Using Relations to Index Biological Document Repositories for Efficient Searching

Lipika Dey **Rohit Goyal**

Department of Mathematics
Indian Institute of Technology Delhi
Hauz Khas, New Delhi-110016
India

lipika@maths.iitd.ac.in *rohitgyl@gmail.com*

Abstract

In this paper we propose a rule-based mechanism using Natural Language Processing techniques for extracting biological relations from biomedical text documents. While the rules identify frequently occurring patterns that can be potential relations, significant relations are identified using statistical analysis. Evaluation of the technique has been done on MEDLINE abstracts obtained from the GENIA corpus. Results indicate that our technique has good potential for other text mining applications also. Preliminary analysis shows that indexing biomedical documents on these relations can facilitate high precision document retrieval.

1. Introduction

The life science industry is an emerging market in which application spaces such as drug discovery, development in the pharmaceutical sector and clinical record management in health care, have become areas of significant research interest. Documents in the scientific literature play an important role in life sciences by serving as a potential source for knowledge discovery. These documents offer a rich repository of information on relationships among biomedical concepts such as genes, proteins, diseases, and a variety of other key topics. Due to the enormity of the repository, retrieving the relevant information is not an easy task. The searching mechanism provided by popular Biomedical search engines like PUBMED is based on pattern matching and does not always exploit biological knowledge to do function-based searching. We highlight this with an example query issued over PUBMED. A query "*STAT activation*" retrieves a list of abstracts. On detailed analysis, it is revealed that some of these, for example PMID: 16731155 do not mention anything about *STAT activation*. The two sentences for which this abstract is chosen as an answer are as follows:

"*Animals were cooled to 15 degrees C or 34 degrees C on cardiopulmonary bypass (pH stat, hematocrit 30%, pump flow 100 mL/kg/minute) followed by 2 hours of low flow (50 mL/kg/minute) or very low flow (10 mL/kg/minute). ...This is associated with delayed capillary reperfusion. Reduced eNOS is also associated with increased white cell activation which may lead to greater neurologic injury.*"

Clearly this is not a relevant document. Marshall *et al.* [1] proposed the use of biomedical pathway relations to index the digital library of abstracts maintained by PUBMED to help users harness the vast amount of information embedded in the documents. Rindflesch *et al.* [2] also suggested that a detailed and comprehensive source of information regarding biomolecular functions could be utilized for effective information retrieval from text documents. Though several systems have been proposed in the past few years to extract biological relations from the text documents, most of the previous tasks in this field, other than [3], have focused on extracting patterns pertaining to a predefined set of relations from biological text. The *a priori* specification was both in terms of the nature of the relations to be extracted as well as the type of relations. Detailed discussion about the earlier systems has been provided in section 2.

A biological relation depicts the role of biological entities in biological processes. A biological process is a series of events accomplished by one or more ordered assemblies of molecular functions. Examples of broad biological process terms are "*molecular activation"* or "*signal transduction*." An example of a biological relation involving a biological process is "*activation of NF-kappa B*."

The emphasis of the proposed work is to discover patterns and trends of relations involving proteins from text documents without any assumption either about the nature or the type of relations. Starting with the protein name

International Conference on Management of Data
COMAD 2006, Delhi, India, December 14-16, 2006

extraction from text using the algorithm PROPER developed by Fukuda *et al.* [4], this work exploits the grammatical dependencies between various verbs/nouns and entities of a sentence, as elucidated by the Stanford English PCFG Parser [5, 6]. This is a statistical parser. We have proposed a rule base to analyze these dependencies and infer different categories of relations among proteins. The set of relations thus extracted are subjected to significance analysis and only the significant ones are retained. The significant relations are thereafter used to index text documents for improving retrieval effectiveness. Though this work was initiated for mining biological relations involving proteins from text, it was later observed that the methodology is easily extensible to recognizing a certain category of biological entities also. Since PROPER uses a set of rules to identify special characters in names, hence some elements, which contain special characters but are not proteins, are also extracted in the process. However, it was observed that usually these elements are part of an entity, which also includes a feature term, but this feature term is not recognized by PROPER. The present work identifies a method for identifying a feature set for recognizing these terms.

