kap104p" can be extracted using the predefined rule A bind to B. To process negative sentences, which describe a lack of interaction, several pattern of regular expression were constructed, such as PROTEIN1.* not (interact|associate|bind|complex).*PROTEIN2.

2.2.2. Discussion

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. Specif-

ically, 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.

2.3. 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 2.1.2. 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 [79–86,8,87–91]. They can be further divided into different types based on the mining units, such as abstracts, sentences and so on.

Approaches proposed in Andrade and Valencia [79] and Marcottle et al. [85] aim to extract protein–protein interactions from a set of abstracts. Andrade and Valencia [79] used a group of relevant documents against a set of random documents to extract domain specific information such as gene functions and interactions. Marcottle et al. [85] was only interested in retrieving a large number of documents that probably contained information about protein–protein interactions. We will discuss it in detail in Section 2.3.1.

The first machine-learning sentence-based information extraction system in molecular biology was described in Craven and Kumlien [81]. 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 [82–84,8] focus on a pair of proteins and detect the relations between them using probability scores. Stapley and Benoit [82] 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 et al. [83] furthered the method to discover relationships using more complicated computation on co-occurrences. Jenssen et al. [84] used a similar approach to find relations between human gene clusters obtained from DNA array experiments. Donaldson et al. [8] constructed PreBIND and Textomy—an information extraction system that uses support vector machines to evaluate the importance of protein–protein interactions.

2.3.1. An example

In this section, we illustrate the process of detecting protein–protein interactions using the method proposed in [85]. The whole process can be divided into three steps.

2.3.1.1. Build the training and testing corpora. The training corpus contains 260 papers cited by the Database of Interacting Proteins (DIP). Testing data which are denoted as *Yeast MEDLINE* were obtained from MEDLINE by querying the PubMed using the term "*Saccharomyces cerevisiae*" in the title, abstract, or MESH terms.

2.3.1.2. Construct discriminating words. The discriminating words are defined as those words which may be useful for discriminating the training abstracts from other abstracts. A dictionary was constructed containing the frequencies of the 60,000 most common words used more than three times in the *Yeast MEDLINE* abstracts. For each word in the training abstracts, the probability $P(n|N,f)$ of finding the observed number of times n given the known dictionary frequency f and the total number of words N in the training abstracts, was calculated from the Poisson distribution as

$$P(n|N,f) \approx e^{-Nf}\frac{(Nf)^n}{n!}$$

In practice, the log of the probability was calculated as $\ln P(n|N,f) \approx -Nf + n\ln Nf - \ln n!$. The 500 words in the training abstracts with the most negative log probability scores were selected as *discriminating words*.

2.3.1.3. Score each abstract in Yeast MEDLINE by its likelihood of discussing protein–protein interaction. Assume that an abstract has N words, the discriminating word set D has M distinct words, n_i denotes the number of occurrences of the discriminating word d_i. At first, modeling the $P(n_i|\text{AbstractSet})$ with a Poisson distribution gives

$$P(n_i|\text{InteractionAbstract}) = \frac{e^{-Nf_{I,i}}(Nf_{I,i})^{n_i}}{(n_i)!}$$

$$P(n_i|\text{NonInteractionAbstract}) = \frac{e^{-Nf_{N,i}}(Nf_{N,i})^{n_i}}{(n_i)!}$$

where the $f_{I,i}$ is the frequency of the discriminating word i in the training abstracts, $f_{N,i}$ is the dictionary frequency of the discriminating word i. Based on the Bayesian form, the following equation can be obtained:

$$\frac{P(\text{InactionAbstract}|n_i)}{P(\text{NonInactionAbstract}|n_i)} = \frac{e^{-Nf_{I,i}}(f_{I,i})^{n_i}}{e^{-Nf_{N,i}}(f_{N,i})^{n_i}} \times \frac{P(\text{InteractionAbstract})}{P(\text{NonInteractionAbstract})}$$

The score is deduced as following:

$$\text{Likehood} = \prod_{i=1}^{M}\left(\frac{\text{P}(\text{InactionAbstract}|n_i)}{\text{P}(\text{NonInactionAbstract}|n_i)}\right) = \prod_{i=1}^{\text{M}}\left(\frac{\text{e}^{-Nf_{I,i}}(f_{I,i})^{n_i}}{\text{e}^{-Nf_{N,i}}(f_{N,i})^{n_i}}\right) \times \left(\frac{\text{P}(\text{InteractionAbstract})}{\text{P}(\text{NonInteractionAbstract})}\right)^{M}$$

