Cellier *et al. Journal of Biomedical Semantics* (2015) 6:27
DOI 10.1186/s13326-015-0023-3

JOURNAL OF BIOMEDICAL SEMANTICS

RESEARCH ARTICLE Open Access

Sequential pattern mining for discovering gene interactions and their contextual information from biomedical texts

Peggy Cellier[1*], Thierry Charnois[2*], Marc Plantevit[3], Christophe Rigotti[4], Bruno Crémilleux[5], Olivier Gandrillon[6], Jiří Kléma[7] and Jean-Luc Manguin[5]

Abstract

Background: Discovering gene interactions and their characterizations from biological text collections is a crucial issue in bioinformatics. Indeed, text collections are large and it is very difficult for biologists to fully take benefit from this amount of knowledge. Natural Language Processing (NLP) methods have been applied to extract background knowledge from biomedical texts. Some of existing NLP approaches are based on handcrafted rules and thus are time consuming and often devoted to a specific corpus. Machine learning based NLP methods, give good results but generate outcomes that are not really understandable by a user.

Results: We take advantage of an hybridization of data mining and natural language processing to propose an original symbolic method to automatically produce patterns conveying gene interactions and their characterizations. Therefore, our method not only allows gene interactions but also semantics information on the extracted interactions (e.g., modalities, biological contexts, interaction types) to be detected. Only limited resource is required: the text collection that is used as a training corpus. Our approach gives results comparable to the results given by state-of-the-art methods and is even better for the gene interaction detection in AIMed.

Conclusions: Experiments show how our approach enables to discover interactions and their characterizations. To the best of our knowledge, there is few methods that automatically extract the interactions and also associated semantics information. The extracted gene interactions from PubMed are available through a simple web interface at https://bingotexte.greyc.fr/. The software is available at https://bingo2.greyc.fr/?q=node/22.

Keywords: Data mining, Sequential pattern mining, Natural language processing, Information extraction, Gene interactions

Introduction

Literature on biology and medicine represents a huge amount of knowledge: more than 24 million publications are currently listed in the PubMed repository [1]. These text collections are large and it is difficult for biologists to fully take benefit from this incredible amount of knowledge. A critical challenge is then to extract relevant and useful knowledge spread in such collections. Text mining and Natural Language Processing (NLP) are rapidly becoming an essential component of various bio-applications. These techniques have widely been applied to extract and exploit background knowledge from biomedical texts.

Among many tasks, a crucial issue is the annotation of a large amount of genetic information. NLP, and Information Extraction (IE) in particular, aim to provide accurate processing to extract specific knowledge such as named entities (e.g., gene, protein) and relationships between the recognized entities (e.g., gene-gene interactions, biological functions). Databases such as BioGRID [2] or STRING [3] store a large collection of interactions derived from different sources and indicate which gene

*Correspondence: Peggy.Cellier@irisa.fr; Thierry.Charnois@lipn.univ-paris13.fr
[1] INSA de Rennes, IRISA, UMR6074, F-35042 Rennes, France
[2] Université de Paris 13, LIPN, UMR7030, F-93430 Villetaneuse, France
Full list of author information is available at the end of the article

interacts with a specified gene. However, these databases do not support more complex requests such as: *which genes inhibit gene X? what is the biological context (e.g., organism, biological information) associated to a gene-gene interaction? what is the kind of interaction between genes X and Y? what is the modality associated to the extracted information (related work, experimental result, etc.)?* These requests are useful for biologists since they enable to faster point out the piece of information they look for. Unfortunately, to the best of our knowledge, no work has been reported yet to support these kinds of requests. That is why in this paper we propose a method to retrieve that kind of information.

Our method automatically discovered a human manageable set of patterns that are then validated by experts to provide linguistic patterns. In other words, thanks to the linguistic patterns, our method not only allows gene interactions but also *semantics information on the extracted interactions (e.g., modalities, biological contexts, interaction types)* to be detected.

The need for linguistic resources (grammars or linguistic rules) is a common feature of the information extraction methods. Indeed, those NLP approaches apply rules such as regular expressions [4] or syntactic patterns [5,6]. However, these rules are handcrafted and thus those methods are time consuming and often devoted to a specific corpus [7].

In contrast, machine learning based methods, for example support vector machines or conditional random fields [8], are less time consuming than rule-based methods. Machine learning methods for gene interaction detection usually tackle the task as a classification problem. Best results are obtained with kernel methods [9-12] and some NLP parsers can be used to provide some features to the classifier [13]. Although they provide good results, machine learning methods still need many features. Also, their outcomes are not really understandable by a user, nor they can be used as linguistic patterns in NLP systems. Furthermore, the annotation process of training corpora requires a substantial investment of time, and cannot be reused in other domains (some new corpora must be annotated for new domains) [7]. A good trade-off is the cross-fertilization of information extraction and machine learning techniques which aims at automatically learning the linguistic rules [14,15]. However, in most cases the learning process is done from text syntactic parsing. For instance, BioContextt [16] or Turku Event Extraction System (TEES) [17] aim at extracting biological events with contextual informations (e.g., species involved, localization, modality) about the biological events. Those systems are based on a syntactic analysis. Therefore, the quality of the learned rules relies on syntactic process results. Still some works such as [18] or [19] do not use syntactic parsing.

For example, Abacha and al. [19] have a corpus based strategy close to [20] and this line of research. They aim at learning patterns from a list of seed terms corresponding to pairs of entities known to be in some target relations. Other works based on pattern matching as AliBaba [21-23], learn surface patterns using sequence alignment of sentences to derive "motifs". This method is based on a list of terms that represent interactions. Only interaction patterns are learned and no new term to symbolize interaction can be discovered. With our method, linguistic patterns are automatically learned to detect interactions (interaction patterns) and also, at the same time, to characterize the interactions (characterization patterns). In addition, the terms and the patterns do not need to be provided. They are automatically extracted by the method. It thus provides new knowledge.

The key idea of our approach is to take advantage of an hybridization of data mining and NLP for Biological Natural Language Processing (BioNLP). Data mining techniques, such as extraction of frequent sequential patterns [24], enable the discovery of implicit, previously unknown, and potentially useful information from data [25]. Our contribution is an original method to automatically produce *patterns* (which can be seen as a kind of linguistic rules) from text collections.

The problem of data mining techniques is that, in general, too many patterns are generated. That is why, our method is based on recursive sequential pattern mining with constraints from the NLP field to tackle the discovery of gene interactions. The patterns output convey a model of the interactions that are enhanced with semantics information (modalities, biological contexts, interactions types).

Only limited resource is required: the text collection (used as a training corpus) which only contains sentences with interactions and where only gene names are tagged but not the interaction. In particular, terms and patterns are automatically discovered from texts without other resources.

To the best of our knowledge, there are few methods that extract the interactions and also provide associated semantics information on the extracted interactions thanks to the discovered patterns which are understable and can be manually modified by a human expert.

In addition, we propose to use background knowledge, a well-established biomedical corpora and a gene interaction database, in order to assess the relevance of patterns, used as linguistic rules, and to help an expert[a] to select them. We describe a validation method based on the idea that the *relevant* rules convey information that must be consistent with the background knowledge. This method is interesting because the validation of rules is currently widely based on human checking which is highly time consuming. Last but not least, we conduct extensive experiments highlighting how our approach enables to discover interactions and their characterizations and we present a discussion of the results.