2. Related work on biological relation mining from text documents

Though relation extraction systems vary significantly in their output formats, most of the biological relation mining systems output labeled relational triplets. Friedman *et al.* [7] proposed the GENIES system, which uses Natural Language Processing techniques to capture complex biological relations from text with the output expressed as nested predicates. This format is precise and can express much of the relational information contained in a text. Relations extracted by GENIES are integrated into the larger GeneWays system [8] after they are "unwound" into simple binary statements like "Interlukin-2 binds Interlukin-2 receptor". The GeneScene parser [9] uses an approach based on finite state automata for extracting relational triplets. The Arizona Relation Parser (ARP) [10] also extracts relational triplets with labelled links connecting labeled entities. The GeneScene parser focuses on closed class words to identify important relations while ARP uses a hybrid syntax and semantic parser. Both extract negation indicators. In the ARP results, the link labels consist primarily of verbs or verb phrases and a negation indicator. Entities are phrases (generally noun) extracted from the text. ARP output supports relation nesting. The system developed by Palakal *et al.* [11] adds tags to text marking the boundaries of identified biological objects and extracts both directional relationships such as "binds" and "inhibits" as well as a hierarchical relations like "is a". Relations are specified by a subject, an object and a relation specification. Rinaldi *et al.* [12] propose a method for discovery of interactions between genes and proteins from the scientific literature, based on a complete syntactic analysis of the corpus. It is based on full parsing of the documents and on a set of rules that map syntactic structures into the relevant relations. Ciaramita *et al.* [3] present an unsupervised model for learning arbitrary relations between concepts of molecular biology ontology to support text mining and manual ontology building.

While the previous works have focused mainly on extracting labeled relational triplets (binary relations) and identifying relations pertaining to specific biological entities like proteins and genes or specific biological processes like "*binds*" and "*activation*", in this work, we have also emphasized on identification of two other types of relations (see section 4.2), and our methodology is applicable to a wide range of biological entities and processes.

3. Obtaining dependency graphs for sentences using the Stanford Parser

The Stanford Parser identifies *typed dependencies* between individual words of a sentence as grammatical relations. Dependencies are particularly useful for developing Natural Language based applications. In order to extract the dependencies from a sentence, the semantic head of the sentence is identified, using a set of rules. Since all the other words in a relative clause depend on the head, the verb is chosen as head when determining dependencies. In general, content words are chosen as heads, while auxiliaries, complementizers, etc. are dependents of them.

dep - Basic dependency
  aux - Auxiliary dependency
  arg - Argument
    subj - Subject
      nsubj - Nominal subject
        nsubjpass - Passive nominal subject

    comp- Complement
      obj - Object
        dobj - Direct object
        iobj - Indirect object
        pobj - Object of preposition
  mod - Modifier
    amod - Adjectival modifier
    det - Determiner
    of, in, by etc. - prepositional modifier

Figure 1. Dependency codes for Stanford Parser

A typed dependency labels dependencies with *grammatical relations*, such as subject or indirect object, since these provide information about the predicate-argument structure. The grammatical relations are arranged in a hierarchy, rooted with the most generic relation, *dependent*. When the

relation between a head and its dependent can be identified more precisely, relations further down in the hierarchy are used to represent the dependency. For example, the dependent relation can be specialized to **aux** (auxiliary), **arg** (argument), or **mod** (modifier). The **arg** relation can be further divided into the **subj** (subject) relation, the **comp** (complement) relation, and so on. Figure 1 depicts the list of some root dependencies that the Stanford parser identifies and the ones that we have used for relation extraction, in a hierarchical fashion. A brief description of the nature of the dependencies is also provided. One can refer to [5] for more details.

For a given sentence, a dependency between two words of the sentence is returned in the following format by the parser:

dependency(word1-position1,word2-position2)

Here, "*word1*" and "*word2*" are the two words of the sentence. "*dependency*" is the grammatical relation that exists between them. "*position1*" and "*position2*" are the respective positions of the two words in the sentence.

For mining biological relations from text documents, the emphasis is on identifying links between biological processes (present as verbs or nouns) and biological entities (present as nouns). Hence, the analysis of dependency structures extracted by the Stanford Parser ensures a good starting point to identify significant biological relations from text documents.