As the ratio between P(InteractionAbstract) and P(NonInteractionAbstract) is constant, it can be omitted from the log calculation.

$$\text{Score} = \sum_{i}^{M}\left(n_i \ln\frac{f_{I,i}}{f_{N,i}} - N*(f_{I,i}-f_{N,i})\right)$$

2.3.2. Discussion

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. Recently, the hidden vector state model (HVS) which was previously proposed for spoken language understanding has been applied to extract protein–protein interactions [92] to strike the balance. The HVS model explores the embedded sentence structures using only lightly annotated corpus, unlike other statistical parsers which need fully annotated treebank data for training. Also the hierarchical information is embedded into the HVS model, which enable the HVS model extract the relations between proteins precisely.

2.4. Performance comparison of existing approaches

The performance of the existing protein–protein interaction extraction methods along with the data corpora they used are listed in Table 2.

As in the area of extracting information about protein–protein interactions, competitive evaluations have played important roles in pushing the fields of IE and NLP. Several evaluations have been held in recent years. BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) [93][3] 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 [94] addressed issues of functional annotation, requiring systems to identify specific text passages that supported Gene Ontology annotations for specific proteins, given full text articles. Later on, the second BioCreAtIvE challenge was held in 2006, focusing on gene mention tagging (finding the mentions of genes and proteins in sentences drawn from MEDLINE abstracts), gene normalization (producing a list of the EntrezGene identifiers for all the human genes/proteins mentioned in a collection of MEDLINE abstracts), and extraction of protein–protein interactions from text (identifying protein–protein interactions from full text papers, including extraction of excerpts from those papers that describe experimentally derived interactions). Genic Interaction Extraction Challenge [95] 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 3.

3. Challenges and possible solutions

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

- *Poor performance.* 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

[3] http://biocreative.sourceforge.net/

[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) [96,97].

Table 2
Performance of existing protein–protein interaction extraction methods and the data corpora used

Category	Result (%)		Corpus	Ref.
	Recall	Precision		
Shallow parsing	—	73	34,343 sentences from abstracts retrieved from MEDLINE using keywords "leucine zipper", "zinc finger", "helix loop helix motif"	[49]
	29	69	2,565 unseen abstracts extracted from MEDLINE with the keywords molecular, interaction and protein for year 1,998 (560k words)	[51]
	57	90	Training set consists of 500 abstracts from MEDLINE. Evaluation set consists of 56 abstracts collected using search strings "protein" and "inhibit"	[52]
	62	89	26 abstracts	[53]
Deep parsing	48	80	492 sentences out of 250,000 abstracts on cytosine in MEDLINE	[56]
	63.9	70.2	The test corpus consists of 100 randomly selected scientific abstracts from MEDLINE	[60]
	—	96	Articles from cell containing 7790 words revealing 51 binary relations	[58]
	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	[62]
	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"	[73]
	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"	[72]
	39.7	44.9	Five different sets of abstracts were used: 1. 1435 MEDLINE abstracts directly referenced from each of the *Drosophila* Swiss-prot entries 2. 4109 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. 6278 MEDLINE abstracts by expanding set (4) using Neighbors to identify all related abstracts	[70,7]
	60	87	3343 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	[78]
	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 1200 sentences	[75]

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.

- *Changeable relations between biological entities.* 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 the 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.
- *Gap between biologists and computational scientists.* Bridging the gap between biologists and computational scientists seems to be crucial to the success of biomedical IE. Currently, this field is dominated by researchers with computational background; however, the biomedical

Table 3
Online annotated corpora for the extraction of protein–protein interactions

Corpus name	Description	URL
GENIA	GENIA 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 1745 protein names, test collection consisting of 101 abstracts with 1966 protein names. The protein names in all the abstracts were annotated manually	www.sics.se/humle/projects/prothalt/
Penninite	The corpus consists of 2258 MEDLINE abstracts in two domains: (1) the molecular genetics of oncology (1158 abstracts); (2) the inhibition of enzymes of the CYP450 class (1100 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/
BioCreAtIvE II Task 3 Corpus	The corpus consists of four parts of data, for four different subtasks. Overall, the training data was derived from the content of the IntAct and MINT databases. The data files of both databases are freely accessible for download and are compliant with the HUPO PSI Molecular Interaction Format. The test set collection will consist of a collection of PubMed article abstracts in a format compliant with the training collection format	biocreative.sourceforge.net/index.html

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.