Method

This section presents our method to produce linguistic rules in order to discover interactions and their characterizations. Figure 1 gives a global view of the process.

Background: sequential pattern mining

Sequential pattern mining is a well-known technique introduced in [24] to find regularities in database of sequences, and for which there are several efficient algorithms (e.g. [26-30]). A *sequence* as used in our method is an ordered list $\langle i_1 \ldots i_m \rangle$, where the elements of the list $i_1 \ldots i_m$ are called items[b].

A sequence $S_1 = \langle i_1 \ldots i_n \rangle$ is *included* in a sequence $S_2 = \langle i'_1 \ldots i'_m \rangle$ if there exist integers $1 \leq j_1 < \ldots < j_n \leq m$ such that $i_1 = i'_{j_1}, \ldots, i_n = i'_{j_n}$. The sequence S_1 is called a *subsequence* of S_2, and we note $S_1 \preceq S_2$. For example, we have $\langle b\ d \rangle \preceq \langle a\ b\ c\ d \rangle$.

A sequence database *SDB* is a set of tuples (sid, S), where *sid* is a sequence identifier and S a sequence. For instance $SDB_1 = \{(1, \langle a\ b\ c\ d \rangle), (2, \langle b\ d\ e \rangle), (3, \langle a\ c\ d\ e \rangle), (4, \langle a\ d\ c\ b \rangle)\}$ is a database of four sequences.

A tuple (sid, S) *contains* a sequence S_α, if S_α is a subsequence of S. The *support* of a sequence S_α in a sequence database *SDB*, denoted $sup(S_\alpha)$ is the number of tuples in the database containing S_α. For example, in SDB_1 $sup(\langle b\ d \rangle) = 2$, since sequences 1 and 2 contain $\langle b\ d \rangle$. Notice that for notational convenience, sometimes the *relative support* is used. In this case, the support $sup(S_\alpha)$ is the relative number of tuples in the database that contain S_α, $sup(S_\alpha) = \frac{|\{(sid,S)\ |\ (sid,S) \in SDB \wedge (S_\alpha \preceq S)\}|}{|SDB|}$.

A *frequent sequential pattern* is a sequence such that its support is greater or equal to a given support threshold *minsup*.

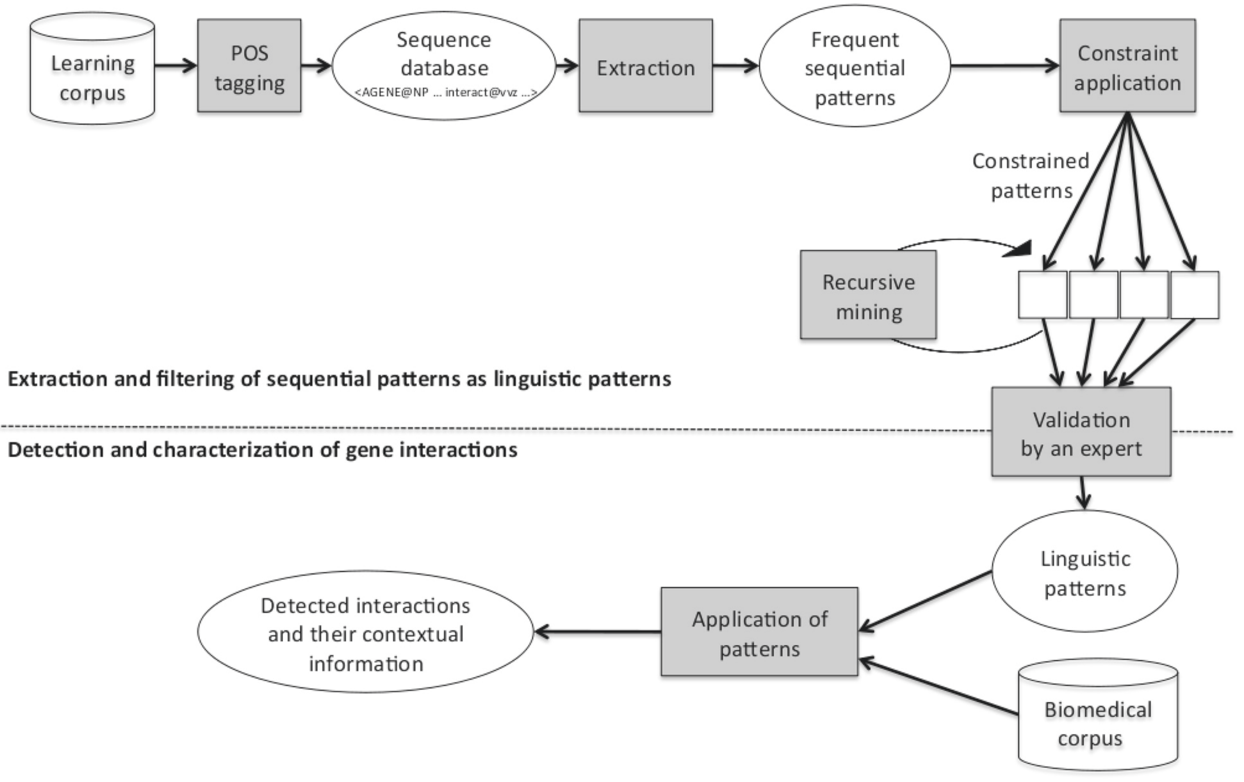
Learning corpus
POS tagging
Sequence database
<AGENE@NP ... interact@vvz ...>
Extraction
Frequent sequential patterns
Constraint application
Constrained patterns
Recursive mining
Extraction and filtering of sequential patterns as linguistic patterns
Validation by an expert
Detection and characterization of gene interactions
Linguistic patterns
Detected interactions and their contextual information
Application of patterns
Biomedical corpus

Figure 1 General framework to extract gene interactions. Figure 1 presents the overall process to detect and characterize gene interactions. There are two steps. The first step is the extraction of patterns. Sequential patterns are mined from a learning corpus that contains sentences representing gene interactions. In order to reduce the number of extracted patterns, constraints and recursive mining are applied. At the end, few sequential patterns corresponding to candidate linguistic interaction or characterization rules remain. A key point is that the sequence of words expressing the interaction in a pattern is automatically discovered. As an example, the sequence of words <*AGENE*, "association"> in the pattern "*AGENE* association with *AGENE*" (see Table 5) where a small gap may appear between *AGENE* and "association" are discovered by our method. This pattern conveys an interaction between two genes (denoted AGENE) in association. The sequential patterns are then analyzed and validated by experts. The ones that are not relevant for interaction detection or characterization are removed. The second step is the application of those validated patterns as linguistic rules to discover and characterize interactions in new corpora.

Extraction of sequential patterns in texts

For the extraction of sequential patterns from biological texts, we use a training corpus which is a set of sentences that contain interactions (but not annotated) and where the genes are identified. In this paper we consider sentences containing interactions and at least two gene names to avoid problems introduced by the anaphoric structures[c] [31]. The training corpus, with tagged gene names, is selected by an expert. The items are combinations of lemma and POS tags[d]. POS tag information is important to disambiguate words (e.g., "*form*" the noun vs "*to form*" the verb). The sequences of the database are the interaction sentences where each word is replaced by the corresponding item. The order relation between items in a sequence is the order of words within the sentence. For example, let us consider two sentences that contain gene interactions: *"Recent studies have suggested that c-myc may be vital for regulation of hTERT mRNA expression and telomerase activity".* and *"Injection of frpHE mRNA in Xenopus embryos inhibited the Wnt-8 mediated dorsal axis duplication".* All gene names are replaced by a specific item, *AGENE*, and the other words are replaced by the combinations of their lemma and their POS tag. An excerpt of the database that contains the sequences associated to those two sentences is given in Table 1. The sequential patterns are extracted from this database.