4. Methodology of finding biological relations from GENIA corpus

In this section, we present the methodology that was implemented by us to mine frequently occurring relations with proteins as the actors. The relations were mined from a set of 1000 MEDLINE abstracts, that were randomly chosen from the GENIA corpus, and the results were verified through manual evaluation.

4.1 Entity Extraction

The entities representing protein names were extracted by implementing the NLP-based entity recognition algorithm PROPER [4]. We provide a brief description of the same.

Protein names have several characteristics that can be used to identify them. Typically these words contain capital letters, numerical figures, and special symbols. These words can be clearly distinguished from general words due to these characteristics. These characteristics provide large amount of information to the reader and can be considered as the core of biological entity names. In this respect, such words that appear in protein names are called "***core-terms***". For example, the following entity names have a core-term each, which is underlined:

(i) Src homology (SH) 2 and SH3 domains

(ii) p54 SAP kinase

Furthermore, as in the following examples, certain key-words can be included in the entity name, that describe the function and characters of a compound word that represents a protein name. For example, the following entity names have a key word each, which is underlined.

(i) EGF receptor

(ii) Ras GTPase-activating protein (GAP)

Such words are called "***feature-terms" (f-terms).*** PROPER uses a fixed list of such feature terms pertaining to proteins that is fed to the algorithm.

The algorithm works in following 3 phases:

(i) Core-term extraction from tokenized texts

(ii) Concatenation of core-terms and f-terms

a. Rebuild "***core-blocks***" (noun phrases without conjunction and preposition)

b. Rebuild dependencies (noun phrases with conjunction and prepositions)

(iii) Demark unnecessary annotations

Using the above-mentioned entity-extraction algorithm, protein names were identified from the 1000 abstracts. Each entity was assigned a unique identifier.

4.2 Characterizing biological relationships

Our aim is to identify and characterize all relations pertaining to protein entities that occur in a corpus. The basic mechanism of relation characterization was based on analyzing the dependency graphs involving the protein entities that emerged from the text documents. Since protein names are compound in nature, hence it is observed that using the protein names directly generates wrong dependency graphs. Hence we introduced a preprocessing step, in which each unique protein name recognized is assigned an identity, and each occurrence of a protein name is replaced by the unique identifier assigned to it.

For example, given a sentence like "*STAT5 was activated by IL-2*", we first replace *STAT5* with the identifier *Entity1* and *IL-2* with *Entity2*. So the sentence that is input to the parser has the form "*Entity1 was activated by Entity2*". The complete list of dependencies output by the parser, inclusive of the position of the word is as follows:

by(activated-3, Entity2-5)

aux(activated-3, was-2)

nsubjpass(activated-3, Entity1-1)

The preposition "*by*" helps in identifying the active and the passive entities taking part in the relation. The biological relation that is present in the sentence, can be captured by analyzing this output as follows:

(i) *STAT5(Entity)—activated(Process)-by→IL- 2(Entity)*

(ii) *activated(Process)-by→IL-2(Entity)*

The directed line towards *IL-2* indicates that the entity *IL-2* takes an active role with respect to the preposition and the process activation i.e. activation is done **by** *IL-2*. If no preposition is relevant then we use an undirected line as in case of *STAT5* and *activation*.

This example illustrates the efficacy of the dependency graph in characterizing biological relations. Analysis of the various relevant dependency graphs that emerged from the corpus of 1000 documents showed that biological relations could be characterized into three classes:

(i) **Type I**: A relation involving a single biological entity and a process. While the biological entity always occurs as a noun, the process occurs in the sentence either as a noun or as a verb followed by a preposition. The dependency structure elucidates the relationship between the process and the entity. There may or may not be another entity in the sentence that is also associated with the same process. For example the sentence "*Activation of NF-kappa B takes place in certain conditions*." contains the Type I relation:

activation(Process)-of → NF-kappa B(Entity)

(ii) **Type II**: A relation involving one biological entity and two processes is termed as a Type II relation. There is a preposition relevant to one of the processes. For example, the sentence "*This led to the inhibition of NF-kappa B activation*" contains the Type II relation:

inhibition(Process1)-of→NF-kappaB(Entity)—activation(Process2)

(iii) **Type III**: A relation that involves two biological entities and a single biological process is categorized as a Type III relation. There may or may not be a preposition associated to the process in the relation. If there is a preposition associated to the process and related to one of the entities, then the Type III relation also contains a Type I relation.

