Biological Relation Extraction and Query Answering from Medline Abstracts using Ontology-Based Text Mining

Muhammad Abulaish
Department of Mathematics
Jamia Millia Islamia (A Central University)
Jamia Nagar, New Delhi – 110 025, India
abulaish@computer.org

Lipika Dey [#]
Department of Mathematics
Indian Institute of Technology, Delhi
Hauz Khas, New Delhi -16, India
lipika@maths.iitd.ernet.in

Abstract

The rapid growth of the biological text data repository makes it difficult for human beings to access required information in a convenient and effective manner. The problem arises due to the fact that most of the information is embedded within unstructured or semi-structured text that computers can not interpret very easily. In this paper we have presented an ontology-based Biological Information Extraction and Query Answering (BIEQA) System, which initiates text mining with a set of ontological concepts stored in a biological ontology, and thereafter mines possible biological relations among those concepts using NLP techniques and co-occurrence-based analysis. The system extracts all frequently occurring biological relations among a pair of biological concepts through text mining. A mined relation is associated to a membership value, which is proportional to its frequency of occurrence in the corpus and is termed a fuzzy biological relation. The fuzzy biological relations extracted from a text corpus along with other relevant information components like biological entities occurring within a relation, are stored in a database. The database is integrated with a query-answering module. The query answering module has an interface, which guides users to formulate biological queries at different levels of specificity.

Keywords: Text mining; Ontology; Biological relation extraction; Biological query processing

1. Introduction

Since Molecular Biology has been a primary research area for the last two decades, the number of text documents disseminating knowledge in this area has gone up tremendously. The information is largely disseminated in textual format and is also available over the electronic media. The sheer enormity of the collection necessitates design of automated content analysis systems, without which, the assimilation of knowledge from this vast repository is becoming practically impossible [2]. The PUBMED database maintains a catalog of 12 million papers and receives hundreds of new papers every day [5]. Given a set of query terms, PUBMED can identify papers containing those terms quite efficiently. However, there is an increasing demand for Information Extraction (IE) systems, which can perform curation. Curation is the process of

[#] **To whom correspondence should be addressed**
Tel. No. +91-11-26591487
Tele-Fax: +91-11-26581005

extracting relevant information components from text documents and automatic construction of knowledge bases. Curation helps in efficient and intelligent retrieval of relevant information from online journal collections [1].

Intelligent retrieval requires analyzing the contextual relationship among query terms and judging the relevance of a document or a portion of a document with respect to a query in the perspective of this relationship. For example, a simple information need in the biological domain may be expressed as *"List all those documents that contain any information about NF-Kappa B"*. A simple pattern-matching technique is sufficient to decide whether a document is relevant to the query or not, based on the occurrence of the term in the document. However, a more complex query in this domain can be framed as "*List all those documents that contain information about any protein molecule that activates NF-Kappa B.*" Four sentences from MEDLINE abstracts, judged as highly relevant using simple pattern matching are shown in the upper stub of Table 1. It can be observed from Table 1 that two of these retrieved from Medline abstract 97032774, are not relevant, since "inorganic lead" is not a protein molecule. But this cannot be judged through simple pattern matching. Moreover, simple pattern matching judges t the relevant sentence shown in the bottom stub of Table 1, as less relevant than the earlier ones. This is not correct since this is a perfect answer to the second query. Machine interpretation of such relevance however is not a straightforward task, since such interpretation requires pattern matching to be enriched with text analysis and biological knowledge.

Table 1. Result of simple pattern matching.

	Medline No.	**Medline Sentence**	**Correctness**
Medline sentences judged relevant by simple pattern matching	MEDLINE: 95190988	LMP-1 *activates NF-kappa B* by targeting the inhibitory molecule I kappa B alpha.	Correct
	MEDLINE: 97423508	Surfactant protein A *activates NF-kappa B* in the THP-1 monocytic cell line.	Correct
	MEDLINE: 97032774	Inorganic lead *activates NF-kappa B* in primary human CD4+ T lymphocytes.	Incorrect
	MEDLINE: 97032774	We demonstrate that Pb at physiologically relevant concentrations *activates NF-kappa B* in primary human CD4+ T lymphocytes.	Incorrect
Missed Medline sentence	MEDLINE: 99252157	A20 can be regulated by the *NF-kappa B* transcription factor, which is known to be *activated by* the EBV LMP-1 protein.	Relevance not correctly judged

This work is motivated by the urge to develop a system which can answer complex queries like the one stated above. The aim is to retrieve all sentences that contain a set of biological concepts stated in a query, in the same context as specified in the query. A system designed to handle the problem of focused information retrieval will have to identify not just patterns but complex information components consisting of patterns and inter-relationships from the documents to judge their relevance to queries. Since text documents are usually unstructured or semi-structured in nature, it is essential that natural language understanding principles be also incorporated for extracting these information components. Recent efforts at consolidating biological and clinical knowledge in the structured form of ontologies [18,22,29], have made the task of locating and annotating biological concepts in text documents relatively easy to handle. A significant amount of research has been directed towards recognizing Biological entities from texts, some of which we shall be discussing in the next section.

In this paper, we have proposed the design of an intelligent Biological Information Extraction and Query Answering (BIEQA) system that identifies relevant portions of text which contain user-specified concepts or entities in a given context. The primary focus of the work lies in locating complex biological information components in text, where an information component comprises of biological concepts or entities and biological relations. However, unlike most of the other systems proposed earlier [25,27] which assume a list of biological relations that can be located in text, our work focuses on mining these relations from a document collection without any prior knowledge about their occurrences within the text. The mined information components are exploited to answer user queries. Sentences in a document, which are found to be relevant for a given query, are retrieved along with their Medline references. The user can look up the entire document separately.

The design of our system exploits ontology-based pattern matching techniques to locate relevant concepts within text and Natural Language Processing (NLP) principles to analyze the inter-relationships between these concepts. The biological relationships that are mined are different from the taxonomical or partonomical relations defined in the underlying biological ontologies. These relations may embody any kind of action or interaction among a pair of biological substances and/ or their locations. Within the text, these may express research findings about possible interaction between two types of biological substances, about chemical reactions in biological substances, or about localization of certain biological activities etc. We have employed deep text-mining principles to locate and identify the frequently occurring biological interactions among various biological entities. Initial reports about the text-mining process used for this work was reported in [16]. In [17], we had reported a mechanism to identify feasible biological relations. Query answering mechanisms including entity and relation-based indexing were not addressed in either of them. The design and processing details of the complete system is presented in this paper along with a performance analysis of the complete system.

The system helps in consolidating knowledge about the feasibility of various kinds of biological interactions among biological entities. The results presented in this paper have been verified through cross-validation over the source collection. The design of the entire system has been validated with experiments conducted over Medline abstracts, available as a part of GENIA corpus 3.01 [10]. The corpus contains 2000 MEDLINE abstracts which are manually tagged according to the GENIA[1] ontology. The input to our system is comprised of these tagged abstracts.

The unique aspects of the proposed BIEQA system are as follows:

- We have proposed a unified approach that integrates NLP and pattern mining techniques to identify all feasible biological relations. Unlike most of the related work [8,19,26], which have described methods for mining a fixed set of biological relations occurring with a set of predefined tags, the proposed system identifies all verbs in a document, and then identifies the feasible biological relational verbs using contextual analysis. The system has been designed to work with a collection of tagged abstracts along with the underlying ontology as

[1] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/topics/Corpus/genia-ontology.html

input. It initiates pattern mining with the set of ontological concepts to extract biological relations among them.

- The extracted binary relations are represented in the form of a triplet <*Left actor, Relation, Right actor*>, where left and right actors are biological entities. All relation triplets are associated to a membership value, which is directly proportional to the frequency of occurrence of that class of relation. The membership value of a relation triplet can be considered for enhancing the underlying ontology to a *fuzzy ontology,* which can encode inter-concept relation of varying strengths [28]. Preliminary reports about the relations extracted appeared in [16].
- Relations with membership values greater than a user-specified threshold are termed as feasible fuzzy biological relations and are stored in a structured format. These relations are thereafter used to allow users to formulate intelligent queries at various levels of specificity.
- The system stores all other relevant information like biological entity names and their tags, their occurrences etc. in a structured knowledge base, which is managed efficiently through novel indexing mechanisms. User queries related with biological entities, generic concepts and relationships among them, are processed over this structured knowledge base. Relevant portions of documents containing the exact information, along with the MEDLINE abstract number are returned to the user.

The rest of the paper is organized as follows. We review some related works on biological information extraction from text documents in section 2. The architecture of the proposed BIEQA system is given in section 3. Functional and design details of individual components of the BIEQA system are presented through sections 4 to 8. A complete performance analysis of the system, along with the details of the evaluation procedure, is presented in section 9. Finally, in section 10, we conclude with a summary and direction for possible enhancements to the proposed system.

2. Related work on biological document analysis

In this section we present an overview of some of the recent research efforts that have been directed towards the problems of extraction of biological entities, biological interactions between them, and query-answering from unstructured text documents. We start with a review on related work in biological entity recognition in section 2.1. Section 2.2 highlights the various approaches to biological relation identification that have been exploited in the past. In section 2.3 we have discussed some of the biological query-answering systems that have gained popularity.

2.1. Related work on named entity recognition from Biological texts

Information extraction from biological documents is largely dependent on the correct identification of biological entities in the documents. These entities are then tagged or annotated for more meaning information extraction. Initially the process of named entity recognition and their tagging were done manually. But the sheer volume of texts arriving everyday has initiated a significant research effort towards automated identification of biological entities in journal articles and tagging them. The various approaches followed can be classified as follows:

- *Rule-based approach*: The systems based on this approach look at the morphological characteristics of names and use Parts-Of-Speech (POS) information and keywords to discover and tag entity names. Fukuda *et al.* [13] have proposed a method called PROtein Proper-noun phrase Extracting Rules (PROPER) to extract material names from sentences using surface clue on character strings in medical and biological documents. PROPER identifies protein names in articles with a recall value of 98.8% and a precision of 94.7%.

- *Dictionary-based approach*: The systems based on this approach identify gene or protein names by matching them to dictionary entries and tag them appropriately. Proux *et al.* [7] identified non-English words in a document as gene terms. Dictionary-based systems can also be built to look up standard knowledge sources like GeneBank[2], PDB[3]etc. for identification of entities. However, a dictionary may not contain recently introduced names and may not cover all spelling variations of gene and protein names.