The choice of a support threshold *minsup* is a well-known problem in data mining. With a high *minsup*, only few very general patterns can be extracted. With a low *minsup*, a lot of patterns can be found. Some interesting words, for example "interaction", are not very frequent so that we set a low value of *minsup*. As a consequence, a huge set of patterns is discovered and it needs to be filtered in order to return only relevant patterns.

Constraints and recursive mining

To reduce the number of extracted patterns, we use a combination of data mining methods. The constraint-based pattern paradigm (e.g., [32]) enables discovering patterns under user-defined constraints in order to drive the mining process towards the user objectives. Recursive mining [33] reduces the number of patterns by extracting their common structures.

Linguistic constraints

In pattern mining, constraints allow the user to define more precisely what should be considered as interesting. The most commonly used constraint is the constraint of frequency (*minsup*). However, it is possible to use different constraints [34]. In our method, in order to extract gene interaction patterns, we use three additional constraints.

The first constraint is that the pattern must contain two gene names, i.e. two *AGENE* items.

The second constraint is that the pattern must contain at least a verb or a noun.

Finally, among the patterns that satisfy the frequency and the two other previous constraints, we retain only the maximal ones with respect to the inclusion order $\preceq$. That last constraint allows the redundancy between patterns to be reduced.

The constraints can be gathered in only one constraint $\mathcal{C}_G$ which is the conjunction of the three constraints. $SAT(C_G)$ is the set of patterns satisfying $\mathcal{C}_G$.

Recursive mining

Even if the new set of sequential patterns, $SAT(\mathcal{C}_G)$, is significantly smaller than the initial set of all extracted sequential patterns without constraints, it can still be too large to be analyzed and validated by experts. To find a limited number of patterns corresponding to general structures among the whole pattern collection, we use the *recursive mining* technique of [33]. The key idea of this post-processing is to reduce the size of the output by successively repeating the mining process on the patterns themselves in order to extract the structure shared by the patterns. More precisely, at each step, the previous set of sequential patterns is used as a new sequential database, and a new extraction is made. The process stops when no

Table 1 Example of a sequence database

ID	Sequence
...	...
S1	*⟨Recent@jjstudy@nnshave@vhpsuggest@vvnthat@in / thatAGENEmay@mdbe@vbvital@jj for@in regulation@nn of@in AGENE mrna@np expression@nn and@cc telomerase@nn activity@nn .@sent ⟩*
S2	*⟨injection@nnof @inAGENEmrna@npin@inxenopus@npembryo@nnsinhibit@vvdthe@dt AGENE mediate@vvd dorsal@jj axis@nn duplication@nn .@sent ⟩*
...	...

Table 1 shows an excerpt of a sequence database which contains two interaction sentences:
S1: *"Recent studies have suggested that c-myc may be vital for regulation of hTERT mRNA expression and telomerase activity."* and
S2: *"Injection of frpHE mRNA in Xenopus embryos inhibited the Wnt-8 mediated dorsal axis duplication.".*

more than k patterns are obtained by the extraction, where k is a parameter set by the user.

Our target is to identify at least one pattern by verb or noun that appears in the patterns in $SAT(\mathcal{C}_G)$. So, for each verb or noun denoted X_i, that appears in $SAT(\mathcal{C}_G)$, we collect the set E_{X_i} of patterns containing X_i, $E_{X_i} = \{s \in SAT(\mathcal{C}_G) \mid \langle X_i \rangle \preceq s\}$. Note that some frequent patterns can contain more than one noun and/or verb (so several X_i). In this case, the pattern is duplicated in the E_{X_i} of each noun and/or verb.

For a given value of k, we apply the recursive mining post-processing technique on each E_{X_i}. At each extraction step we select only the patterns that satisfy $\mathcal{C}_G$, and use a relative minimum support threshold $minsup = \frac{1}{k}$. That threshold value and the maximality constraint guarantee that recursive mining process terminates in finite steps as proved in [35].

At the end of this post-processing of all E_{X_i}, the number of sequential patterns cannot exceed $n \times k$ where n is the number of verbs and nouns occurring in $SAT(\mathcal{C}_G)$.

Selection and categorization of patterns

The sequential patterns are then analyzed and validated by experts. The ones that are not relevant for interaction detection or characterization are removed. The remaining ones are selected as linguistic extraction rules [36]. A selected pattern is classified with respect to the kind of information conveyed by the pattern. There are two main classes of patterns: *interaction patterns* and *characterization patterns*. The first class indicates what kind of interaction between genes is found (e.g., inhibition). The second class is *characterization patterns*. It is built by the experts and can be completed with other classes if other kinds of information extraction rules are found. There are two kinds of *characterization patterns*: *modality patterns* and *biological context patterns*. Examples of patterns are discussed in Section “Extracted sequential patterns”.

When the experts have validated and classified all patterns in the different categories, they are applied as linguistic rules to discover and characterize interactions in new corpora.

In practice, this step is not time consuming as shown in the following and can be helped by using background knowledge to support pattern validation as proposed in the Section “About validation of sequential patterns as linguistic extraction rules”. Detection with sequential patterns representing interactions, modalities or biological contexts is much more elaborated than just a cooccurrence detection. Indeed, the order of the words and the context are important, they provide semantics information. For instance, the sub-categorization of the verb given by the POS tagging indicates the passive or active verb and identifies the direction of the interaction. Prepositions can give this information when the pattern does not contain a verb, for example: $\langle$*activation@nn of@in AGENE by@in AGENE*$\rangle$.

Note the genericity of the approach, indeed the extracted patterns allow genetic interactions to be discovered as well as physical protein interactions.

Results

In this section, we present the experiments and results. First, the training corpus is detailled. Then, the sequential pattern extraction is described. Finally, the results of the application of the extracted patterns on testing corpora are presented.

Training corpus

Genes can interact with each other through the proteins they synthesize. Moreover, although there are conventions, the same word can represent a gene name and the protein synthesized by the gene. Biologists know from the context if the sentence is about protein or gene. To discover the linguistic patterns of interactions between genes, we merge two different corpora containing genes and proteins, to create the training corpus. The first corpus contains sentences from PubMed abstracts, selected by Christine Brun[e] as sentences containing gene interactions. It contains 1,806 sentences. That corpus is available as a secondary data source for the learning tasks “Protein-Protein Interaction Task (Interaction Award Sub-task, ISS)” from BioCreAtIvE Challenge II [8]. The second corpus [37] contains 2,995 sentences mentionning interactions between genes selected by an expert. The union of those two corpora results in a dataset containing 4,801 sentences about gene interactions.

Sequential pattern extraction

Data mining task

As previously mentionned, the extraction of sequential patterns from the training corpus needs the computation of POS tags. For this task, we use the *treetagger* tool [38].

In addition, for the data mining task, *minsup* is set to 10. It means that a sequential pattern is frequent if it appears in at least 10 sentences (i.e. in more than 0.2% of sentences). Indeed, with that threshold some irrelevant patterns are not taken into account while many patterns of true gene interactions are discovered. Note that other experiments, not reported here, have been conducted with greater *minsup* values (15 and 20). With those greater *minsup*, some relevant patterns for interaction detection are lost.