For example "STAT5 was activated by IL-2" is a sentence that has the Type III relation -

STAT5(Entity)—activated(Process)-by → IL-2(Entity),

as well as the Type I relation -

activated(Process) –by → IL-2(Entity).

While a lot of work has already been reported on Type III relations [13, 3], the Type I and II relations proposed here, have not received much focus earlier. Hence, in this work we report detailed statistics and results about occurrence patterns of Type I and II relations. However, for the sake of completeness of the rule set, we provide rules for identifying all the three types of relations.

4.3 Extracting relations from corpus

Having identified the basic categories of relations that have to be extracted from text documents, the next task was to design appropriate dependency-graph traversal mechanisms that would be able to recognize instances of these relations within texts. We now present a detailed discussion on the proposed set of rules that are used to extract all relations belonging to one of the three categories from the corpus.

We denote all subject related dependency codes (subj, nsubj, nsubjpass etc.) by "**Sub**", all object related dependency codes (obj, dobj etc.) by "**Obj**" and preposition codes (to, from, by etc.) by "**Preposition**" in the rule set. We do this because as far as the biological relations are concerned we do not need to differentiate between the different kinds of subjects and objects. Remaining codes are the same as given in Figure 1.

Rule 1 - On traversing the dependency set, if there exists a dependency involving one entity and one process, satisfying the following conditions: [**Preposition**(Process, Entity1)], then it characterizes the following Type I relation:
Process-Preposition → Entity1

For example, this rule would fire for the sentence, *"These results suggest that phosphorylation at one or both of these residues is critical for activation of NF-kappa B*", which has the following relevant dependencies:

at(phosphorylation-5, one-7)

of(both-9, residues-12)

that(suggest-3, critical-14)

of(activation-16, Entity1-18)

for(critical-14, activation-16)

det(residues-12, these-11)

aux(critical-14, is-13)

or(phosphorylation-5, both-9)

det(results-2, These-1)

nsubj(critical-14, phosphorylation-5)

nsubj(suggest-3, results-2)

The relation extracted is- *activation-of → Entity1* where *Entity1* stands for *NF-kappa B.*

Rule 2 - On traversing the dependency set, if there exist three dependencies involving one entity and the same process, satisfying the following conditions - [**Subj**(word1, Process) & **Obj**(word1, word2) & **Preposition**(word2, Entity1)], then [**Subj**(word1, Process) & **Obj**(word1, word2)] give the following intermediate relation:

word2 — Process

[**Preposition**(word2, Entity1) & word2—Process] characterize the following Type I relation:

Process-Preposition → Entity1

This rule is capable of capturing more complex notions about the biological process. For example, this rule fires for the sentence "*The activation takes place in Entity1*." which has the following dependencies:

in(place-4, Entity1-6)

dobj(takes-3, place-4)

det(activation-2, The-1)

nsubj(takes-3, activation-2)

The relation extracted is *activation-in → Entity1*.

Rule 3 - On traversing the dependency set, if there exist two dependencies involving one entity but two different processes, satisfying the following conditions- [**amod**(Process1, Entity1) & **Preposition**(Process2, Process1)], then they characterize the following Type II relation:

Process2-Preposition → Entity —Process1

We place a constraint on the word *Process2* that it should not be an adjective. This constraint can be implemented using "Part of Speech (POS)" analysis, for which we use the Stanford POS Tagger [14, 15].