- *Machine-learning based approach*: Machine-learning based techniques like Hidden Markov Model [20], Naïve Bayes [4] and Support Vector Machines (SVMs) [11] have been successfully applied to identify and classify gene/ protein names in text documents. HMM-based method [20] was reasonably successful, where the HMM is trained using bigrammes from training documents. The features describing the words are mainly character-based - digit, symbol, punctuation mark, etc. No domain knowledge and linguistic-based feature is used in [20]. HMM post-processing corrects tags by comparing the tags of the different occurrences of the same word through the corpus, thereby increasing the accuracy. On application of this technique on 100 abstracts from GENIA corpus, [20] reports best results for correct identification of protein names as 76% and gene names as 47%. These are also found to be the most frequently occurring tags. Nobata *et al.* [4] compares the performance of Naïve Bayesian (NB) approach to Decision Trees (DT) on the same 100 abstracts mentioned earlier, using term lists and typed head nouns and chunking (shallow parsing). The NB method performs better on gene names (84%) while the DT method yields better results on protein names (85%) and other categories. Su *et al.* [14] have proposed a corpus-based approach for automatic compound extraction, which considers only bigrammes and trigrams. However, there are instances of medical and biological entities containing up to seven words, and bigram or trigram based approaches fail in such situations. Kazama *et al.* [11] presents an application of SVMs to the task of Named Entity (NE) recognition in the GENIA corpus. The class of non-entity words of the corpus is split according to the POS tag information in order to make learning by SVMs tractable and this improves the accuracy.

- *Hybrid approach*: Hybrid approaches have combined dictionary-based and rule-based approaches for multi-word gene/ protein name recognition. Hanish *et al.* [6] proposes a hybrid approach including the use of dictionaries and hand-coded rules in combination with Robust Linear Programming (RLP) based optimization. Though results obtained are encouraging, but the problem of non-specific synonyms is not fully solved. Rindflesch *et al.* [24] have developed a system named ARBITER that uses dictionaries and rules to identify binding terms in biomedical texts with a recall value of 72% and precision of 79%.

- *Statistical analysis*: Statistical analyses aims at clustering abstracts for keyword identification [15]. Term identification and classification methods based on statistical learning can

[2] http://www.ncbi.nlm.nih.gov/Genbank/

[3] http://pmd.ddbj.nig.ac.jp/

generalize to handle new knowledge types and representations more effectively than the methods based on dictionaries and hand-constructed heuristic rules [19].

2.2. Earlier work on extraction of biological relations from text documents

Though, named-entity recognition from biological text documents has gained reasonable success, reasoning about contents of a text document however needs more than identification of the entities present in it. Context of the entities in a document can be inferred from an analysis of the inter-entity relations present in the document. Hence, it is important that the relationships among the biological entities present in a text are also extracted and interpreted correctly. We present an overview of some of the earlier works reported in this area.

- *Co-occurrence based approach:* In this approach, after the biological entities are extracted from a document, relations among them are inferred based on the assumption that two entities in the same sentence or abstract are related. Negation in the text is not taken into account. Jenssen *et al.* [25] collected a set of almost 14,000 gene names from publicly available databases and used them to search Medline. Two genes were assumed to be linked if they appeared in the same abstract; the relation received a higher weight if the gene pair appeared in multiple abstracts. For the pairs with high weights i.e. with five or more occurrences of the pair, it was reported that 71% of the gene pairs were indeed related. However, the primary focus of the work is to extract related gene pairs rather than studying the nature of these relations.
- *Linguistics-based approach:* In this approach, usually shallow parsing techniques are employed to locate a set of handpicked verbs or nouns. Rules are specifically developed to extract the surrounding words of these predefined terms and to format them as relations. As with the co-occurrence based approach, negation in sentences is usually ignored. Sekimizu *et al.* [27] collected the most frequently occurring verbs in a collection of abstracts and developed partial and shallow parsing techniques to find the verb's subject and objects. The estimated precision of inferring relations is about 71%. Thomas *et al.* [12] modified a preexisting parser based on cascaded finite state machines to fill templates with information on protein interactions for three verbs – *interact with, associate with, bind to*. They calculated recall and precision in four different manners for three samples of abstracts. The recall values ranged from 24 to 63% and precision from 60 to 81%. The PASTA system is a more comprehensive system that extracts relations between proteins, species and residues [23]. This system fills templates representing relations among these three types of elements. This work reports precision of 82% and a recall value of 84% for recognition and classification of the terms, and 68% recall and 65% precision for completion of templates. Craven and Kumlien [19] have proposed identification of possible drug-interaction relations between protein and chemicals using a bag of words approach applied at the sentence level. This produces inferences of the type: *drug-interactions (protein, pharmacologic-agent)*, which specify an interaction between an agent and a protein. Ono *et al.* [26] reports a method for extraction of *protein-protein interactions* based on a combination of syntactic patterns. They employ a dictionary look-up approach to identify proteins in the document to analyze, and then select sentences that contain at least two proteins, which are then parsed with POS matching rules. The rules are triggered by a set of keywords, which are frequently used to name protein interactions (e.g., *'associate', 'bind'* etc.). Rinaldi *et al.* [8] have proposed an approach towards automatic extraction of a pre-defined set of seven relations in the domain

of Molecular Biology, based on a complete syntactic analysis of an existing corpus. They extract relevant relations from a domain corpus based on full parsing of the documents and a set of rules that map syntactic structures into the relevant relations. Friedman *et al.* [3] have developed a natural-language processing system, GENIES, for the extraction of molecular pathways from journal articles. GENIES uses the MedLEE parser to retrieve target structures from full-text articles. GENIES identifies a predefined set of verbs using templates for each one of these, which are encoded as a set of rules. This work [3] reports a precision of 96% for identifying relations between biological molecules from full-text articles.

It can be observed that most of the systems discussed above have been developed to extract a pre-defined set of relations. These relations have been chosen since they occur very frequently in Biological documents. However, each system is tuned to work with a particular set of relations and does not address the problem of relation extraction in a generic way. For example the method of identification of interaction between genes and gene products cannot work for extraction of enzyme interactions from journal articles, or automatic extraction of protein interactions from scientific abstracts, since tags will differ.

2.3. Review of biological information extraction systems

While entity recognition and tagging are back-end jobs for biological information retrieval, the front end comprises of a query-answering system. A number of biological information retrieval systems have been designed specially for extracting information from MEDLINE abstracts.

Textpresso [9] is an ontology-based biological information retrieval system. Textpresso analyzes tagged biological documents. Two types of tags are used for tagging text elements manually. The first set of tags defines a collection of biological concepts and the second set of tags defines a set of relations that can relate two categories of biological concepts. A tag is defined by a collection of terms including nouns, verbs etc. that can be commonly associated to the concept. Portions of the document containing a relevant subset of terms are marked by the corresponding biological concept or relation tag. The search engine allows the user to search for combinations of concepts, keywords and relations. With specific relations like commonly occurring gene-gene interactions etc. encoded as a relation tag, Textpresso enables the user to formulate semantic queries. The recall value of the system is reported to vary from 45% to 95%, depending on whether the search is conducted over abstracts or full text documents.

Uramoto *et al.* [21] have proposed a text-mining system, MedTAKMI, for knowledge discovery from biomedical documents. The system dynamically and interactively mines a large collection of documents with biomedically motivated categories to obtain characteristic information from them. The MedTAKMI system performs entity extraction using dictionary lookup from a collection of two million biomedical entities, which are then used along with their associated category names to search for documents that contain keywords belonging to specific categories. Users can submit a query and receive a document collection in which each document contains the query keywords or their synonyms. The system also uses syntactic information with a shallow parser to extract binary (a verb and a noun) and ternary (two nouns and a verb) relations that are used as keywords by various MedTAKMI mining functions like keyword-based and full text searching, hierarchical category viewer, chronological viewer, etc.

Our system is based on a hybrid approach that combines co-occurrence based and linguistics-based techniques for mining biological information from text documents. Linguistic analysis is employed in conjunction with co-occurrence based principles, to extract feasible biological relational verbs along with their morphological variants that are present in the corpus. Co-occurrence is also used to compute the strengths of the extracted relations. Unlike the other works mentioned in this section, no prior knowledge is used to identify biological relational verbs. The abstract collection is indexed on biological concepts and entities and also the extracted relation set, to enable efficient query answering.

3. System architecture

We now present the complete architecture of the proposed BIEQA system, which is designed to mine information about biological relations relating two biological concepts, when the information is embedded within free form but ontologically tagged text. The extracted relations are used to assist users in extracting information from text documents in a more efficient way. Our system is characterized by three key functionalities:

- *Extraction of information components from texts* – The input to the system comprises of biological abstracts in which biological entities have been tagged with their respective ontology tags. The system is designed to extract *information components* from these, where an information component comprises of two entities, their tags and a relation binding these two entities in a document.
- *Compiling a collection of feasible inter-concept relations* – The *information components* extracted from the documents are subjected to feasibility analysis. All feasible biological relations are stored in a structured knowledge-base. These relations are also used to index the document collection.
- *Query Answering* – The system is equipped with a query answering mechanism which accepts user queries and retrieves relevant sentences from the document collection. User queries are constructed in a guided manner and may contain both simple and complex information requirements ranging from requirement specification about presence of an entity name, to presence of a pair of entities or concepts bound by a particular biological relation.

Though the system design is fairly generic, all experiments and evaluation has been conducted on the GENIA corpus [10,29]. Hence we first give a brief overview of the GENIA ontology and the corpus.