More than 32 million of frequent sequential patterns are extracted, with *minsup* equals to 10. This number is large but the extraction takes only 15 minutes (the extraction tool is *dmt4sp* [39]). The application of constraints significantly reduces the number of sequential patterns. Indeed, the number of sequential patterns satisfying the constraints is about 65,000. Note that the application of the constraints was not time consuming and takes less than two minutes. However, the number of remaining patterns is still prohibitive for analysis and validation by human experts.

The recursive mining also reduces significantly the number of sequential patterns. From the extracted patterns, we build a subset of patterns for each noun or verb. The number of built subsets is 515 (365 for nouns, 150 for verbs). The recursive mining of each subset exhibits at most *k* sequential patterns to represent that subset. In this experiment, we set the parameter *k* to 4. It allows several patterns to be kept for each noun or verb in order to cover a sufficient number of different cases (for example 4 patterns corresponding to 4 syntactic constructions with the verb *inhibit@vvn* are computed). At the end of the recursive mining, there remain 667 sequential patterns that can represent interactions or their characterizations[f]. That number, which is significantly smaller than the previous one, guarantees the feasibility of an analysis of those patterns by experts. The recursive mining of those subsets is also very fast and takes about 2 minutes.

Extracted sequential patterns

The 667 remaining sequential patterns were analyzed by two experts in 90 minutes.

The patterns are grouped together by noun or verb, the experts have thus to classify 380 groups. But some nouns or verbs are repeated with different POS tagged information (e.g., *analyze@vvd* and *analyze@vvn*), so these groups are not considered independently by the experts and it helps the validation task (for instance both versions of verb "analyze"İ are pruned together). Actually, there are 285 different nouns and verbs. Moreover, at this point of the validation, the patterns are roughly split into three sets by the experts: "interaction patterns", "characterization patterns"İ and "not relevant".

Finally, the experts validated 232 sequential patterns for interaction detection, 231 patterns for characterization of interactions and they removed the remaining (i.e. 204 unuseful patterns). Indeed, the latter do not convey information about interactions, in particular there are generic verbs like "appear"İ and "contain". Among the first group of 232 patterns, some explicitly give the type of the interactions. For example, ⟨*AGENE interact@vvz with@in AGENE*⟩, ⟨*AGENE bind@vvz to@to AGENE*⟩, ⟨*AGENE deplete@vvn AGENE*⟩ and ⟨*activation@nn of@in AGENE by@in AGENE*⟩ describe well-known interactions (binding, inhibition, activation). Note that when the patterns are applied, zero or several words may appear between two consecutive items of the pattern. For example, the pattern ⟨*AGENE interact@vvz with@in AGENE*⟩ matches the sentence "<gene_name=MYC> interacts with <gene_name=STAT3>". and also the sentence "<gene_name=MYC> interacts with genes in particular <gene_name=STAT3>"[g].

Other patterns represent more general interactions and express the fact that a gene plays a role in an activity of another one. Representative patterns of this kind are for instance ⟨ *AGENE involve@vvn in@in AGENE*⟩, ⟨*AGENE play@vvz role@nn in@in the@dt AGENE*⟩ and ⟨*AGENE play@vvz role@nn in@in of@in AGENE*⟩. Note that the *"involve"* verb and the *"play role in"* phrase were not reported in [40,41] and [21].

The second group of 231 patterns for characterization represents other kinds of semantics information: modalities or biological context, for instance, ⟨*in@in fibroblast@nns AGENE AGENE*⟩ or ⟨*the@dt possibility@nn that@in/that AGENE AGENE*⟩. Figure 2 depicts the taxonomy that we define and use in our experiments for the characterization patterns. That taxonomy was built with the help of the extracted patterns. The *modality patterns* express the confidence in the detected interactions. Modality can be seen as a kind of uncertainty [42]. We define four levels of confidence: *Assumption*, *Observation*, *Demonstration* and *Related work*, and another subclass representing the *Negation* (patterns denoting evidence of absence of interaction). For example, the sentence "It suggests that <gene_name=MYC> interacts with <gene_name=STAT3>" has a lower confidence than "It was demonstrated that <gene_name=MYC> interacts with <gene_name =STAT3>". The *biological context patterns* indicate information about the biological context of interactions, for example the disease or the organism involved in the interaction. That class is split into four subclasses: *organism*, *component*, *biological situation* and *biological relation*. The subclass *organism* represents the organisms involved in the interaction (e.g.,

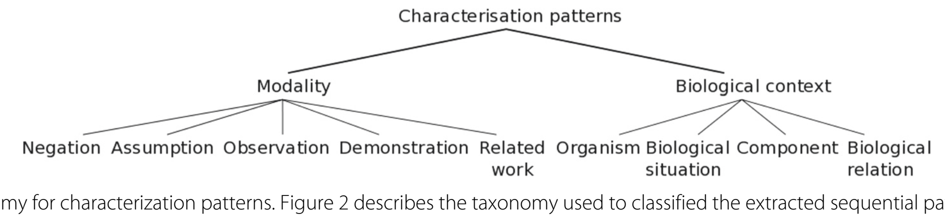

Figure 2 Taxonomy for characterization patterns. Figure 2 describes the taxonomy used to classified the extracted sequential patterns.

"mouse", "human"). The subclass *component* represents the anatomy/biological components (e.g. "breast" or "fibroblast"). The subclass *biological situation* gives the framework of interactions, for example, "cancer", "tumor" or "in vitro". The last subclass gives, when applicable, the *biological relation* (e.g., "homology").

The sequential patterns obtained are linguistic extraction rules that can be used on biomedical texts to detect and characterize interactions between genes. Note that to be applied, those patterns do not need a full syntactic analysis of a sentence.

Indeed, the matching process tries to instantiate each element of the pattern in the given sentence. For each pattern, every possible matching within the sentence is tested and not only the first one.

Application: detection and characterization of gene interactions

We have evaluated the quality of the sequential patterns found in the previous section as information extraction rules. In this section, we present the experimental settings and the results.

Testing corpora and evaluation criteria

Testing Corpora We have considered three well-known testing datasets (cf Table 2 and Table 3): *AIMed* [43], *BioInfer* [44], *HPRD50* [45] and a fourth testing corpus extracted from PubMed [1] (more information is given in the next section). Note that, in *AIMed*, *BioInfer* and *HPRD50*, the names of genes are already identified and tagged. More information about those corpora can be found in [46].

Construction of the PubMed corpus In order to test the sequential patterns extracted in the previous section as linguistic extraction rules to characterize interactions, we need a testing corpus.

We have built a testing corpus that is a subset of abstracts from the PubMed database. It is built in two steps which are described below. The first step is the selection of abstracts from PubMed. In the PubMed database, each paper has an identifier called PMID (PubMed IDentifier). For each official acronym of gene in the HUGO [47] dictionary, a request is sent to PubMed in order to get all PMID of papers that contain the gene. An index of genes and their associated papers is thus created. Then the inverted index is computed, i.e. the index that associates to each PMID the list of genes. From that second index, the PMIDs that do not have at least two gene names in their list are pruned. Indeed, as we are looking for interactions between genes, it implies that at least two genes are mentioned in the text. There remains 624,519 PMIDs. The second step is the named-entity recognition. Sometimes, the gene name used to index an abstract and the gene name that appears in the abstract text are different. Indeed, a gene can be represented by different synonymous forms. It is thus important to identify the gene in the text; that task is called Named-Entity Recognition (NER). We propose to use a "dictionary-based" approach [48]. Although that kind of approaches usually has a good precision, it does not provide a good recall. We propose some improvements to increase the recall.