This constraint eliminates wrong instances of Type II relations as shown below, since biological processes occur either as nouns or as verbs:

necessary-for → IL-7R — transduction

We also place a constraint on word *Process1* that it should not be one of the newly identified *feature-terms* (Identification of new *feature-terms* is explained in section 5). This is also done to eliminate wrong type II relations like:

transcription-of → immunoglobulin — gene

For example, this rule fires for the sentence "*Therefore, inhibition of NF-kappa B activation may be an effective strategy for acquired immunodeficiency syndrome therapy.*" which has the following relevant dependencies (other dependency outputs of the Stanford Parser for this sentence have not been shown for the sake of clarity):

of(inhibition-3, activation-6)

amod(activation-6, Entity1-5)

The relation extracted is:

inhibition-of→Entity1— activation, where *Entity1* stands for *NF-kappa B.*

Rule 4 – On traversing the dependency set, if there exist two dependencies involving two different entities, but the same process satisfying the following conditions- [**Obj**(Process, Entity2) & **Subj**(Process, Entity1)], then they characterize an instance of a Type III relation represented by:

Entity1—Process— Entity2

For example this rule would fire for the sentence "*LMP-1 activates NF-kappa B by targeting the inhibitory molecule I kappa B alpha.*" which has the following dependencies:

by(activates-2, targeting-5)

det(Entity3-7, the-6)

nsubj(activates-2, Entity2-1)

dobj(targeting-5, Entity3-7)

dobj(activates-2, Entity1-3)

The relation extracted is:

Entity1— activates— Entity2, where *Entity 1* stands for *LMP-1* and *Entity2* stands for *NF-kappa B*.

It may be noted that this relation is not accompanied by a preposition. Though other dependencies are also extracted from the sentence involving *Entity3*, at present no rule is fired for this pattern and hence is not considered.

Rule 5 - On traversing the dependency set, if there exist three dependencies involving two different entities but the same process, satisfying the following conditions- [**Subj**(word, Entity1) & **Preposition1**(word, Process) & **Preposition2**(Process, Entity2)], then [**Subj**(word, Entity1) & **Preposition1**(word, Process)] , gives the following intermediate relation:

Process— Entity1

[**Preposition2**(Process, Entity2) & Process — Entity1] characterize the following Type III relation:

Entity1— Process-Preposition2 → Entity2

For example, this rule would fire for the sentence "*It can be concluded that IL-6 is responsible for the activation of STAT proteins in a primary T cell response.*" which has the following relevant dependencies (other dependency outputs of the Stanford Parser for this sentence have not been shown for the sake of clarity):

for(responsible-8, activation-11)

of(activation-11, Entity2-13)

nsubj(responsible-8, Entity1-6)

The relation extracted is:

Entity1— activation-of→Entity2, where *Entity1* stands for *IL-6* and *Entity2* stands for *STAT proteins*.

Rule 6 - On traversing the dependency set, if there exist two dependencies involving two different entities but the same process, satisfying the following conditions- [**Preposition**(Process, Entity2) & **Subj**(Process, Entity1)], then they characterize the following Type III relation-

Entity1— Process- Preposition → Entity2

For example, this rule would fire for the sentence "*HIV-1 expression was activated from J delta K cells by treatment with phorbol myristate acetate (PMA)*", which has the following relevant dependencies (other dependency outputs of the Stanford Parser for this sentence have not been shown for the sake of clarity):

from(activated-3, Entity2-5)

nsubjpass(activated-3, Entity1-1)

The relation extracted is:

Entity1 — activated-from → Entity2 where *Entity1* stands for *HIV-1 expression* and *Entity2* stands for *J delta K cells*.

Rule 7 – On traversing the dependency set, if there exist three dependencies involving two different entities but the same process satisfying the following conditions: [**Subj**(word, Entity1) & **Obj**(word, Process) & **Preposition**(Process, Entity2)], then [**Subj**(word, Entity1) & **Obj**(word, Process)] give the following intermediate relation:

Process— Entity1

[**Preposition**(Process, Entity2) & Process—Entity1] characterize the following Type III relation:

Entity1—Process-Preposition → Entity2

For example, this rule would fire for the sentence "*An interferon-gamma activation sequence mediates the transcriptional regulation of the IgG Fc receptor type IC gene.*" which has the following dependencies:

of(regulation-6, Entity2-9)

amod(regulation-6, transcriptional-5)

det(Entity1-2, An-1)

dobj(mediates-3, regulation-6)

nsubj(mediates-3, Entity1-2)

det(Entity2-9, the-8)

det(regulation-6, the-4)

The relation extracted is:

Entity1— regulation-of → Entity2, where *Entity1* stands for *interferon-gamma activation sequence* and *Entity2* stands for *IgG Fc receptor type IC gene.*

5. Extending the rule set to extract non-protein entity names containing feature terms

Biological entities other than proteins also consist of *core-terms* and *feature-terms* in most cases. The entity names that do not contain *feature-terms* but only special characters are already being extracted by PROPER. However, since PROPER uses a feature term set for recognizing proteins only, it does not recognize other entities correctly.