The GENIA ontology was proposed by Tateisi *et al.* [10,29]. This ontology used substances and sources (substance locations) as a base to fix the class of molecular biological entities and relationships among them. GENIA ontology stores knowledge about Molecular Biology substances and their locations in a structured format and has been widely accepted as a baseline for categorizing Molecular Biological concepts. In GENIA, substances are classified according to their chemical characteristics rather than their biological roles. The substances are classified into families, complexes, individual molecules, subunits, domains and regions. The sources are classified into natural and cultured sources that are further classified as an organism (human), a tissue (liver), a cell (leukocyte), a sub-location of a cell (membrane or a cultured cell line (HeLa)). Organisms are further classified into multi-cell organisms, virus, and mono-cell organisms other than virus. The GENIA corpus also maintained by the same group, contains

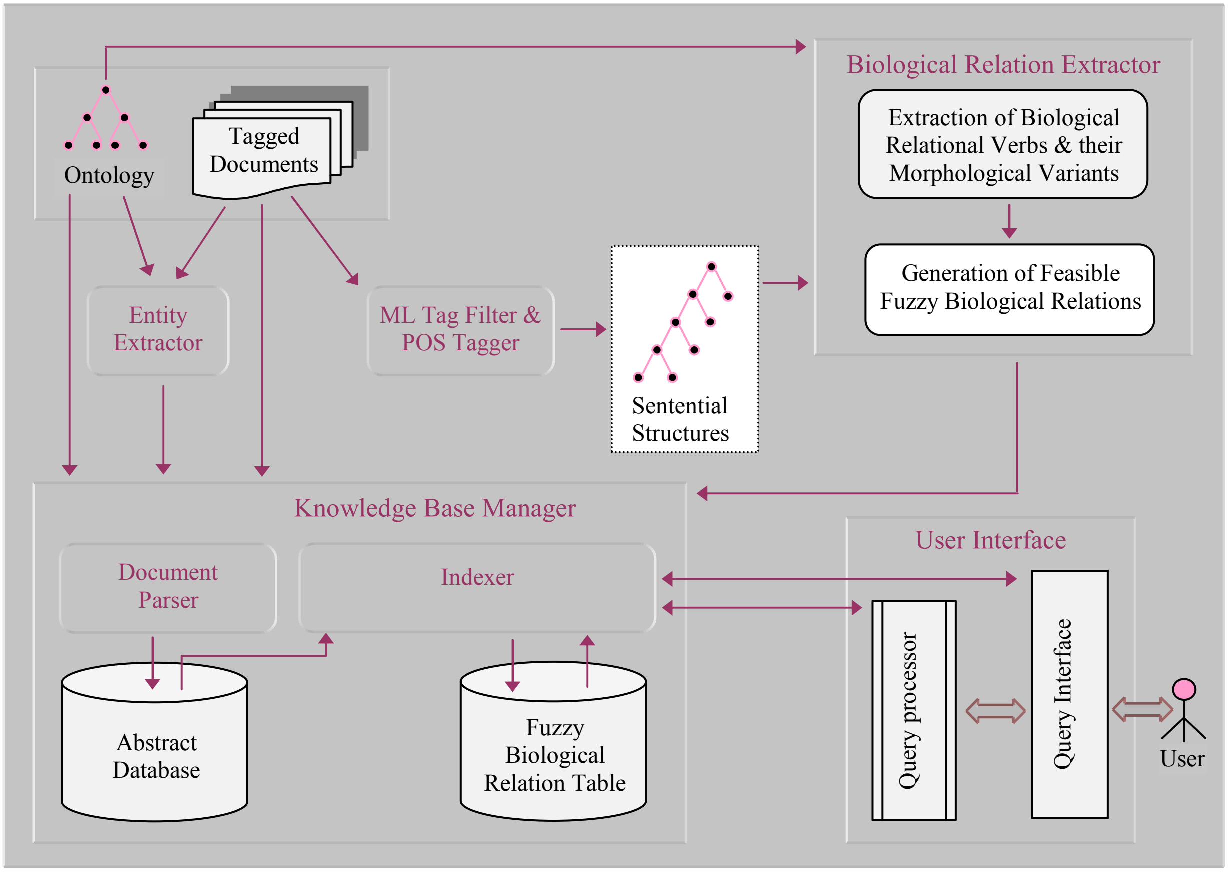

Biological Relation Extractor
Extraction of Biological Relational Verbs & their Morphological Variants
Generation of Feasible Fuzzy Biological Relations
Ontology
Tagged Documents
Entity Extractor
ML Tag Filter & POS Tagger
Sentential Structures
Knowledge Base Manager
Document Parser
Indexer
Abstract Database
Fuzzy Biological Relation Table
User Interface
Query processor
Query Interface
User

Fig. 1. BIEQA system architecture.

2000 MEDLINE abstracts which have been manually annotated according to the GENIA ontology. Each abstract is also annotated with Meta Language tags to identify individual documents, sentences, titles etc.

Fig. 1 presents the complete architecture of the system, which comprises of five main modules. An overview of each module and their overall interactions is presented in this section.

- *Entity Extractor* –This module accepts tagged biological abstracts as input and extracts entity names from the text. Since entities may be embedded within single as well as nested tags, we have designed a set of preprocessing rules to extract biological entities from the tagged text documents. The module is equipped to handle extraction of individual entity names from complex combinations which occur in conjunction with special characters like forward slash (/), hyphen (-) etc. We have also incorporated rules to handle special words like "and", "or" etc. occurring within entity names. Such occurrences indicate the presence of multiple entities in a concise form. The detailed design of this module is presented in section 4.
- *Meta Language (ML) Tag filter & Parts-Of-Speech (POS) Tagger* – The function of this module is to filter the ML tags from input documents. Since the occurrences of entities present in the documents are already noted by the first module, during the filtering process, the ontological tags associated to these entities are retained, while the actual entity names are filtered out along with the ML tags. The filtered documents are passed through a POS tagger that assigns parts-of-speech to different words. POS tagged sentences are thereafter parsed to

extract grammatical and entity-relation association information from them. Based on its parse structure, each sentence is stored as a record which instantiates a pre-defined tree structure to store the various components of the sentence.

- *Biological Relation Extractor* – This module uses the tree-structured records generated by the earlier module in collaboration with the underlying ontology to extract biological relations. A biological relation is characterized by a root verb and its morphological variants. The system employs statistical techniques along with shallow parsing to mine binary relations defined by relational verbs co-occurring with two biological entities within a sentence. These relations are stored as triplets in the form of <*left actor (=entity + tag), biological relation, right actor (=entity + tag)*>. This relation triplet is also termed as an *information component*. The same relational verb may be associated to multiple ontological tag- pairs. Hence each relation is associated with a strength value. Strength reflects the degree of co-occurrence of a pair of ontology tags with a particular relational verb. The relations with strength greater than a user-specified threshold are thereafter maintained as *feasible fuzzy biological relations* in the knowledge base.
- *Knowledge Base Manager* –The Knowledge Base Manager maintains the abstract collection. It stores information about the occurrence of each feasible relation triplet in the collection. It contains a d*ocument parser* which locates *feasible information components* within sentences in the abstracts. The knowledge-base manager associates with each information component the Medline number and the sentence number in which it occurs. This module indexes the document collection using feasible relations and entities. The collection is used by the query processor to create an intelligent query interface and also to answer queries efficiently.
- *Query Interface* – The relation triplets are utilized to build a query-interface through which users can pose queries at multiple levels of specificity. User interface guides the user to formulate feasible queries and displays the relevant results. At the back end of the interface is a query parser, which interacts with the knowledge base manager to retrieve relevant portions of documents from the knowledge base.

The design and working principles of the different components are explained in the following sections.

4. Entity extractor

Since the proposed system works with tagged biological abstracts in which entities are tagged according to the GENIA ontology, the entity extractor is designed to recognize simple and complex entities that occur within these tags. This is not a general-purpose entity recognizer. The focus is to identify entities correctly from both simple and nested tags. For example, given a tagged sentence like "*We demonstrate, through the deletion of the <cons sem="G#DNA_domain_or_region">human <cons sem="G#protein_ molecule"> UDG </cons> promoter sequences</cons>, that expression …",* our aim is to extract the entities *"UDG"* and *"human UDG promoter sequences"* along with their ontological tags *protein molecule* and *DNA_domain_or_region* respectively. While these are simple extraction tasks, some more complex tasks include extraction of individual entity instances from complex representations.

Rules

1. <W>-<S><D> → <W>-<D>
2. <w1><S>(<w2>)-<D> → <w_1>-<D><S>(<w_2>-<D>)
3. <W><D>/<D> → <W><D>/<S><W><D>
4. <w_1>-, <w_2>-,.., and <w_n>-<w_m> → <w_1>-<w_m>, <w_2><w_m>,.., and <w_n><w_m>
5. <w_1>-, <w_2>-,...,and <w_n>-<w_m> → <w_1>-<w_m>, <w_2><w_m>,…, or <w_n><w_m>
6. <w_m>-<w_1>, -<w_2>,.., and -<w_n> → <w_m>-<w_1>, <w_m>-<w_2>,.., and <w_m><w_n>
7. <w_m>-<w_1>, -<w_2>,..., or -<w_n> → <w_m>-<w_1>, <w_m>-<w_2>,..., or <w_m><w_n>
8. <w_1>-<w_2> AND <w_2>: induc*, activat*, bind*, inhibit*, regulat*, transcri*, phospho*, suppress*, indepen*, specifi* → <w_1><S>-<S><w_2>

Examples

1. interleukin- 5 → interleukin-5
2. interleukin (IL)-1 beta → interleukin-1 (IL-1) beta
3. gp350/220 → gp350/ gp220
4. B- and T-cell → B-cell and T-cell
5. B- or T-cell → B-cell or T-cell
6. PKC-alpha, -epsilon, and –zeta → PKC-alpha, PKC-epsilon, and PKC-zeta
7. PKC-alpha, -epsilon, or –zeta → PKC-alpha, PKC-epsilon, or PKC-zeta
8. alpha-activated → alpha - activated

Legend

<W> = (a-z, A-Z)$^+$; $w = w_1 + w_2 + \ldots + w_n \in W$; <D> = (0-9)$^+$;
<S> = {White space characters}

Fig. 2. Preprocessing rules.

Gavrilis *et al.* [5] have proposed a rule set for pre-processing biological documents to extract entities. The proposed entity extractor enhances this rule set and also includes some new rules to identify entities correctly. The extractor is specially equipped to handle extraction of individual entity names from complex combinations which occur in conjunction with special characters like forward slash (/), hyphen (-) etc. It also handles the use of connectors like "and", "or" etc. which usually describe multiple entities in a concise form. The rules extract individual instances from concatenations and commonly used abbreviations. For example, concatenated instances like *B- and T-cell* or *gp350/220* are converted into entity names *B-cell* and *T-cell*, and *gp350* and *gp250* respectively. Fig. 2 illustrates the complete set of rules implemented by our entity extractor to locate and extract entity names from tagged documents. The improper use of white space characters is also handled by our rule set. The rule set was identified after analysing the tagged MEDLINE abstracts of the GENIA corpus. This mechanism extracts entities from this corpus with 100% accuracy.