- *Self-contradictory extracted knowledge.* 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.
- *Obstacles in NLP.* 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 [98]. (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.
- *Development of gold standard for evaluation systems.* The development of the 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 discovery 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] Ioannis Xenarios, Lukasz Salwinski, Xiaoqun Joyce Duan, Patrick Higney, Sul-Min Kim, Eisenberg David. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 2002;30(1):303–5.
[3] Bader Gary D, Betel Doron, Houge Christopher WV. BIND: the biomolecular interaction network database. Nucleic Acids Res 2003;31(1):248–50.
[4] Hermjakob Henning, Montecchi-Palazzi Luisa, Lewington Chris, Mudali Sugath. IntAct: an open source molecular interaction database. Nucleic Acids Res 2004;1(32Database issue):452–5.
[5] von Mering Christian, Jensen Lars J, Snell Berend, Hooper Sean D. STRING: known and predicted protein–protein associations, integrated and transferred across organisms. Nucleic Acids Res 2005;33(Database issue):433–7.
[6] Limsoon Wong. PIES, a protein interaction extraction system. In: Proceedings of the Pacific symposium on biocomputing, Hawaii, USA, 2001. p. 520–531.
[7] Blaschke Christian, Valencia Alfonso. The frame-based module of the SUISEKI information extraction system. IEEE Intel Syst 2002;17(2):14–20.
[8] Donaldson Ian, Martin Joel, Bruijn Berry de, Wolting Cheryl, Lay Vicki. PreBIND and textomy-mining the biomedical literature for protein–protein interactions using a support vector machine. BMC Bioinformat 2003;4(11).
[9] Chiang Jung-Hsien, Yu Hsu-Chun, Hsu Huai-Jen. GIS: a biomedical text-mining system for gene information discovery. Bioinformatics 2004;20(1):120–1.
[10] Syed Toufeeq Ahmed, Deepthi Chidambaram, Hasan Davulcu, 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. p. 54–61.