We have identified an enhanced set of feature terms, using the dependency output of the Stanford Parser that can recognize other non-protein biological entities also. Earlier, Naraynaswami *et al.* [16] had proposed a list of *feature-terms* for various categories of biological entities. However, that is very limited and the method proposed by us contains this set as a subset. Using this enhanced entity set, we also extracted a set of relations involving these entities using the earlier methodology.

Consider for example the entity *"Jurkat line"* which is type of cell line, in this case "*Jurkat"* is a *core-term* and *"line"* is a *feature-term*. Now the existing algorithm would identify only *Jurkat* as an entity and leave out *line* because line is not in the feature list. A similar example is "*HIV enhancer*". We observe that when we input a sentence to the parser in which only *Jurkat* has been identified as an entity then one of the dependencies that the Stanford Parser outputs is:

amod(line, Entity1)

where, *Entity1* is the label for *Jurkat*, i.e. *Jurkat* has been identified as an adjectival modifier for *line*.
So we analyzed the list of all those dependencies output by the Stanford Parser that were of the form **amod**(Word, Entity) and it was observed that in 65% of the cases "*Word*" qualified as a feature-term for "*Entity*".

Table 1 shows the list of some of the *feature-term*s that we extracted using the proposed methodology. The newly identified list of *feature-terms* was added to the existing list of *feature terms* and the whole procedure of relation extraction as described in section 4 was repeated.

6. Statistical analysis of biological relations extracted

After all the potential relations had been extracted from our set of 1000 abstracts (as mentioned earlier, these were randomly chosen form the GENIA corpus), the following statistical analysis was carried out to test the feasibility of the relations. The feasibility test was conducted using either "G-test" which is based on log-likelihood ratio or "Fisher's Exact Test". G-test is a good statistic to measure the significance of those relations for which the minimum expected frequency requirements are greater than or equal to 5 for a 2*2 contingency table. For those relations whose minimum expected frequency was found to be less than 5, Fisher's Exact Test was used for significance analysis. In

both cases the confidence level used is 95%.

Table 1. List of newly extracted feature-terms

Feature-term	Example
Enhancer	HS2 enhancer
Line	Jurkat line
Box	X2 box
Cell	K562 cell
fragment	DNA fragment
clone	cDNA clone
virus	Epstein-Barr virus
lineage	B-cell lineage
construct	LTR construct
homodimer	p50 homodimer
Element	cis-acting element
Provirus	HIV provirus
superfamily	Ig superfamily
Gene	immunoglobulin gene

A total of 6520 unique patterns were extracted as potential Type I relations, from the corpus using the rules defined earlier. Only those patterns for which the occurrence frequency was greater than 3, were subjected to significance analysis. Table 2 lists some of the Type I relations, which passed the significance tests. The patterns that passed the significance tests are accepted as biological relations.

A total of 1320 unique patterns were identified as potential Type II relations from the corpus using the rules defined earlier. Since Type II relations are not as frequent as Type I relations, only those patterns which occurred with frequency greater than or equal to 2, were considered for significance test. Table 3 lists some of the Type II relations, which were accepted as significant.

Since the relations identified as significant are to be used for indexing the document collection, hence these were subjected to manual verification. Each occurrence of the pattern was traced to its origin in a sentence and the sentence was analyzed for correctness. Table 4 summarizes the precision of the results by checking for their biological relevance.

An example of wrong Type I biological relation that got selected is:

effects-of → iNO

However *effects* is not a valid biological process.