5. Meta language tag filter and parts of speech tagger

The function of this module is to enable extraction of information components from text documents. The tagged abstracts contain Meta Language tags like <*cons*>, <*title*> etc. To begin with, all ML tags are stripped off from these documents. Stop words are also removed. We have used a sub-set of the stop words used by PubMed database. The prepositions have been purposefully retained, since they contain important information about relations and the entities they bind.

During the filtering process, each sentence is transformed to an equivalent form, in which the ontological tags representing biological concept names are retained, while the tagged entities are filtered out. The transformed sentences are subjected to POS analysis, whereby each word is assigned a parts-of-speech. We have used a web-based Tagger that has been developed by the specialized information services division of the National Library of Medicine[4]. Finally, every sentence and thereby a complete abstract is converted into a binary tree structure which is recursively defined as follows:

```
Struct Sentence
Begin
        string  Root;
        struct sentence * Lchild;
        struct sentence * Rchild;
End
Struct Sentence Document [number_of_sentence];
```

The contents of a sentence are stored in the tree by distributing the POS tags in the following way:

Root (R): contains the right most verb of the sentence
Lchild (L_c): points to the sub-tree representing sentence segment that is to the left of the word stored at R.
Rchild (R_c): points to the sub-tree representing sentence segment that is to the right of the word stored at R.

The root of each sub-tree contains the right-most verb in the corresponding sentence-segment. The process of segmentation continues until no more verbs are found in the segment.

The equivalent context-free grammar for the scheme is given as follows:

Document (D) $\rightarrow S^*$

$S \rightarrow L_c R R_c \mid \in$

$L_c \rightarrow L_c R R_c \mid (E+N+A+J+R)^* \mid \in$

$R_c \rightarrow (E+N+A+J+R)^*$

$R \rightarrow V$

where, S represents a sentence, and N, A, J, R and V represent Noun, Adverb, Adjective, Preposition, and Verb respectively, as assigned by the POS tagger. The POS tagger assigns a POS tag to each Biological tag also. However, the system ignores the POS tag assigned and replaces it by symbol E, to indicate the positions of entities.

Fig. 3 shows a sample tagged sentence picked up from the GENIA corpus. The filtered sentence along with the POS tags assigned by the POS tagger to different words is shown on bottom left. The resulting tree structure that is created is shown in bottom right of Fig. 3. This structure encodes all relevant information, which can be effectively exploited by the Biological relation extractor whose working principle is elaborated in the next section.

[4] http://tamas.nlm.nih.gov

We demonstrate, through the deletion of the <cons sem="G#DNA_domain_or_region">human <cons sem="G#protein_ molecule"> UDG </cons> promoter sequences</cons>, that expression of <cons sem="G#protein_molecule">E2F-1 </cons> activates the <cons sem= "G# DNA_domain_or_region"><cons sem="G#protein_molecule">UDG</cons> promoter</cons> through several <cons sem ="G# DNA_domain_or_region"><cons sem="G#protein_family_or_ group ">E2F</cons> sites</cons>.

```
demonstrate, through  deletion of
    V            R        N     R
DNA-domain-or-region expression of
      E                       N      R
protein-molecule activates DNA-domain-or-
      E                V            E
region through DNA-domain-or-region.
         R                 E
```

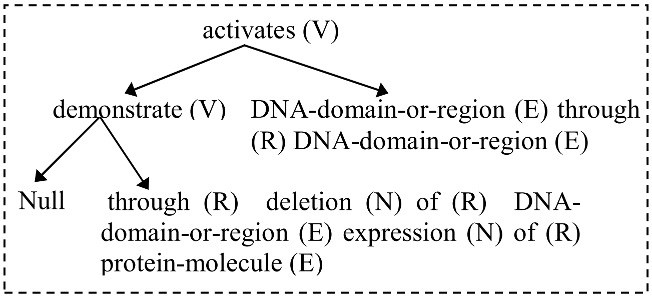

activates (V)
demonstrate (V)
DNA-domain-or-region (E) through (R) DNA-domain-or-region (E)
Null
through (R) deletion (N) of (R) DNA-domain-or-region (E) expression (N) of (R) protein-molecule (E)

Fig. 3. A sentence from MEDLINE: 95197524, it's stripped and POS tagged form and generated Binary tree.

6. Biological relation extractor

A biological relation is assumed to be binary in nature, which associates two biological entities. As already defined in section 3, an information component arising out of a biological relation is a triplet represented as *<left actor (=entity + tag), biological relation, right actor (=entity + tag)>.* The process of identifying biological relations is accomplished in two stages. During the first phase, prospective information components which might embed biological relations within them are identified from the sentences. During the second stage, a feasibility analysis is employed to identify correct biological relations. These stages are elaborated further in the following sub-sections.

6.1. Extraction of information components

A biological relation is usually manifested in a document as a relational verb. The biological actors associated to a relation can be inferred from the biological entities located in the proximity of the relational verb. Since the entities are ontologically tagged, this module exploits the ontological tags and the POS tags assigned by the POS Tagger, to identify relevant information components.

A biological relation is characterized by verb. A verb may occur in a sentence in its root form or as a variant of it. Different classes of variants of a relational verb are recognized by our system. The first of this class comprises of *morphological variants* of the root verb, which are essentially modifications of the root verb itself. In English language the word *morphology* is usually categorized into "inflectional" and "derivational" morphology. *Inflectional morphology* studies the transformation of words for which the root form only changes, keeping the syntactic constraints invariable. For example, the root verb *"activate"*, has three inflectional verb forms: "*activates*", "*activated*" and "*activating*". *Derivational morphology* on the other hand deals with the transformation of the stem of a word to generate other words that retain the same concept but may have different syntactic roles. Thus, "*activate*" and "*activation*" refer to the concept of "making active", but one is a verb and the other one a noun. Similarly, *inactivate*, *transactivate*, *deactivate* etc. are derived morphological variants created with addition of pre-fixes. Presently the system does not consider derivational morphology, and only inflectional variants of a root verb are recognized.

```
Algorithm: Extract_Information_Components()
Input: Sentence Trees
Output: List of information components $L_{IC}$
Steps:
1    Initialize $L_{IC}$ with Null values
2    Ptr = ROOT    // start from root node
3    If (Ptr ≠ Null)     // If the tree is non-empty
4      Check_Surrounding_Tags(Ptr) // Search the sub-tree pointed by Ptr for
                                   // a possible left and right entity tags
5      If the node has surrounding entity tags
6         Store the node value along with surrounding tags into $L_{IC}$
7         Check_Preposition(Ptr->Rchild)
8         If a preposition found
9            Calculate its distance with the node value
10           If distance is 1, OR 2 with middle word as an adverb
11              Store the proposition into $L_{IC}$ under a separate field
12              Go to step 17
13           End If
14        End If
15        Ptr = Ptr -> Lchild // Consider the next sub-tree
16        Go to step 3 // Repeat the above steps for the next sub-tree
17     Else  // the node has no surrounding tags
18        Ptr = Ptr -> Lchild
19        Go to step 3
20     End if
21   End if
22   Stop
```

Fig. 4. Algorithm – Extract_Information_Components ( ).

In the context of biological relations, we also observe that the occurrence of a verb in conjunction with a preposition very often changes the nature of the verb. For example, the functions associated to the verb *activates* may be quite different from the ones that can be associated to the verb form *activates in,* in which the verb *activates* is followed by the preposition *in*. Thus our system also considers a third category of biological relations, which are combinations of *root verbs or their morphological variants, and prepositions* that follow these. Typical examples of biological relations identified in this category include "*activated in*", "*binds to*", "*stimulated with*" etc. This category of relation can take care of special biological interactions involving substances and sources or localizations. To recognize relations correctly, all prepositions at distance one or two from a relational verb are considered. This increases the accuracy of the system in identifying biological relations, since it has been found that very often the text is interjected with adverbs following the main verb. Using the proposed approach, the adverbs are eliminated from consideration since they do not play any role in the main biological interaction, rather is used by the author to emphasize on the strength of the associated biological verb. One such sample sentence is shown below, in which the biological relation to be identified is *expressed in*, though the words occur in the text separated by the adverb *exclusively.*

```
MEDLINE:95016436 - A family of <cons sem="G#protein_family_or_group">serine
proteases</cons> expressed exclusively in myelo-<cons sem="G#cell_type">
monocytic cells</cons> specifically processes the <cons sem="G#protein_
subunit">nuclear factor-kappa B subunit p65</cons> in vitro and may impair
human <cons sem="G#other_name"><cons sem="G#virus">immunodeficiency virus
</cons> replication</cons> in these cells.
```

The relation extractor module identifies all possible information components containing a relation and its associated actors by traversing the binary tree built earlier. The working principle of the Algorithm Extract_Information_Components is explained by the following steps:

- List of information components L_{IC} is initialized to *Null*.
- The binary sentential tree generated earlier is traversed in post-order fashion to locate and extract information components. Starting at the left-most leaf node, if both left and right siblings contain biological tags, the verb represented at the parent of these siblings is assumed to represent a biological relation. If the right child of a node contains a preposition within distance 1 or 2 from the node verb, then the preposition is associated to the verb in the parent node, and the verb-preposition pair is identified as a possible biological relation. If right child does not contain a preposition, only the verb which may be a root verb or an inflectional-variant of it, is identified as a possible biological relation. A unique combination of a possible biological relational verb identified this way along with the biological tags occurring in the neighborhood of these verbs, are added to list of information components L_{IC}. Function Extract_Information_Components(), shown in Fig. 4, summarizes this process formally.

The above process of considering only those verbs which co-occur with biological tags in their vicinities eliminates a large number of irrelevant verbs from being considered as biological relations. However, our aim is not just to identify possible relational verbs but to identify generic biological relations. Hence we engage in further statistical analysis to identify feasible relation triplets. After all possible information components are identified; the module performs a tag co-occurrence analysis, to eliminate the infeasible tag pairs.