Table 2 Results of the application of the extracted patterns

Corpus	# Sentences	Recall	Precision	*f − score*	*f − score* presented in [11] details given in Table 3
AIMed	1955	78.6	35.6	**49**	[34.7, 41.5]
BioInfer	1100	46.5	25.3	32.8	[15.9, 40.6]
PubMed	200	75.0	83.0	78.7	–
HPRD50	145	66.8	46.7	55.0	[38.3, 69.8]

Table 2 gives the list of the four testing corpora used to evaluate the proposed approach, and the results of the evaluation. The meaning of the columns is: the name of the corpus, the number of the sentences in the corpus, the recall score of the proposed approach applied on the corpus, the precision score of the proposed approach applied on the corpus, the*f-score* of the proposed approach applied on the corpus. The last column indicates the range of the *f-scores* presented in [11] with also a cross-corpus validation.

First, all genes associated to the PMID of an abstract are searched into that abstract using official acronyms from the HUGO dictionary. With that approach only 48.1% of abstracts have at least two recognized genes. In addition, we identified 182 official acronyms as common English words (e.g., AGO, AS, BAD)[h]. In order to reduce the number of mistakes, they are considered as gene names only when they are in uppercase.

Second, in order to improve the number of recognized genes in abstracts, other fields of the HUGO dictionary are used: old acronyms, alias acronyms, and complete names. With that improvement, 61.7% of abstracts have at least two recognized genes. Note that this improvement is mainly due to the alias acronyms (+ 9%).

The last improvement is the use of significant parts of the complete official name. The official name is often long, and authors do not write it completely. Instead of looking for the complete name, we look for significant parts of it. We identify three common kinds of significant parts: a word ending by "in" (e.g., "insulin"), a word ending by "ase" (e.g., "transferase") and a word followed by "protein" (e.g., "AE binding protein 1"). For instance, for gene "alkaline ceramidase 3", the significant part is "ceramidase" and thus to recognize this gene name in texts, only "ceramidase" would be used. The plural forms are also taken into account (e.g., "caspases", "kinases"). With that improvement, 66.1% of the 624,519 abstracts have at least two recognized genes and form the testing corpus.

Evaluation criteria for the extraction of gene interactions and their contextual information We evaluate

Table 3 Details of information presented in paper [11]

Method	SL	SL	SpT	SpT	kBSPS	kBSPS	edit	edit	APG	APG
Training corpus	**(AIMed)**	**(BioInfer)**	**(AIMed)**	**(BioInfer)**	**(AIMed)**	**(BioInfer)**	**(AIMed)**	**(BioInfer)**	**(AIMed)**	**(BioInfer)**
AIMed	-	41.5	-	34.7	-	40.3	-	39.6	-	37.9
BioInfer	40.6	-	24.3	-	24.8	-	15.9	-	22.5	-
HPRD50	59.0	61.8	43.2	51.3	51.0	69.8	38.3	62.4	61.6	62.1

The acronyms used in this table are the ones used in paper [11]: SL: Shallow linguistic kernel; SpT: Spectrum tree kernel; kBSPS: k-band shortest path spectrum kernel; edit: Edit distance kernel; APG: All-paths graph kernel. See paper [11] for more details.

our approach with a cross-corpus evaluation to show the genericity of the proposed approach. It means that we extract the patterns from a corpus and apply them on the other four corpora [11]. Note that in the literature many approaches are evaluated with a cross-validation, which means that a corpus is split in several parts, one part is used to learn and the rest is used to apply.

It is thus much more difficult to get good results with a cross-corpus evaluation than a cross-validation.

Indeed, there is more heterogeneity between corpora (i.e., corpus characteristics are different) than between parts of a single corpus [11].

We use the f-*score* function as an evaluation measure, which is defined as $f\text{-}score = \frac{2\times Precision\times Recall}{Precision+Recall}$.

Detection of gene interactions

We have applied the 232 extracted sequential patterns as linguistic extraction rules to detect interactions on the four corpora.

All corpora used for evaluation have all gene names readily tagged. This means that our results only measure the performance of gene interaction extraction and are not influenced by the issue of named entity recognition. Therefore, to compute the f-score, a true positive is a couple of mentioned gene names in the sentence (i.e. the gene names given in the tags) which are in interaction and detected as an interaction by our method. Table 2 gives the results. We did not have any gold standard reference to evaluate the results for the testing corpus from PubMed. Since we cannot implement an automatic validation, we randomly took 200 sentences among the sentences of the PubMed testing corpus. Then, we carried out a POS tagging and assessed the performances of the extraction rules to detect interactions in the 200 sentences[i]. The f-*score* for the gene interaction detection for the testing corpus is 78.7. In Table 2, the last column indicates the range of the f-*scores* presented in [11] with a cross-corpus validation. Several kernel-based approaches are presented in [11], the range allows to show the worst and the best results among all those methods. Note that the best result of the ranges is not achieved in practice by the same method. Our approach gives results comparable to the results given by state-of-the-art methods and it is even better for the gene interaction detection in AIMed. This last result is important because AIMed is the largest corpus and the most commonly used in the literature. Moreover, our approach is simple and allows more information that just the presence of an interaction to be extracted. Indeed, thanks to the patterns, semantics information can also be extracted, contextual information (see next section) but also information about the kind of interaction (e.g., inhibition, binding) and the direction of the interaction.

Characterization of gene interaction

The method also gives information about modality and about the biological context: biological situation, component, organism, biological relation. For that characterization task, there exist some methods dealing with the subtask of the detection of sentences containing uncertainty [42] (modality can be seen as a kind of uncertainty) but few adress the biological characterization problem. It was thus difficult to compare our result for the interaction characterization with a gold standard. We randomly took 200 sentences containing at least two gene names among the sentences of the testing corpus extracted from PubMed. Those sentences are not the same ones that are used to evaluate the interaction detection but they come from the same testing corpus PubMed. Out of 200 interactions, there are 149 characterizations (71 modalities and 78 biological context). The sentences have been annotated by a computer scientist with specialisation in NLP and a biologist. Then, we evaluated the precision and recall. The characterization patterns are applied on a pair of genes that is already detected as in interaction. We evaluate the characterization at the interaction level. The precision is 88% and the recall is 69% (f-*score*= 77). Several reasons explain why the recall is not greater and are discussed in the next section.

Discussion

In this section, the results of the previous section are discussed from a qualitative point of view and we present a process to support pattern validation.

About interaction detection

In the experiments a linguistic pattern is matched against a whole sentence at a time. That wide scope may introduce ambiguities in the detection of interactions, and false positives, when more than two genes appear in a sentence. For example, in sentence *"FGF-7 recognizes one FGFR isoform known as the FGFR2 IIIb isoform or keratinocyte growth factor receptor (KGFR), whereas FGF-2 binds well to FGFR1, FGFR2, and FGFR4 but interacts poorly with KGFR".* an interaction between FGFR2 IIIb and FGFR1 is detected. Actually, there is no interaction between those two genes, they only appear in two different propositions of the same sentence. FGFR1 interacts with FGF-2 in the second proposition but since there is no limitation of the scope, an interaction between FGFR1 and FGFR2 IIIb is also detected. Several cases are possible: when several binary interactions are present in the sentence or when the interaction is n-ary ($n \geq 3$). The case of n-ary interactions can be solved with a training data set containing n-ary interactions. The other cases can be treated by introducing limitations of pattern scope, for example cue-phrases (e.g., *but*, *however*).