Table 2. List of significant Type I biological relations

Process	Entity
Activation-of	NF-kappa B
Activation-of	PKC
Dephosphorylation-of	PRb
Phosphorylation-of	HS1
Phosphorylation-of	Rap1 protein
Degradation-of	I kappa B alpha
Binding-to	DNA

Table 3. List of significant Type II biological relations

Process 1	Entity	Process 2
Inhibition-of	NF-kappa B	Activation
Blockade-of	HL-60	Differentiation
Inhibition-of	HIV-1	Replication
Induction-by	IL-2	Stimulation
Suppression-of	SLA-DR	Induction
Inhibition-of	IgE	Production
Modulation-of	NF-kappa B	Activation

Table 4. Precision of Type I and II relations, identified from GENIA corpus

Relation Type	Total Feasible Relations	Precision
Type I	17	70.6%
Type II	35	60%

Similarly, an example of wrong Type II relation that got selected is:

leading-to → NF-kappa B— activation, where *leading* is not a valid biological process.

It is observed that this problem is more dominant in Type II relations. We can improve the precision of extracting Type II relations by starting with a target set of biological processes, thereby constraining the Type II relations to contain only those processes which are a part of this target set.

7. Indexing biological documents on relations

The statistically selected feasible relations are used to index the corpus. It helps in searching for abstracts in response to complex user queries like "*inhibition of NF-kappa B*

activation", which is a Type II query for our system. Table 5 shows three sentences that are retrieved by our system in response to this query.

Table 5. Sample sentences retrieved for the query *"inhibition of NF-kappa B activation"*

Sentences
The inhibition of NF-kappa B activation by antioxidants and specific protease inhibitors may provide a pharmacological basis for interfering with these acute processes.
These results are the first to suggest that surfactant's suppressive effects on inflammatory cytokine production may involve transcriptional regulation through inhibition of NF-kappa B activation.
In contrast, long-term treatment with oxLDL prevented the lipopolysaccharide-induced depletion of I kappa B-alpha, accompanied by an inhibition of both NF-kappa B activation and the expression of tumor necrosis factor-alpha and interleukin-1 beta genes.

A total of 20 documents were found to contain this information correctly, of which 15 were correctly retrieved by our system. Each of these sentences was judged for correctness manually. No other sentences were retrieved by our system for this query. Hence it can be said that this query is processed with 100% precision and 75% recall. Table 6 summarizes the precision and recall of 3 other queries tested over the GENIA corpus.

Table 6. Precision and recall of 3 queries

Query	**Precision**	**Recall**
Activation of PKC	100%	80%
Phosphorylation of HS1	100%	71%
Degradation of I kappa B alpha	100%	70%

On feeding the query "*inhibition of NF-Kappa B activation*" to PUBMED, one of the top 5 documents is the document with PMID: 16857678, which is not relevant to the query as it can be judged from the content presented in Figure 2. Hence we can say that indexing repositories of biological documents like PUBMED on biological relations can help in retrieving documents with high precision.

8. Conclusions and future work

A new NLP based technique to characterize biological relations has been proposed in this paper. This is achieved through a rule-based analysis of a biological corpus, where the rules study the relationship between biological entities and processes. The patterns extracted as potential relations are subjected to statistical significance test. The method is capable of extracting biological relations with 70% accuracy. A methodology for extending PROPER has been proposed to extract a larger set of biological entities. Preliminary results show that the concept of indexing biological documents on relations can yield promising results by retrieving documents with high precision.

> *Activation of NF-kappaB and autophagy are two processes involved in the regulation of cell death, but the possible cross-talk between these two signaling pathways is largely unknown. Here we show that NF-kappaB activation mediates repression of autophagy in TNFalpha-treated Ewing sarcoma cells. This repression is associated with an NF-kappaB-dependent activation of the autophagy inhibitor mTOR. In contrast, in cells lacking NF-kappaB activation, TNFalpha treatment upregulates the expression of the autophagy-promoting protein Beclin 1, and subsequently induces the accumulation of autophagic vacuoles. Both of these responses are dependent on reactive oxygen species (ROS) production and can be mimicked in NF-kappaB-competent cells by the addition of H2O2. Small interfering RNA-mediated knock down of Beclin 1 and atg7 expression, two autophagy-related genes, reduced TNFalpha- and ROS-induced apoptosis in cells lacking NF-kappaB activation and in NF-kappaB-competent cells, respectively. These findings demonstrate that autophagy may amplify apoptosis when associated with a death signaling pathway. They are also evidence that inhibition of autophagy is a novel mechanism of the anti-apoptotic function of NF-kappaB activation. We suggest that stimulation of autophagy may be a potential way bypassing the resistance of cancer cells to anticancer agents that activate NF-kappaB.*