6.2. Identifying feasible biological relations

Though a large number of commonly occurring verbs are eliminated by the earlier step, it is found that further processing is necessary to consolidate the final list of relations. During the consolidation process, we take care of two things. Firstly, since various forms of the same verb represent a basic biological relation in different forms, so the feasible collection is extracted by considering only the unique root forms after analyzing the complete list of information components. Again, each relation can occur in conjunction with multiple tag-pairs, while some tag pairs may not ever co-occur. Hence, in the second phase of feasibility study, all feasible triplet combinations are compiled. The core functionalities of the biological relation finding module are summed up in the following steps.

- Let L_{IC} be the collection of verbs or verb-preposition pairs, which are extracted as part of information components. L_{IC} is the collection of possible biological interactions. However, each verb can occur in more than one form in this list. For example, the verb "*activate*" may occur in the form of "*activate*, *activates*, *activated* or *activated in*" etc, all of them essentially representing the biological interaction "*activation*" in some form. Function

Find_Root_Verbs(), shown in Fig. 5, analyzes L_{IC} to determine the set of unique root forms from this collection. The frequency of occurrence of each root verb is the sum-total of its occurrence frequencies in each form. All root verbs with frequency less than a user-given threshold are eliminated from further consideration. The surviving verbs are termed as *most-frequently occurring* root verbs and represent important *biological relations*.

- Once the frequent root verb list is determined, function *Find_Morphological_Variants(),* shown in figure 6, operates on L_{IC} and identifies the complete list of all biological relation verbs including frequent root verbs, their morphological variants and their co-occurrence with prepositions.
- For each inferred biological relation verb form, the frequency of occurrence of each form in

```
Algorithm: Find_Root_Verbs()
Input: List $L_{IC}$ of information components
Output: List $L_{RV}$ of root verbs
Steps:
i ← 0  // initialize i with 0
While there are records in $L_{IC}$ do  // create a list $L_V$ of verbs
   $L_V$[i] ← $L_{IC}$[i].Verb
   i ← i+1  // consider next record
End While
i ← 0  // re-initialize i with 0
$L_{UV}$[i] ← $L_V$[i]  // copy first element of $L_V$ into $L_{UV}$
While there are values in $L_V$ do  // create a list $L_{UV}$ of unique verbs
   i ← i+1  // consider next element
   If $L_V$[i] is not in $L_{UV}$ then
      $L_{UV}$[i] ← $L_V$[i]
   End If
End While
Filter out verbs from $L_{UV}$ with a prefix as cross-, extra-, hydro-, micro-
, milli-, multi-, photo-, super-, trans-, anti-, down-, half-, hypo-,
mono-, omni-, over-, poly-, self-, semi-, tele-, dis-, epi-, mis-, non-,
pre-, sub-, de-, di-, il-, im-, ir-, un-, up-
Filter out verbs from $L_{UV}$ with a suffix as -able, -tion, -ness, -less, -
ment, -ally, -ity, -ism, -ous, -ing, -er, -or, -al, -ly, -ed, -es, -ts,
-gs, -ys, -ds, -ws, -ls, -rs, -ks, -en
Count the frequency of occurrences of the elements of $L_{UV}$ in $L_V$ and
arrange them in descending order of their frequencies. For frequency
count allow partial match to increase the frequency of an element
Plot a line chart by using the frequency counts of the verbs in $L_{UV}$ and
their rank order
Fix the cut-off value by observing the nature of the graph and create a
list $L_{RV}$ by retaining the verbs having frequency count greater than or
equal to the cut-off value
Stop
```

Fig. 5. Algorithm – Find_Root_Verbs ( ).

```
Algorithm: Find_Morphological_Variants()
Input: List $L_{RV}$ of root verbs; List $L_{IC}$ of information components
Output: List of Morphological variants $L_{MV}$
Steps:
i ← 0  // initialize i with 0
While there are records in $L_{IC}$ do
   MV ← $L_{IC}$[i].Verb
   Search_Root_Verb_List(MV, $L_{RV}$) // search $L_{RV}$ for a partial match of MV
   If MV has a match in $L_{RV}$ then
      $L_{MV}$[i] ← MV
   End If
   If $L_{IC}$[i].preposition <> Null then
      $L_{MV}$[i] ← $L_{MV}$[i] + $L_{IC}$[i].preposition
   End If
   i ← i+1 // consider next record of $L_{IC}$
End While
Stop
```

Fig. 6. Algorithm – Find_Morphological_Variants ( ).

conjunction with a unique biological tag pair is computed. A relation triplet is a unique combination of a particular variant of a biological relational verb and a pair of biological tags. At this point, only those triplets may be retained as *frequent* which have a frequency greater than a user-given threshold. In the present implementation, we have retained all relation triplets with non-zero frequency.

The frequently occurring relations extracted from a corpus may be considered for enhancing a known ontology, provided the strength of the association between a relation and its associated biological actors can be determined. Since a particular relation may occur with multiple tag pairs, we have computed the relative frequency of occurrence of a relation triplet to compute fuzzy membership of a relation. The relation along with the fuzzy membership may be considered while enhancing a given ontology to include a new biological relation between two entities.

6.3. Feasible Biological relations identified from the GENIA corpus

In this section we shall present the entire collection of biological relations identified from the GENIA corpus following the methodology described earlier. In order to check the consistency of the relative frequencies of various verbs and their variants, that can possibly denote biological interactions, we have employed the following cross-validation technique. The entire GENIA corpus consisting of 2000 Medline abstracts was divided randomly into three subsets – one containing 640 documents and the other two containing 680 documents each. The relation extraction process was applied on each subset separately. For each subset, the list of frequently-occurring root verbs was identified as those verbs which co-occur with two biological entities on both sides and have more than 15 occurrences in the corpus. It was observed that the root verb collection was fairly stable and the relative frequencies of the various root verbs extracted also remained identical in all the three corpus subsets. Hence, the relative frequencies of various root verbs can be concluded to reflect the actual distribution of various biological interactions in the corpus.

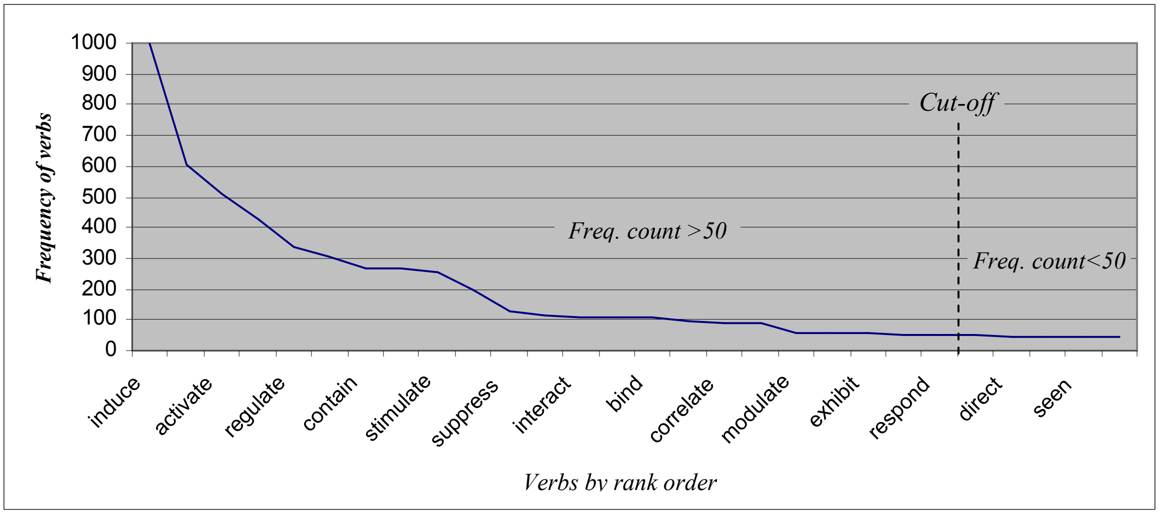


Fig. 7. A plot of the frequency of relational verbs occurrence and their rank order.

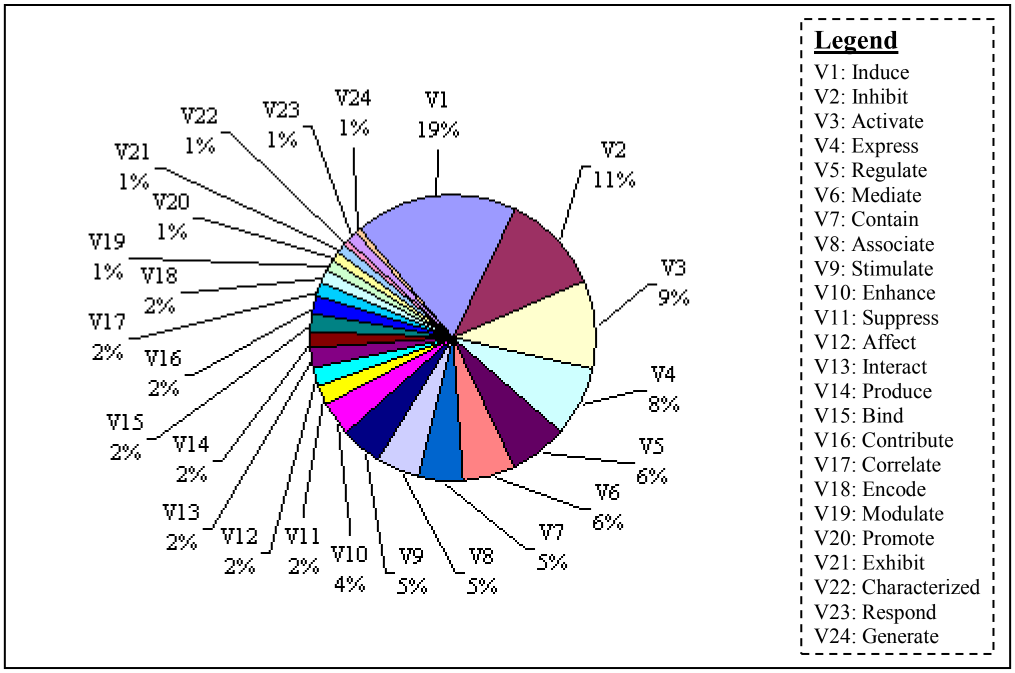


Fig. 8. Relative frequencies of occurrence of extracted root biological relational verbs.