[11] Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. In: Proceedings of Pacific symposium biocomputing. 2000. p. 517–28.
[12] David PA, Buxton Corney Bernard F, Langdon William B, Jones David T. BioRAT: extracting biological information from full-length papers. Bioinformatics 2004;20(17):3206–13.
[13] Rzhetsky Andrey, Iossifov Ivan, Koike Tomohiro, Krauthammer Michael, Kra Pauline. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. J Biomed Informatics 2004;37(1):43–53.
[14] Brigitte Mathiak and Silke Eckstein, Five steps to text mining in biomedical literature. In: Data mining and text mining for bioinformatics european workshop, 2004.
[15] Ding J, Berleant D, Nettleton D, Wurtele E. Mining MEDLINE: abstracts, sentences, or phrases. In: Proceedings of the Pacific symposium on biocomputing, Hawaii, USA, 2002. p. 326–37.
[16] Krauthammer Michael, Nenadic Goran. Term identification in the biomedical literature. J Biomed Informatics 2004;37(6):512–26.
[17] Pearson H. Biology's name game. Nature 2001;411(6838):631–2.
[18] Chen Lifeng, Liu Hongfang, Friedman Carol. Gene name ambiguity of eukaryotic nomenclatures. Bioinformatics 2005;21(2):248–56.
[19] Leser Ulf, Hakenberg Jörg. What makes a gene name? Named entity recognition in the biomedical literature. Brief Bioinform 2005;6(4):257–69.
[20] Drabkin Harold J, Hollenbeck Christopher, Hill David P, Blake Judith A. Ontological visualization of protein–protein interactions. BMC Bioinform 2005(29).
[21] Peri Suraj, Daniel Navarro J, Amanchy Ramars. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 2003;13:2363–71.
[22] Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G. MINT: a Molecular INTeraction database. FEBS lett 2002;513(1):135–40.
[23] Chitta Baral, Graciela Gonzalez, Anthony Gitter, Craig Teegarden, Amanda Zeigler. CBioC: beyond a prototype for collaborative annotation of molecular interactions from the literature. In: Computational systems bioionformatics conference, 2007.
[24] Chen Hao, Sharp Burt M. Content-rich biological network constructed by mining PubMed abstracts. BMC Bioinform 2004;8(5):147.
[25] Doms Andreas, Schroeder Michael. GoPubMed: exploring PubMed with the gene ontologys. Nucleic Acids Res 2005;33:W783–6.
[26] Hoffmann Robert, Valencia Alfonso. A gene network for navigating the literature. Nature Gene 2004;36:664.
[27] van Rijsbergen CJ. Information Retrieval. http://www.dcs.gla.ac.uk/Keith/Preface.html, 1999.
[28] Hersh William. Evaluation of biomedical text-mining systems: lessons learned from information retrieval. Brief Bioinformat 2005;6(4):344–56.
[29] Andrade MA, Bork P. Automated extraction of information in molecular biology. FEBS Lett 2000;476(1-2):12–7.
[30] Hirschman Lynette, Park Jong C, Tsujii Junichi, Wong Limsoon, Wu Cathy H. Accomplishments and challenges in literature data mining for biology. Bioinformatics 2002;18(12):1553–61.
[31] Shatkay Hagit, Feldman Romen. Mining the biomedical literature in the genomic era: an overview. J Comput Biol 2003;10(6):821–55.
[32] Jensen Lars Juhl, Saric Jasmin, Bork Peer. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Gene 2006;7:119–29.
[33] William R. Hersh. Information retrieval: a health and biomedical perspective. 2003.
[34] Jeffrey T. Chang, Using machine learning to extract drug and gene relationships from text. PhD thesis, Stanford University, September 2003.
[35] 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.
[36] K. Bretonnel Cohen and Lawrence Hunter. Natural language processing and systems biology. In: Dubitzky Werner, Azuaje Francisco, editors. Series: computational biology, vol. 5. 2004.