Figure 2. Abstract retrieved from PUBMED for query, "*inhibition of NF-Kappa B activation*"

Recall value computation for biological relation based retrieval is a complex task. In future, we propose to do this for all the significant relations identified. This would also involve identifying new relations that were not covered by the existing set of rules. The rule base for relation extraction can also be enhanced for eliminating noise and ensuring good recall of relations as we observed cases where a relation was missed by the rule base. It was observed that relations were also missed due to incorrect identification of the biological entities or due to the entities not being recognized at all. Using a combination of NLP and dictionary based approach may yield better results for entity extraction.

A detailed analysis of extraction of Type III relations and comparison of our methodology with existing ones needs to be done. Also a complete indexing is under implementation with the entire set of biological relations

identified.

References

[1] B. Marshall, K. Quiñones, H. Su, S. Eggers and H. Chen. Visualizing Aggregated Biological Pathway Relations, *JCDL'05*, 2005.

[2] T.C. Rindflesch, L. Tanabe, J. N. Weinstein and L. Hunter. EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature, *Proceedings of the Pacific Symposium on Biocomputing,* 2000.

[3] M. Ciaramita, A. Gangemi, E. Ratsch, J. Sari and I. Rojas. Unsupervised Learning of Semantic Relations between Concepts of Molecular Biology Ontology, *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, Scotland, 2005.

[4] K. Fukuda, T. Tsunoda, A. Tamura and T. Takagi. Toward Information Extraction: Identifying Protein Names from Biological Papers, *Proceedings of the Pacific Symposium on Biocomputing,* 1998.

[5] M. Marneffe, B. MacCartney and C. D. Manning. Generating Typed Dependency Parses from Phrase Structure Parses, *5th International Conference on Language Resources and Evaluation,* 2006.

[6] Stanford Parser, *http://nlp.stanford.edu/software/lexparser.shtml*

[7] C. Friedman, P. Kra, M. Krauthammer, H. Yu and A. Rzhetsky. GENIES: A Natural-Language Processing System for the Extraction of Molecular Pathways from Journal Articles, *Bioinformatics,* 17, 74-82, 2001.

[8] A. Rzhetsky, I. Iossifov, T. Koike, M. Krauthammer, P. Kra, M. Morris, H. Yu, P.A. Duboue, W.Weng, W.J. Wilbur, V. Hatzivassiloglou, and C. Friedman. Geneways: A System for Extracting, Analyzing, Visualizing, and Integrating Molecular Pathway Data, *Journal of Biomedical Informatics,* 37, 43-53, 2004.

[9] G. Leroy, H. Chen, and J. D. Martinez. A Shallow Parser Based on Closed Class Words to Capture Relations in Biomedical Text, *Journal of Biomedical Informatics,* 36, 145-158, 2003.

[10] D. M. McDonald, H. Chen, H. Su and B. Marshall. Extracting Gene Pathway Relations using a Hybrid Grammar: the Arizona Relation Parser, *Bioinformatics*, 20, 3370-3378, 2004

[11] M. Palakal, M. Stephens, S. Mukhopadhyay, R. Raje and S. Rhodes. Identification of Biological Relationships from Text Documents using Efficient Computational Methods, *Journal of Bioinformatics and Computational Biology,* 1, 307-342, 2003.

[12] F. Rinaldi, G. Schneider, K. Kaljurand, J. Dowdall, C. Andronis, A. Persidis and O. Konstanti. Mining Relations in the GENIA Corpus, *Proceedings of the Second European Workshop on Data Mining and Text Mining for Bioinformatics*, 2004.

[13] M. Abulaish and L. Dey. Biological Relation Extraction and Query Answering from Medline Abstracts using Ontology-Based Text Mining, to appear in *Data and Knowledge Engineering*, 2006.

[14] K. Toutanova and C. D. Manning. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger, *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 2000.

[15] Stanford Tagger, http://nlp.stanford.edu/software/tagger.shtml

[16] M. Narayanaswamy, K. E. Ravikumar and K. Vijay-Shanker. A Biological Named Entity Recognizer, *Pacific Symposium on Biocomputing,* 8, 427-438, 2003.