Thereafter we compiled the frequent root-verb collection from the entire corpus and computed their relative frequency of occurrences. In order to retain only frequently occurring interaction verbs, a verb is termed as *frequent* and retained provided it has a minimum of 50 occurrences in the entire collection. Using this cut-off, the system extracted 24 root verbs as frequent from the entire corpus. Fig. 7 shows the overall distribution of the frequent verbs extracted from the entire collection. These verbs are identified as representative elements for various biological relations. It may be noted that the most frequent set of verbs and thereby the biological relations identified by our system, include those seven that were identified by Sekimizu and Tsujii [27] as relevant and were also used by Rinaldi *et al.* [8]. Fig. 8 represents the relative frequencies of the various biological relational verbs identified by the system from the entire corpus.

6.4. Relating Biological relations to Biological substances and locations

A biological relation expresses how two biological concepts interact. Hence, a stronger characterization may be achieved by analyzing the frequencies of each biological interaction identified earlier along with the biological tags it co-occurs with in the collection. The frequencies reflect the possibility of two types of entities co-occurring in a particular context and hence can provide insight into possible inter-actions between various concepts. The distribution details can also be used for efficient query processing.

Considering the frequent feasible relations in conjunction with valid tag pairs, we have mined 4162 unique biological relation triplets from the GENIA corpus. Each feasible biological relation is stored in the knowledge base as a quadruple of the form <T1, V, T2, S>, where T1 and T2 are ontology-defined tags, V denotes a biological relation (a verb or its morphological variant) and S is the strength of the relation. Obviously, two tags T1 and T2 may define two different tuples even with the same verb V between them, where their roles will be reversed and hence the strengths may not be same. For example, *T1 activates T2* does not imply that *T2 activates T1*. Thus <T1, V, T2, S> and <T2, V, T1, S'> are stored as different patterns. The strength S of a relation triplet is a function of the frequency of its occurrence along with a pair of tags. The relation quadruples with non-zero fuzzy membership values are called *feasible fuzzy biological relations* as already mentioned in section 1.

Fig. 9 shows some of the *feasible fuzzy relations* along with their membership values. This figure also illustrates the many-many relationship between a pair of biological tags and a feasible biological relation. On analysis of the relation triplets, it is observed that though there are 36×36 tag pairs that are feasible, a large number of tag pairs do not occur within the neighborhood of any frequent biological relation extracted earlier. This information can be used to restrict the

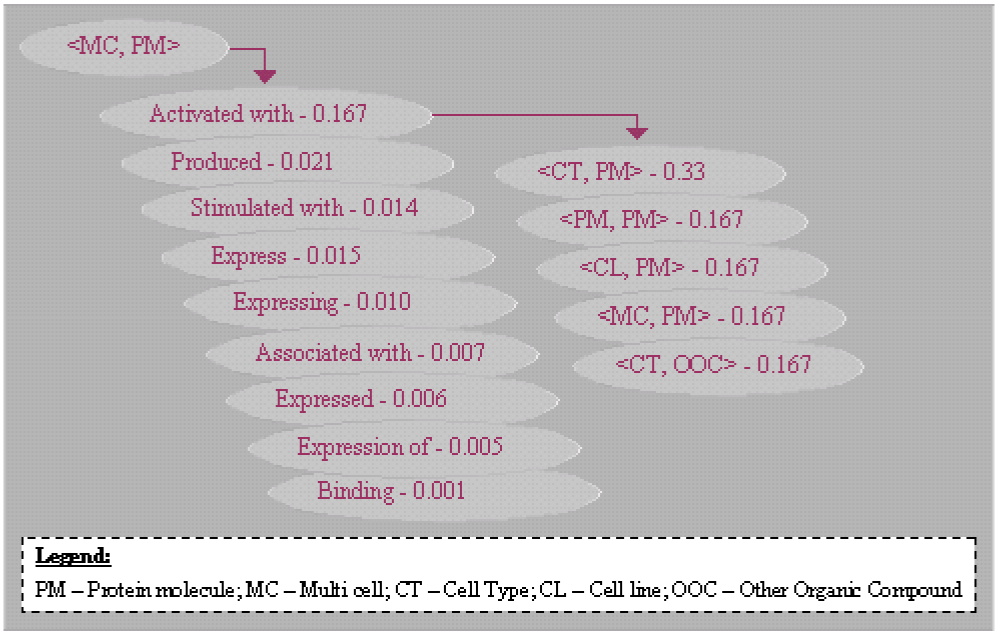


Fig. 9. Feasible Fuzzy Biological Relations along with their strength of associations with corresponding Tag-pairs.

users to formulate queries using valid tag pairs only. Table 2 shows a partial listing of tag pairs that do not co-occur in any sentence. Details about how this analysis was carried out, has been earlier reported in [16]. A complete list of all feasible relation triplets extracted by the BIEQA system from the GENIA corpus is available on http://www.geocities.com/mdabulaish/BIEQA/.

Table 2. A partial List of Non-related Biological Tag pairs in GENIA corpus 3.01

Ordered Tag Pairs	Ordered Tag Pairs
<protein_family_or_group, RNA_domain_or_region>	<protein_complex, RNA_substructure>
<cell_component, RNA_domain_or_region>	<Carbohydrate, Polynucleotide>
<DNA_family_or_group, Carbohydrate>	<Carbohydrate, Atom>
<Lipid, protein_substructure>	<protein_substructure, RNA_substructure>
<Nucleotide, Carbohydrate>	<Polynucleotide, RNA_domain_or_region>
<Tissue, DNA_substructure>	<RNA_substructure, protein_complex>
<amino_acid_monomer, DNA_substructure>	<RNA_substructure, protein_domain_or_region>
<amino_acid_monomer, RNA_domain_or_region>	<Virus, RNA_substructure>

Table 3. Fuzzy relation extraction statistics of the Biological Relation Extractor Module

Attribute	Value	Attribute	Value
# Ontology tags	36	# Root verbs along with morphological variants as biological relations	246
# Possible ordered tag pairs	1296	# possible <tag, relation, tag> triplets (taking only valid tag pairs)	258420
# Related tag pairs	1180	# Extracted Valid relation triplets having non-zero membership value	4162
# Selected root verbs	24		

The study of feasibility of various biological relations in the context of a pair of biological tags is one of the major contributions of the present work. The information extracted about the occurrences of biological relations in the corpus is exploited to design an efficient query answering system. The overall statistics of biological interactions extracted from the GENIA corpus 3.01 is summarized in Table 3.

6. Knowledge base manager

A primary goal of our system is to assist scientists to get information about relevant snippets of documents that contain information regarding biological interactions between different biological substances and/or their locations. Hence, the main function of this module is to create and maintain a knowledge base that contains sentences occurring in Medline abstracts indexed by the information components mined earlier. The knowledge base manager consists of a *document parser* that parses the Medline sentences for locating information components and stores them locally in an *abstract* database. In order to answer user queries efficiently, the knowledge base manager also consists of an *Indexer* that employs a data-cube like indexing scheme to index each sentence based on its information components. Documents are also indexed on the entities occurring in them, through a trie structure. Details of the *document parser* and *indexer* modules are given in the following subsections.

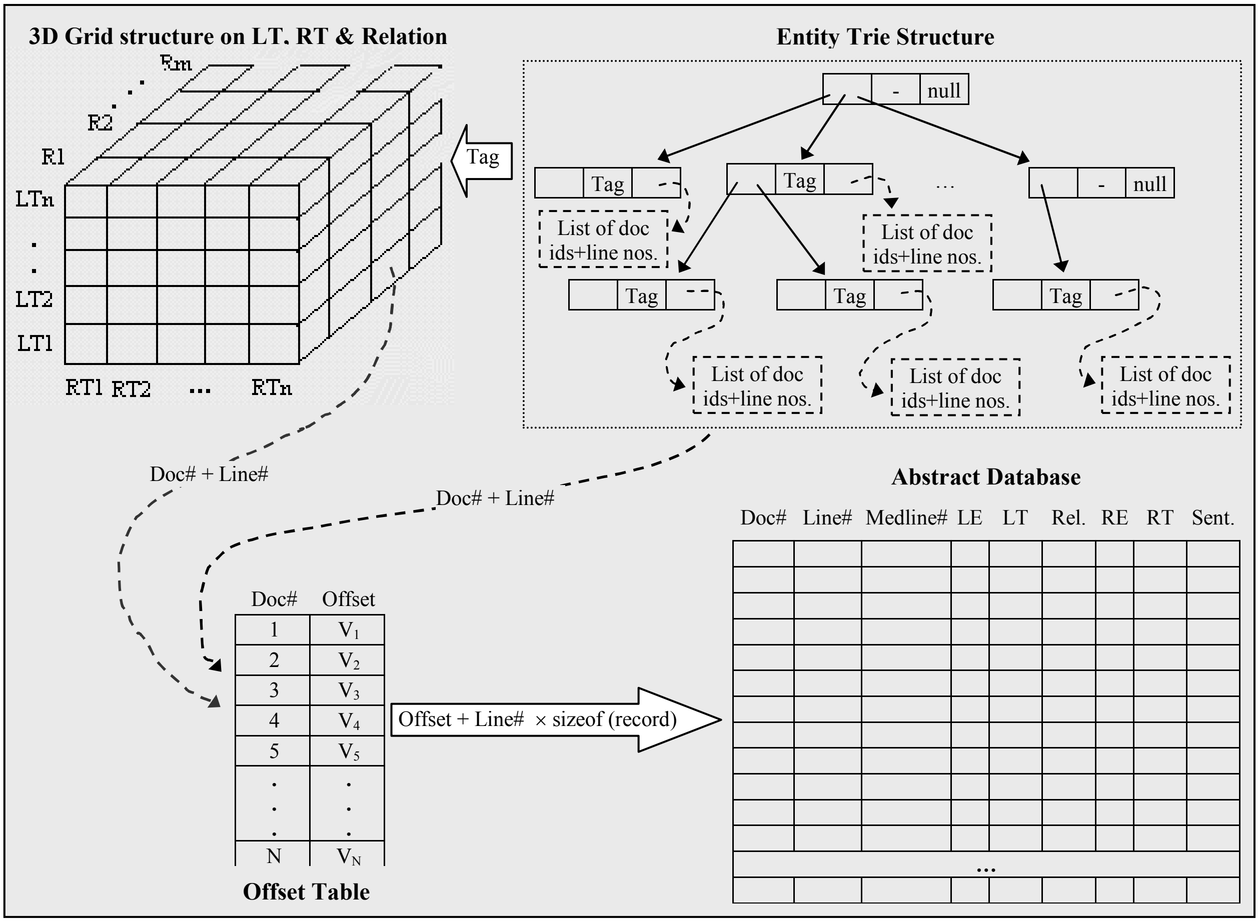

3D Grid structure on LT, RT & Relation
Rm
R2
R1
LTn
LT2
LT1
RT1 RT2 ... RTn
Tag
Entity Trie Structure
- null
Tag
Tag
- null
List of doc ids+line nos.
List of doc ids+line nos.
Tag
Tag
Tag
List of doc ids+line nos.
List of doc ids+line nos.
List of doc ids+line nos.
Doc# + Line#
Doc# + Line#
Abstract Database
Doc# Line# Medline# LE LT Rel. RE RT Sent.
Doc# Offset
1 V1
2 V2
3 V3
4 V4
5 V5
N VN
Offset + Line# × sizeof (record)
Offset Table