False negatives depend on the absence of some nouns or verbs of interaction in the patterns. For example, the noun "modulation" is not discovered whereas the verb "modulate" appears in sequential patterns. This suggests that the use of linguistic resources (e.g. lexicon or dictionary), manually or semi-automatically, would improve interaction patterns and thus interaction detection.

About interaction characterization

The false negatives, which are dependent on the absence of some patterns, are also an important problem for interaction characterization.

For example, in our experiments in the sentence *"<gene_name=BRCA1> interacts in vivo and in vitro with the Rb-binding proteins, <gene_name=RBBP7> and <gene_name=RBBP4>[...]"* the biological situation "in vitro" is detected whereas "in vivo" is not detected. Indeed, there is no sequential pattern extracted from the training corpus that contains "in vivo". That case is considered as true positive for in vitro interaction and as false negative for in vivo interaction. The recall (69%) is strongly dependent on the number of false negatives. Note that the false negatives mainly come from biological contexts not sufficiently represented (about 92%). It is explained by the difficulty to have a training corpus that contains all biological context (e.g, body parts as "liver", "pituitary gland", diseases). As for interaction detection, using a specialized lexicon would increase the vocabulary and thus the number of patterns and would improve those results.

About validation of sequential patterns as linguistic extraction rules

Section "Method" shows how the sequential patterns are automatically extracted from a corpus. Those patterns are then analyzed and validated by two experts as linguistic extraction rules. But sometimes, the needed resources (e.g., time, expert) can be missing or the number of sequential patterns can be too large to be easily managed by a human. In those cases, for the selection and validation of patterns, we propose an automatic process based on the use of background knowledge. The selection is thus less accurate than a manual selection but can be automatic.

The automatic validation process is based on two steps.

First, each sequential pattern is applied on a corpus called *rule validation corpus*. It provides for each pattern the following information: the genes detected as interacting and the associated sentences.

Second, a gene interaction database is used as an *oracle* to assess the patterns. In our method, the rule validation corpus comes from the PubMed papers and the gene interaction database is BioGRID. Our idea is that the relevant patterns, when applied on the validation corpus, retrieve interactions that must be consistent with the gene interaction database. An interaction detected by a sequential pattern is considered as a false positive if the interaction does not exist in the gene interaction database, else it is a true positive (same gene names and same PMID)[j].

A pattern with a high number of true positives is likely to be interesting.

Table 4 gives an excerpt of the information provided for each pattern. It contains the number of interactions

Table 4 Examples of information about the application of information extraction rules

Information extraction rule	Number of retrieved interactions	Number of true positives
AGENE AGENE the@dt response@nn	6	1
AGENE AGENE serine@nn	3	3
AGENE reveal@vvd *AGENE*	0	undefined
AGENE association@nn with@in *AGENE*	6	4
AGENE bind@vvz to@to *AGENE*	8	5

Table 4 gives an excerpt of provided information about patterns extracted from the PubMed corpus. The meaning of the columns is: sequential pattern, number of interactions detected by the pattern and number of detected interactions that are correct with respect to the oracle, i.e. interactions that also exist in BioGRID. The first pattern can be read as "a *gene* followed by a *gene* then by the word *the* and the word *response*". This pattern detects 6 interactions and 1 is in BioGRID. The second pattern can be read as "a *gene* followed by a *gene* then by the word *serine*". It detects 3 interactions that are all in BioGRID. The third pattern can be read as "a *gene* followed by the verb *reveal* in past tense, then by a *gene*". This pattern does not detect interactions in the rule validation corpus, thus no information is provided to evaluate it. The fourth pattern can be read as "a *gene* followed by the noun *association*, then by the word *with* and a gene name". It detects 6 interactions out of which 4 are in BioGRID. The fifth pattern can be read as "a *gene* followed by the verb *bind* in present tense, then by the word *to* and a gene name". This pattern detects 8 interactions and 5 of them are in BioGRID. For example, it detects that the following complex sentence "Cbl is a cytosolic protein that is rapidly tyrosine phosphorylated in response to Fc receptor activation and binds to the adaptor proteins Grb2, CrkL, and Nck." contains an association between two signalling molecules (Cbl and Grb2).

detected by the pattern and the number of detected interactions that are correct with respect to the oracle, i.e. interactions that also exist in BioGRID. For example, the fifth pattern can be read as "a *gene* followed by the verb *bind* in present tense, then by the word *to* and a gene name". This pattern detects 8 interactions and 5 of them are in BioGRID. For example, it detects that the following complex sentence "Cbl is a cytosolic protein that is rapidly tyrosine phosphorylated in response to Fc receptor activation and binds to the adaptor proteins Grb2, CrkL, and Nck". contains an association between two signalling molecules (CBL and GRB2).

Those measures can be used to automatically select patterns as linguistic information extraction rules.

To end up with a more speculative note, this step could also be interesting even when there is an expert to select the patterns by providing more information to help them. In addition, a pattern with low number of true positives can retrieve sentences that really contain interactions. This can be the case if the interaction is not reported in BioGRID or is reported but the gene names in the sentence are other gene names than the ones used in BioGRID. Therefore, it is interesting to provide, for each pattern, the detected interaction sentences. Table 5 gives an excerpt of interactions detected by the sequential pattern: "*AGENE* association with *AGENE*". For instance, the pattern detects that in the paper with PMID 10204582, an interaction between genes SHC1 and CRKL is mentioned (the pattern matches a sentence of the abstract) but according to BioGRID, there is no interaction between SHC1 and CRKL. The discovered interaction in the sentences in paper 10204582 is thus unexpected in BioGRID. The pattern also detects that in this paper, an interaction between genes CBL and CRKL is mentioned, and indeed, according to BioGRID, there is an interaction between CBL and CRKL mentioned in paper 10204582. It is interesting to note that the three genes involved harbour all three similar biological functions (they are all signalling molecules) and that their association is fully relevant as exemplified by the strong functional connectivity detected in the STRING database between those three genes. Therefore the extracted interaction between genes SHC1 and CRKL is fully relevant even if it does not appear in BioGRID. Of course, more systematic studies should be undertaken to ascertain this, but this is beyond the scope of the present paper.

Conclusions

We have proposed an original approach to help experts to design linguistic information extraction rules by automatically extract sequential patterns filtered by linguistic constraints and recursive mining. Unlike existing methods, our approach is independent of syntactic parsing and only requires the training corpus as external resource to learn patterns (note that interaction clues are not annotated in the training corpus). The patterns representing interactions and their characterizations are automatically discovered from texts. An advantage of the use of sequential patterns as linguistic rules is that they are understandable and manageable by an expert. If needed, the expert can easily modify the proposed rules or add new ones. To the best of our knowledge, there are few methods that automatically extract interaction patterns and also characterization patterns (i.e., patterns for contextual information about the discovered interactions).