[37] Yandell MD, Majoros WH. Genomics and natural language processing. Nat Rev Gene 2002;3(8):601–10.
[38] Hunter Lawrence, Bretonnel Cohen K. Biomedical language processing: What's beyond PubMed? Mol Cell 2006;21(5):589–94.
[39] Cardie Claire. Empirical methods in information extraction. AI Magazine 1997;18(4):65–80.
[40] Blaschke Christian, Hoffmann Robert, Oliveros Juan Carlos, Valencia Alfonso. Extracting information automatically from biological literature. Comp Funct Genomics 2001;2(5):310–3. 2.
[41] Hamish Cunningham. In: Encyclopedia of language and linguistics. 2nd ed. Information Extraction Automatic. Elsevier; 2005.
[42] Skusa Andre, Rüegg Alexander, Köhler Jacob. Extraction of biological interaction networks from scientific literature. Brief Bioinformatics 2005;6(3):263–76.
[43] Berry De Bruijn and Joel Martin. Literature mining in molecular biology. In: Proceedings of the EFMI workshop on natural language processing in biomedical application, 2002. p. 1–5, 2002.
[44] Krallinger Martin, Erhardt Ramon Alonso-Allende, Valencia Alfonso. Text-mining approaches in molecular biology and biomedicine. Drug Discov Today 2005;10(6):439–45.
[45] Spasic Irena, Ananiadou Sophia, McNaught John, Kumar Anand. Text mining and ontologies in biomedicine: making sense of raw text. Brief Bioinformatics 2005;6(3):239–51.
[46] Cohen Aaron M, Hersh William R. A survey of current work in biomedical text mining. Brief Bioinformatics 2005;6(1):57–71.
[47] Sophia Ananiadou and John Mcnaught. Text mining for biology and biomedicine. 2006.
[48] Shatkay Hagit, Craven M. Biomedical text mining. Cambridge, Massachussets: MIT Press; 2007.
[49] Takeshi Sekimizu, Hyun S. Park, Junichi Tsujii. Identifying the interaction between genes and gene products based on frequently seen verbs in MEDLINE abstracts. In: Workshop on genome informatics, vol. 9. 1998. p. 62–71.
[50] TC. Rindflesch, L Hunter, Aronson AR. Mining molecular binding terminology from biomedical text. In: Proceedings of AMIA symposium, 1999. p. 127–31.
[51] James Thomas, David Milward, Christos Ouzounis, Stephen Pulman, Mark Carroll. Automatic extraction of protein interactions from scientific abstracts. In: Proceedings of the Pacific symposium on biocomputing, Hawaii, USA, 2000. p. 541–52.
[52] J. Pustejovsky, J. Castano, J. Zhang, M. Kotecki, Cochran B. Robust relational parsing over biomedical literature: extracting inhibit relations. In: Proceedings of the Pacific symposium on biocomputing, Hawaii, USA, 2002, p. 362–373.
[53] Leroy Gondy, Chen Hsinchun, Martinez Jesse D. A shallow parser based on closed-class words to capture relations in biomedical text. J Biomed Informatics 2003;36(3):145–58.
[54] Kristie Seymore, Andrew McCallum, Roni Rosenfeld. Learning hidden Markov model structure for information extraction. In: AAAI 99 workshop on machine learning for information extraction, 1999.
[55] Souyma Ray, 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), 2001. p. 1273–9.
[56] Jong C. Park, Hyun Sook Kim, Jung Jae Kim. Bidirectional incremental parsing for automatic pathway identification with combinatory categorical grammar. In: Proceedings of the Pacific symposium on biocomputing, vol. 6. Hawaii, USA, 2001. p. 396–407.
[57] Akane Yakushiji, Yuka Tateisi, Yusuke Miyao, Jun ichi Tsujii. Event extraction from biomedical papers using a full parser. In: Proceedings of the Pacific symposium on biocomputing, vol. 6. Hawaii, USA, 2001. p. 408–419.
[58] Friedman Carol, Kra Hong Yu Pauline, Krauthammer Michael, Rzhetsky Andrey. GENIES: a natural-language processing system