Figure 10. Index structures created on Biological entities, tags and relations

7.1. Document parser

The document parser analyzes each sentence of the Medline abstracts to locate and extract feasible biological relations contained in them, if any. Relevant information about each useful sentence of the Medline abstract collection is stored in fixed-size records in a local database named as the *abstract* database. This is used to answer user queries posed at multiple levels of specificity.

All sentences containing at least one feasible biological relation component is stored in the abstract database, whose schema is shown in Table 3. Each document is assigned a unique identifier and a sentence is identified by its line number within the document. Thus each sentence is uniquely identified by the tuple (doc #, line #). If a sentence contains an information component, the relation and its left and right actors are extracted and stored in the abstract database. The Medline reference number is also stored in the database for possible future use in retrieving the complete document. Table 4 shows some entires from the abstract database table. If a sentence contains more than one information component, the parser locates all the components, which are then stored as different entries in the database. Whenever a new tagged document is incorporated into the Medline collection, it is analyzed to extract all information components occurring in it. Relevant entries are added to the database with the offset tables and other indexing structures appropriately updated.

Table 4. Schema of *Abstract* database along with some instances

Doc id	Line No.	LE	LET	Relation	RE	RET	Medline#	Sentence
1	8	T cells	Cell_type	Binds	peri-kappa B site	DNA_domain_or_region	MEDLINE: 95333264	While a nuclear factor(s) from both peripheral blood monocytes and T cells binds the peri-kappa B site, electrophoretic…
3	1	CD4 coreceptor	protein_molecule	interacts with	Non-polymorphic regions	protein_domain_or_region	MEDLINE: 95347379	The CD4 coreceptor interacts with non-polymorphic regions of major …
3	5	NF-AT	protein_molecule	Induce	Calcium flux	other_name	MEDLINE: 95347379	Lack of full activation of NF-AT could be correlated to a dramatically reduced capacity to induce calcium flux and…
5	3	cis-acting elements	DNA_family_or_group	Mediate	mouse IL-2R alpha gene	DNA_domain_or_region	MEDLINE: 95256242	Here we map the cis-acting elements that mediate interleukin responsiveness of the mouse IL-2R alpha gene using…

7.2. Indexer

The main indexing structure has been implemented to retrieve relevant sentences in answer to a user query efficiently. The basic query template is designed based on the earlier design of information component and is represented as a triplet *<Left Entity/Tag/*, Relation/*, Right Entity/Tag/*>*. The aim is to retrieve all sentences which contain a relation with matching entities or tags. A * denotes any match. Since a query can be posed by specifying entities or tags or relations, the abstract database has been multiply indexed with these elements.

The indexing structure has two components. The first one is the *information component indexer*, which is implemented as a 3D structure and locates sentences on the basis of biological tags occurring in a particular context of a biological relation. The second component is an *entity trie*, which indexes the abstract database on biological entities.

The *information component indexer* is implemented as a three-dimensional (3D) ragged grid array. The three axes of this structure represent the Left Actor, the Right Actor and the relation. This 3D grid array index structure is shown in Fig. 10. Every cell of the 3D structure stores a list of unique ids, generated from document number and line number, where each line in the list contains an information component comprised of the corresponding tags and relation.

The ragged grid array index structure has been chosen since the number of feasible relations differs from pair to pair. We have observed that using a ragged structure rather than a fixed array index structure drastically reduces the overall memory requirement. For example, since there are 36 biological tags defined in GENIA and 246 feasible biological relations have been mined from the documents, this leads to the possibility of storing 36×246×36 (=318816) relation triplets. Suppose 2 bytes are required to store a unique id created using the document number and line number. If a relation triplet has 10 occurrences on an average in the corpus, the memory space used to store the entire information in a fixed array structure will be approximately 6.08 MB, with many cells having no entries. In contrast, since our tag co-occurrence analysis has already found that only 4162 triplets are feasible, hence each tag pair is allocated just enough space to store only the feasible relations between them. With the same assumptions as stated earlier, the memory requirement in this case is found to be approximately 81.29 KB.

The cells in the grid array index structure contain a combination of document# and line# that is used to retrieve a relevant sentence from the *abstract* database. Since the Medline abstracts contain different number of lines, and a line may contain more than one information component. Hence an offset table is used to store the base address of each document. The address at which the first sentence of a document occurs in the database is called the base address of the document. The offset table is used to expedite the process of locating relevant sentences containing information components within the abstract database. To retrieve line 'L' of document 'D', the document number D is used as an index for the offset table where the base address B (say) of D is stored. The location of L is then computed as:

$Location(L) = B + L \times SizeOf(\mathrm{Re}cord)$, where, $SizeOf(\mathrm{Re}cord)$ is defined in the abstract database.

A *trie structure* on biological entity names is also maintained by the system, which stores information about entity occurrences in documents. For a query containing only entity names, the trie structure is used to locate specific lines that contain relevant information. Each entity name stored in the trie structure contains information about its biological tag and points to a list of sentences containing the entity. Each sentence is identified as earlier, by (doc #, line #). This combination is used to locate a sentence through the offset table.

The use of the Trie structure to answer entity based queries makes the search process very fast. The speed up in search is achieved due to the following:

- A trie structure offers space efficiency for storing Biological names since a sub-string is stored only once. This is very gainful for handling biological entities since very often they have common substrings extended by digits or alphabets to create new entity names. Common examples of this type of entities are Interleukin-1 and Interleukin-2, CD-28 and CD-20 etc. Other indexing structures like B Tree, B+ tree etc. would have stored these sub-strings repeatedly.
- The maximum number of comparisons required to find whether an entity exists in the Knowledge base, is same as the length of the entity name. If a query contains an entity name which is not found in the trie, further search is abandoned, since it is anyway futile.

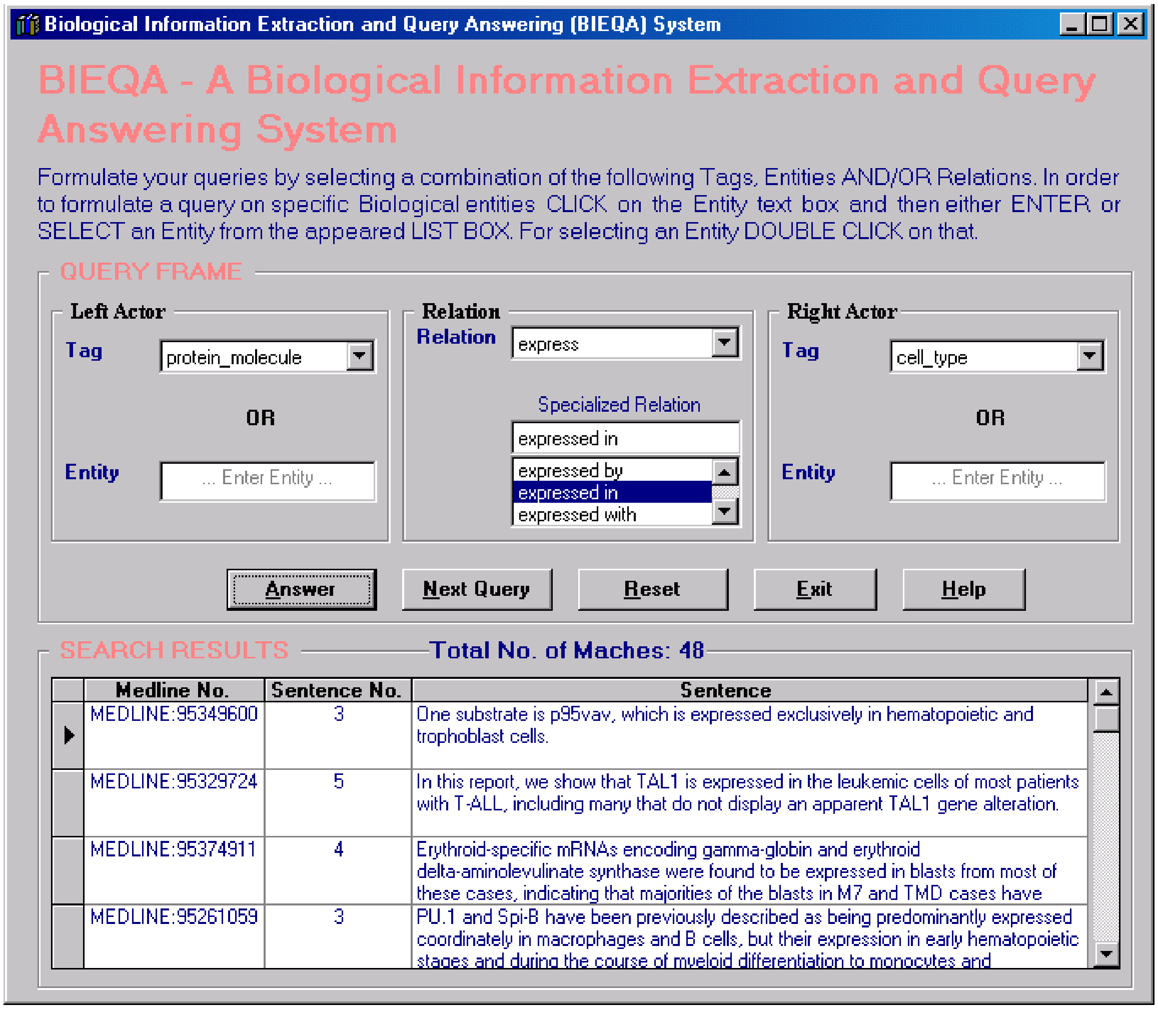
For a query containing both entity names and biological tags or relations, the query processor maps the entities to their respective tags, and then uses the 3D grid structure to locate relevant sentences. We illustrate the search process through a sample query <*protein_molecule, activated by, HOX11*>. While processing this triplet, HOX11 is identified as a known entity whose occurrence in the database can be traced to some sentences. The left actor here is the biological tag *protein_ molecule,* which is to be related through the relation *activated by* to another *protein_molecule,* since *HOX11* is of type *protein_molecule.* Hence the grid cell corresponding to <*protein_molecule, activated_by, protein_molecule*> is looked up for pointers to specific lines that contain relevant information. Those sentences which occur in the output of both entity trie structure and 3D indexing structure are retrieved. Considering another query, <*NF-Kappa B, expressed in, T-Cells*>*,* for which both actors are entity names, a set of relevant sentence numbers containing one or both entities are retrieved from the entity trie structure. Since *NF-*