Table 5 Example of information for a sequential pattern

PMID	Gene 1	Gene 2	BioGRID verdict	Sentence
10204582	SHC1	CRKL	not in BioGRID	These results suggest a fundamental role for the tyrosine phosphorylation of Cbl, CrkL, SLP-76, and <**gene_name="SHC1"**> and the **association** of Cbl **with** <**gene_name="CRKL"**>, SLP-76, and Nck in Fc gammaRI signaling in human macrophages.
10204582	CBL	CRKL	in BioGRID	PP1, a specific inhibitor of Src kinases, inhibited the Fc gammaRI-induced respiratory burst, as well as the tyrosine phosphorylation of <**gene_name="CBL"**> and its inducible **association with** <**gene_name="CRKL"**>.

Table 5 gives 2 interactions (highlighted in bold in the table) detected by the sequential pattern: "*AGENE* association with *AGENE*".
The meaning of the columns is: the id number of the paper in PubMed, the genes that interact, the verdict of the oracle and the sentence where the interaction is recognized. For instance, the pattern detects that in paper 10204582, an interaction between genes SHC1 and CRKL is mentioned, because the pattern matches a sentence of the abstract of the paper, but according to BioGRID, there is no interaction between SHC1 and CRKL and the discovered interaction in the sentences in paper 10204582 is unexpected because not in BioGRID and interesting. The pattern also detects that in this paper, an interaction between genes CBL and CRKL is mentioned, and indeed, according to BioGRID, there is an interaction between CBL and CRKL mentioned in paper 10204582.

The experiments show how our approach enables to discover interactions and their characterizations. Our approach gives results comparable to the results given by state-of-the-art methods and is even better for the gene interaction detection in AIMed. The main advantages of our approach are that, first, semantics information are extracted in addition to the information about the presence of an interaction; second, the patterns, used as extraction linguistic rules, are automatically discovered. Further work will look how enhance the extracted patterns thanks to other information sources (e.g., specialized dictionaries).

Availability of software and supporting data

The extracted gene interactions from PubMed are available at https://bingotexte.greyc.fr/. The evaluation corpora from PubMed are available at https://cremilleux.users.greyc.fr/jbms/. The software (SMBio) that allows sequential patterns of gene interactions to be extracted is available at https://bingo2.greyc.fr/?q=node/22. The list of the 182 official acronyms identified as common English words is available at: https://bingotexte.greyc.fr/ambig_names.

Endnotes

[a] In the rest of the paper, the term ”expert” is used for a linguist or a biologist; both skills are useful to validate rules.

[b] Notice that this is a simplified form of sequences, while in the general sequential pattern mining framework, a sequence is a list of *sets* of items, and not only a list of items.

[c] An anaphoric structure is the use of a linguistic unit, such as a pronoun, to refer back to a gene name.

[d] POS (Part-Of-Speech) tags are grammatical information about words. For example, *nn* means common noun and *vvp* means verb in non-3rd personal singular present. The exhaustive list of POS tags can be found at: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

[e] *Institut de Biologie du Développement de Marseille-Luminy*.

[f] The maximum number of patterns per verb and noun is 4, thus the maximum number of patterns after applying recursive mining is 2060. The number of patterns in the results is only 667 because some verbs and nouns are represented by less than 4 patterns.

[g] Note that such a functional relationship between MYC and STAT-3 can be illustrated by the fact that the expression of the c-myc gene is under the control of the STAT-3 signalling pathway (see [49] for a review).

[h] The list is available at: https://bingotexte.greyc.fr/ambig_names.

[i] Discovered interactions for the whole testing corpus are available at https://bingotexte.greyc.fr/.

[j] BioGRID provides information about gene interactions and the PMID of the articles where the interactions are mentioned. In order to measure the accuracy of the patterns, we take into account these two pieces of information.

Abbreviations

NLP: Natural language processing; NER: Named-entity recognition; IE: Information exctraction; BioNLP: Biological natural language processing; POS: Part-Of-Speech; minsup: support threshold; PMID: PubMed IDentifier.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PC, TC, MP, CR and BC developped the interaction detection and characterization method, and the validation step. Experiments were mainly conducted by MP, PC and TC. JK helped to design the evaluation. JLM and TC developped the named-entity recognition method. OG provided the biological expertise. JLM built the website interface. All authors read and approved the manuscript.

Acknowledgements

This work is partly supported by the ANR (French Research National Agency) funded project Hybride ANR-11-BS002-002 and progam "Investissements d'Avenir" ANR-10-LABX-0083.

Author details

[1] INSA de Rennes, IRISA, UMR6074, F-35042 Rennes, France. [2] Université de Paris 13, LIPN, UMR7030, F-93430 Villetaneuse, France. [3] Université Lyon 1, LIRIS, UMR5205, F-69622 Lyon, France. [4] INSA de Lyon, LIRIS, UMR5205, F-69621 Lyon, France. [5] Université de Caen, GREYC, UMR6072, F-14032 Caen, France. [6] Université Lyon 1, CGMC, UMR5534, F-69622 Lyon, France. [7] Faculty of Electrical Engineering, Czech Technical University, Prague, Czech Republic.

Received: 25 July 2013 Accepted: 22 April 2015
Published online: 18 May 2015

References

1. PubMed. http://www.ncbi.nlm.nih.gov/pubmed/.
2. BioGRID. http://thebiogrid.org/.
3. STRING. http://string-db.org/.
4. Giuliano C, Lavelli A, Romano L. Exploiting shallow linguistic information for relation extraction from biomedical literature. In: Conference of the European Chapter of the Association for Computational Linguistics. Trento, Italy; 2006. p. 401–8.
5. Rinaldi F, Schneider G, Kaljurand K, Hess M, Romacker M. An environment for relation mining over richly annotated corpora: the case of genia. BMC Bioinformatics. 2006;7(Suppl 3):S3.
6. Fundel K, Küffner R, Zimmer R. RelEx - relation extraction using dependency parse trees. Bioinformatics. 2007;23(3):365–71.
7. Hobbs JR, Riloff E. Information extraction In: Indurkhya N, Damerau FJ, editors. Handbook of Natural Language Processing, Second Edition. Boca Raton, FL: CRC; 2010.
8. Krallinger M, Leitner F, Rodriguez-Penagos C, Valencia A. Overview of the protein-protein interaction annotation extraction task of BioCreative II. Genome Biol. 2008;9(Suppl 2):S4.
9. Zhang Y, Lin H, Yang Z, Li Y. Neighborhood hash graph kernel for protein-protein interaction extraction. J Biomed Inform. 2011;44(6):1086–92.
10. Polajnar T, Damoulas T, Girolami M. Protein interaction sentence detection using multiple semantic kernels. J Biomed Semantics. 2011;2:1.
11. Tikk D, Thomas PE, Palaga P, Hakenberg J, Leser U. A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. PLoS Comput Biol. 2010;6(7):1–19.