for the extraction of molecular pathways from journal articles. Bioinformatics 2001;17:S74–82.
[59] Novichkova Svetlana, Egorov Sergei, Daraselia Nikolai. MedScan, a natural language processing engine for MEDLINE abstracts. Bioinformatics 2003;19(13):1699–706.
[60] Temkin Joshua M, Gilder Mark R. Extraction of protein interaction information from unstructured text using a context-free grammar. Bioinformatics 2003;19(16):2046–53.
[61] Jing Ding, Daniel Berleant, Jun Xu, Andy Fulmer W. Extracting biochemical interactions from MEDLINE using a link grammar parser. In: 15th IEEE International conference on tools with artificial intelligence (ICTAI'03), 2003.
[62] Daraselia Nikolai, Yuryev Anton, Egorov Sergei, Novichkova Svetalana, Nikitin Alexander, Mazo llya. Extracting human protein interactions from MEDLINE using a full-sentence parser. Bioinformatics 2004;20(5):604–11.
[63] Shengyang Tan, Chee Keong Kwoh. Cytokine information system and pathway visualization. In: International joint conference of InCoB, AASBi and KSBI (BIOINFO2005), 2005.
[64] Marios Skounakis, Mark Craven, Souyma Ray. Hierarchical hidden markov models for information extraction. In: Proceedings of the 18th International joint conference on artificial intelligence, 2003.
[65] Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, James Dowdall. Mining relations in the GENIA corpus. In: second european workshop on data mining and text mining for bioinformatics, 2004.
[66] Fundel Katrin, Kuffner Robert, Zimmer Ralf. RelEx-Relation extraction using dependency parse trees. Bioinformatics 2007;23(3):365–71.
[67] Rinaldi Fabio, Schneider Gerold, Kaljurand Kaarel, Hess Michael, Andronis Christos, Konstandi Ourania, et al. Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach. Artif Intel Med 2007;39:127–36.
[68] Dan Klein and Christopher D. Manning, Accurate unlexicalized parsing. In: The 41st meeting of the association for computational linguistics, 2003. p. 423–430.
[69] See-Kiong Ng and Marie Wong. Toward routine automatic pathway discovery from on-line scientific text abstracts. In: Proceedings of the 12th national conference on artificial intelligence, 1999.
[70] Blaschke Christian, Andrade Miguel A, Ouzounis Christos, Valencia Alfonso. Automatic extraction of biological information from scientific text: protein–protein interactions. In: Proceedings of the seventh International conference on intelligent systems for molecular biology. AAAI Press; 1999. p. 60–7.
[71] Valencia A, Blaschke C. The potential use of SUISEKI as a protein interaction discovery tool. In: Workshop on genome informatics, vol. 12. 2001. p. 123–134.
[72] Ono Toshihide, Hishigaki Haretsugu, Tanigam Akira, Takagi Toshihisa. Automated extraction of information on protein–protein interactions from the biological literature. Bioinformatics 2001;17(2):155–61.
[73] Gondy Leroy and Hsinchun Chen. Filling preposition-based templates to capture information from medical abstracts. In: Pacific symposium biocomputing, 2002. p. 350–361.
[74] Proux Denys, Rechenmann Franois, Julliard Laurent. A pragmatic information extraction strategy for gathering data on genetic interactions. In: Proceedings of the eighth International conference on intelligent systems for molecular biology. AAAI Press; 2000. p. 279–85.
[75] Huang Minlie, Zhu Xiaoyan, Hao Yu. Discovering patterns to extract protein–protein interactions from full text. Bioinformatics 2004;20(18):3604–12.
[76] Hong-Woo Chun, Young-Sook Hwang, Hae-Chang Rim. Unsupervised event extraction from biomedical literature using co-occurrence information and basic patterns. In: The lecture notes in artificial intelligence, 2005. p. 777–86.
[77] Hao Yu, Zhu Xiaoyan, Huang Minlie, Li Ming. Discovering patterns to extract protein–protein interactions from the literature: Part II. Bioinformatics 2005;21(15):3294–300.
[78] Tu Minh Phuong, Doheon Lee, 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.
[79] Andrade Miguel A, Valencia Alfonso. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. Bioinformatic 1998;14(7):600–7.
[80] Mark Craven. Learning to extract relations from MEDLINE. In: Proceedings of the AAAI 1999 workshop on machine learning for information extraction, 1999. p. 25–30.
[81] Mark Craven, Johan Kumlien. Constructing biological knowledge bases by extracting information from text sources. In: Proceedings of the 7th International conference on intelligent systems for molecular biology. Heidelberg, Germany, 1999. p. 77–86, .
[82] Stapley B, Benoit G. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. In: Proceedings of the Pacific symposium on biocomputing. Hawaii, USA, 2000. p. 529–40.
[83] Stephens M, Palakal M, Mukhopadhyay S, Raje R, Mostafa J. Detecting gene relations from MEDLINE abstracts. In: Proceeding of the Pacific symposium on biocomputing, vol. 6. Hawaii, USA, 2001. p. 483–95.
[84] Jenssen TK, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. Nat Genetics 2001;28(1):21–8.
[85] Marcotte Edward M, Xenarios Ioannis, Eisenberg David. Mining literature for protein–protein interactions. Bioinformatics 2001;17(4):359–63.
[86] Udo Hahn, Martin Romarker. Rich knowledge capture from medical documents in the MEDSYNDIKATE system. In: Proceedings of the Pacific symposium on biocomputing. Hawaii, USA, 2002. p. 338–49.
[87] Jae-Hong Eom, Byoung-Tak Zhang. PubMiner: machine learning-based text mining system for biomedical information mining. In: 11th International conference, AIMSA 2004, Varna, Bulgaria, Proceedings, 2004.
[88] Barbara Rosario, Marti Hearst. Multi-way relation classification: application to protein–protein interaction. In: HLT-NAACL'05. Vancouver, 2005.
[89] Mooney Raymond J, Bunescu Razvan. Mining knowledge from text using information extraction. SIGKDD Explor Newsl 2005;7(1):3–10.
[90] Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun K. Ramani, Yuk Wah Wong. Comparative experiments on learning information extractors for proteins and their interactions. J Artif Intel Med 2005;139–155.
[91] Hong woo Chun, Yoshimasa Tsuruoka, Jin-Dong Kim, Rie Shiba, Naoki Nagata, Teruyoshi Hishiki, Jun'ichi Tsujii. Extraction of gene-disease relations from MedLine using domain dictionaries and machine learning. In: The Pacific symposium on biocomputing (PSB), 2006. p. 4–15.
[92] Deyu Zhou, Yulan He, 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.
[93] Hirschman Lynette, Yeh Alexander, Blaschke Christian, Valencia Alfonso. Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics 2004;6(Suppl 1).
[94] Christian Blaschke, Alexander Yeh, Evelyn Camon, Marc Colosimo, Rolf Apweiler, Lynette Hirschman, et al. Do you do text? Bioinformatics 2005;21(23):4199–200.
[95] Nédellec C. Learning language in logic–genic interaction extraction challenge. In: Learning language in logic workshop (LLL05), 2005. p. 31–37.
[96] Ashburner Michael, Ball Catherine A, Blake Judith A, Botstein3 David. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genetics 2000;25(1):25–9.
[97] Lomax Jane. Get ready to GO! A biologist's guide to the Gene Ontology. Brief Bioinformatics 2005;6(3):298–304.
[98] Salton Gerald. Automatic text processing. Addison-Wesley series in Computer Science 1989.