12. Tikk D, Solt I, Thomas PE, Leser U. A detailed error analysis of 13 kernel methods for protein-protein interaction extraction. BMC Bioinformatics. 2013;14:12.
13. Miyao Y, Sagae K, Sætre R, Matsuzaki T, Tsujii J. Evaluating contributions of natural language parsers to protein-protein interaction extraction. Bioinformatics. 2009;25(3):394–400.
14. Nédellec C. Machine learning for information extraction in genomics - state of the art and perspectives. In: Text Mining and Its Applications: Results of the NEMIS Launch Conference. Studies in Fuzziness and Soft Computing. Berlin Heidelberg: Springer; 2004. p. 99–118.
15. Schneider G, Kaljurand K, Rinaldi F. Detecting protein-protein interactions in biomedical texts using a parser and linguistic resources. In: International Conference on Intelligent Text Processing and Computational Linguistics. LNCS, vol. 5449. Berlin, Germany: Springer; 2009. p. 406–17.
16. Gerner M, Sarafraz F, Bergman CM, Nenadic G. Biocontext: an integrated text mining system for large-scale extraction and contextualization of biomolecular events. Bioinformatics. 2012;28(16):2154–61.
17. Björne J, Ginter F, Pyysalo S, Tsujii J, Salakoski T. Scaling up biomedical event extraction to the entire pubmed. In: Proceedings of the 2010 Workshop on Biomedical Natural Language Processing. Uppsala, Sweden: Association for Computational Linguistics; 2010. p. 28–36. http://www.aclweb.org/anthology/W10-1904.
18. Hakenberg J, Leaman R, Vo NH, Jonnalagadda S, Sullivan R, Miller C, et al. Efficient extraction of protein-protein interactions from full-text articles. IEEE/ACM Trans Comput Biol Bioinform. 2010;7(3):481–94.
19. Ben Abacha A, Zweigenbaum P. Automatic extraction of semantic relations between medical entities: a rule based approach. J Biomed Semantics. 2011;2(Suppl 5):S4.
20. Hearst MA. Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th Conference on Computational Linguistics - Volume 2. COLING '92. Nantes, France; 1992. p. 539–45.
21. Hakenberg J, Plake C, Royer L, Strobelt H, Leser U, Schroeder M. Gene mention normalization and interaction extraction with context models and sentence motifs. Genome Biol. 2008;9(Suppl 2):14.
22. Palaga P, Nguyen L, Leser U, Hakenberg J. High-performance information extraction with alibaba. In: Proc. of the 12th Int. Conf. on Extending Database Technology: Advances in Database Technology. EDBT '09. New York, NY, USA: ACM; 2009. p. 1140–1143.
23. Hakenberg J, Schroeder M, Leser U. Consensus pattern alignment to find protein-protein interactions in text. In: Proc. Second BioCreative Challenge Evaluation Workshop. Madrid, Spain; 2007.
24. Agrawal R, Srikant R. Mining sequential patterns. In: International Conference on Data Engineering. IEEE Computer Society; 1995. p. 3–14.
25. Frawley WJ, Piatetsky-Shapiro G, Matheus CJ. Knowledge discovery in databases: An overview. In: Knowledge Discovery in Databases. Anaheim, CA, USA: AAAI/MIT Press; 1991. p. 1–30.
26. Srikant R, Agrawal R. Mining sequential patterns: Generalizations and performance improvements. In: International Conference on Extending Database Technology. London, UK: Springer-Verlag; 1996. p. 3–17.
27. Pei J, Han B, Mortazavi-Asl B, Pinto H. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In: International Conference on Data Engineering. Washington, DC, USA: IEEE Computer Society; 2001. p. 215–24.
28. Zaki M. Spade: An efficient algorithm for mining frequent sequences. Mach Learn. 2001;42(1/2):31-60.
29. Wang J, Han J. Bide: Efficient mining of frequent closed sequences. In: Proc. of the 20th Int. Conf. on Data Engineering. ICDE '04. Boston, MA, USA: IEEE Computer Society; 2004. p. 79.
30. Nanni M, Rigotti C. Extracting trees of quantitative serial episodes. In: Knowledge Discovery in Inductive Databases 5th Int. Workshop KDID'06, Revised Selected and Invited Papers. Berlin, Germany: Springer; 2007. p. 170–88.
31. Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. Frontiers of biomedical text mining: current progress. Brief Bioinform. 2007;8:358–375.
32. Pei J, Han B, Lakshmanan LVS. Mining frequent itemsets with convertible constraints. In: Proc. of the Int. Conf. on Data Engineering. Washington, DC, USA: IEEE Computer Society; 2001. p. 433–42.
33. Crémilleux B, Soulet A, Kléma J, Hébert C, Gandrillon O. Discovering Knowledge from Local Patterns in SAGE Data. Hershey, Pennsylvania, USA: IGI Publishing; 2008.
34. Ng RT, Lakshmanan LVS, Han J, Pang A. Exploratory mining and pruning optimizations of constrained association rules. In: SIGMOD International Conference on Management of Data. New York, NY, USA: ACM Press; 1998. p. 13–24.
35. Cellier P, Charnois T, Plantevit M, Crémilleux B. Recursive sequence mining to discover named entity relations. In: International Symposium on Advances in Intelligent Data Analysis. LNCS, vol 6065. Berlin, Germany: Springer; 2010. p. 30–41.
36. Cellier P, Charnois T, Plantevit M. Sequential patterns to discover and characterise biological relations. In: International Conference on Intelligent Text Processing and Computational Linguistics. Berlin, Germany: LNCS; 2010. p. 537–48.
37. Rosario B, Hearst MA. Multi-way relation classification: application to protein-protein interactions. In: Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada; 2005. p. 732–9.
38. Schmid H. Probabilistic part-of-speech tagging using decision trees. In: International Conference on New Methods in Language Processing. Manchester, UK; 1994. p. 44–9.
39. DMT4SP tool. http://liris.cnrs.fr/~crigotti/dmt4sp.html.
40. Temkin JM, Gilder MR. Extraction of protein interaction information from unstructured text using a context-free grammar. Bioinformatics. 2003;19:2046-53.
41. Hao Y, Zhu X, Huang M, Ming L. Discovering patterns to extract protein-protein interactions from the literature : Part ii. Bioinformatics. 3294.
42. Farkas R, Vincze V, Mora G, Csirik J, Szarvas G. The conll-2010 shared task: Learning to detect hedges and their scope in natural language text. In: Conference on Computational Natural Language Learning: Shared Task. Uppsala, Sweden; 2010.
43. Bunescu R, Ge R, Kate RJ, Marcotte EM, Mooney RJ, Ramani AK, et al. Comparative experiments on learning information extractors for proteins and their interactions. Artif Intell Med. 2005;33(2):139–55.
44. Pyysalo S, Ginter F, Heimonen J, Björne J, Boberg J, Järvinen J, et al. Bioinfer: a corpus for information extraction in the biomedical domain. BMC Bioinformatics. 2007;8(1):50.
45. Fundel K, Küffner R, Zimmer R. Relex—relation extraction using dependency parse trees. Bioinformatics. 2007;23(3):365–71.
46. Pyysalo S, Airola A, Heimonen J, Bjorne J, Ginter F, Salakoski T. Comparative analysis of five protein-protein interaction corpora. BMC Bioinformatics. 2008;9(Suppl 3):6. doi:10.1186/1471-2105-9-s3-s6.
47. HGNC (HUGO Gene Nomenclature Committee). http://www.genenames.org/.
48. Tsuruoka Y, Tsujii J. Improving the performance of dictionary-based approaches in protein name recognition. J Biomed Inform. 2004;37(6):461–70.
49. Aggarwal BB, Kunnumakkara AB, Harikumar KB, Gupta SR, Tharakan ST, Koca C, et al. Signal transducer and activator of transcription-3, inflammation, and cancer: how intimate is the relationship? Ann NY Acad Sci. 2009;1171(Natural Compounds and Their Role in Apoptotic Cell Signaling Pathways):59–76.