Kappa B is a *protein molecule*, while *T Cells* is an instance of *cell type*, the cell corresponding to <*protein molecule, expressed in, cell type*> is looked up for obtaining a list of possible relevant sentences. The sentences which occur in the output of both the indexing structures are finally retrieved. Involving the 3D structure for answering this query reduces the overall retrieval time by fixing the context within a single desired relation. If a query does not contain any entity name, the entity trie structure is not consulted at all.

7. Query interface

In this section we will present the design of the query interface module, which processes user queries and displays relevant sentences retrieved, along with Medline reference numbers extracted from the abstract database. The Medline reference numbers may be used to extract the complete abstract.

Query processing is a two-step process - acceptance and analysis of the user query and finding the relevant answers from the structured knowledge base. Fig. 11 shows a snapshot of the user interface. A query is represented by a triplet <*Left Entity/Tag/**, *Relation/**, *Right Entity/Tag/**>. The middle element of the triplet is either a valid biological relation that has been identified by


Biological Information Extraction and Query Answering (BIEQA) System

BIEQA - A Biological Information Extraction and Query Answering System

Formulate your queries by selecting a combination of the following Tags, Entities AND/OR Relations. In order to formulate a query on specific Biological entities CLICK on the Entity text box and then either ENTER or SELECT an Entity from the appeared LIST BOX. For selecting an Entity DOUBLE CLICK on that.

QUERY FRAME

Left Actor

Tag: protein_molecule

OR

Entity: ... Enter Entity ...

Relation

Relation: express

Specialized Relation: expressed in

expressed by
expressed in
expressed with

Right Actor

Tag: cell_type

OR

Entity: ... Enter Entity ...

SEARCH RESULTS — Total No. of Maches: 48

Medline No.	Sentence No.	Sentence
MEDLINE:95349600	3	One substrate is p95vav, which is expressed exclusively in hematopoietic and trophoblast cells.
MEDLINE:95329724	5	In this report, we show that TAL1 is expressed in the leukemic cells of most patients with T-ALL, including many that do not display an apparent TAL1 gene alteration.
MEDLINE:95374911	4	Erythroid-specific mRNAs encoding gamma-globin and erythroid delta-aminolevulinate synthase were found to be expressed in blasts from most of these cases, indicating that majorities of the blasts in M7 and TMD cases have
MEDLINE:95261059	3	PU.1 and Spi-B have been previously described as being predominantly expressed coordinately in macrophages and B cells, but their expression in early hematopoietic stages and during the course of myeloid differentiation to monocytes and

Fig. 11. Query interface.

the system as feasible and frequent, or *. Left and right elements of the triplets are either an entity name or an ontology tag or *.

The template allows the user to formulate feasible queries at multiple levels of specificity. A user may specify an entity name and ask for documents containing it. However, a user can be more generic and post a query specifying requirements at ontology concept level, rather than entity names. In that case, all sentences containing an entity which is an instance of the corresponding biological tag, in the appropriate context is judged as relevant. A query can also contain mixture of concept tags and entity names and/or a specific relation. A * in any field represents a wild card entry and any match is considered as successful. Thus a sample query can be formulated as a pattern <*Protein_molecule, activated by, interleukin-1*>, which is a combination of generic concepts like *protein_molecule*, specific instances like *interleukin-1* and a specific relation *activated by*, which should relate these elements in the document.

The list of *feasible fuzzy relations* along with their membership values is used by the system to assist the users during query formulation. Since feasible biological relations have already been identified, hence a user cannot submit an infeasible relation-tag pair requirement as a query. Thus guided query formulation does not allow the user to specify a query like <*Nucleotide, mediate, lipid*>, since it is known that the tag pairs specified in the triplet do not occur in association with the mentioned relation in the document collection.

Behavior wise, the default *relation* list that is displayed to the user initially contains all root relations. Similarly default left and right tag lists initially display all biological tags present in the underlying ontology. After selecting a root relation, the user may further refine the query by choosing morphological variants of the relations. When the user selects a specific tag or relation name, only the feasible elements for the remaining part of query triplet are displayed by the system in decreasing order of fuzzy membership values. On choosing a tag or a tag pair, only the feasible relations associated with these are displayed in order of decreasing strength of association with the selected tags. Similarly, for a chosen relation, only valid tag pairs are filtered and displayed according to their decreasing degree of association strength. The same holds for morphological variants of relations also. The next section presents a discussion on overall system performance.

9. System performance analysis

The performance of the whole system is analyzed by taking into account the performance of the *Biological relation extraction* process and that of the *Biological information extraction* process for answering user queries, separately. Details of the evaluation processes and analysis of results obtained are given in the following subsections.

9.1. Evaluation of Biological Relation Extraction Process

The aim of the biological relation extraction process is to identify relevant verbs signifying biological interactions and also their associated biological entities or tags from Medline abstracts. We have already explained the extraction process in section 6. The extraction process has been applied over the whole tagged corpus to extract all biological relations present in the corpus along with their actors and frequencies. A partial list of extracted relations has been presented in section 6.

We now present detailed discussion about how we evaluate the correctness of the biological relation triplets extracted by analyzing the original sentences in which these relational verbs occur. Since the relation triplets have been extracted from the entire GENIA corpus, to evaluate the correctness of the extraction process, we have randomly selected 10 different relation-triplets and 500 documents from the corpus for manual verification.

A relation-triplet is said to be "correctly identified" if its occurrence within a sentence along with its left and right tags is grammatically correct, and the system has been able to locate it in the right context. To judge the performance of the system, it is not enough to judge the extracted relations only, but it is also required to analyze all the correct relations that were missed by the system. The system was evaluated for its recall and precision values for each of the selected relation triplets. Recall and precision for this purpose are defined as follows:

$$recall = \frac{\#times\ a\ relation\ triplet\ is\ correctly\ identified\ by\ the\ system\ (True\ Positives)}{\begin{array}{l}\#times\ the\ relation\ triplet\ actually\ occurs\ in\ the\ corpus\ with\ the\ specified\ tags\\ in\ the\ correct\ sense\ (True\ Positives + False\ Negatives)\end{array}}$$

$$precision = \frac{\#times\ a\ relation\ triplet\ is\ correctly\ identified\ by\ the\ system\ (True\ Positives)}{\begin{array}{l}\#times\ the\ relation\ triplet\ is\ identified\ by\ the\ system\ as\ correct\ (True\ Positives +\\ False\ Positives)\end{array}}$$

Since the presence of meta-tags and nested tags makes it very difficult to manually locate all possible relations, an evaluation software was written in VC++ that exhaustively checks the corpus for possible occurrences of the required relation. This software identifies all matches for a given relation name from the evaluation corpus, based on pattern matching. For each relation to be judged, the evaluation software takes in the root relation and performs partial string matching to extract all possible occurrences of the relation. This ensures that various nuances of English language grammar can also be taken care of. For example, if the root relation used in any query is "*activate*", all sentences containing *activates*, *inactivate*, *activated by*, *activated in* etc. are extracted. Each sentence containing an instance of the pattern is presented to the human evaluator in two forms. In one form, the biological tags occurring in it are retained, while the other one presents a clear view after stripping off the tags. The second form makes it easier for the evaluator to judge the grammatical correctness of the relation in association to the tags or entities around it. Sample representations generated by the software are shown in Fig. 12. Each occurrence of the relation is judged for correctness by the evaluator, and the correct instances are marked. The marked instances are stored by the evaluation software and later used for computing the precision and recall values.

Table 5 summarizes the performance of the system for 10 different relation triplets. The precision value of the system reflects its capability to identify a relational verb along with the correct pair of tags within which it is occurring. The precision of the proposed system is found to be 92.7%. Recall value reflects the capability of the system to locate all instances of a relation within the corpus. The recall value of this module is 83.89%, which can be improved. On analysis, it was found that most of the misses occur when a biological tag/entity has been used earlier in the sentence or in an earlier sentence, and is referred to in the vicinity of the relation using a pronoun. Since pronouns are not tagged, hence these are not recognized by the system.

```
Tagged: … <cons sem="G#protein_molecule"><cons sem="G#protein_molecule">IL-
      4</cons>     Stat</cons>     is     activated     by     <cons
      sem="G#protein_molecule">JAK3</cons> …

Stripped: …. IL-4 Stat is activated by JAK3 …

Tagged:  …  <cons  sem="G#protein_molecule">LMP-1</cons>  activates  <cons
      sem="G# protein_molecule">NF-kappa B</cons> by targeting the …

Stripped: … LMP-1 activates NF-kappa B by targeting the …
```

Fig. 12. Sample outputs of the evaluator program.

Misses also occur in the case where a relevant tag occurs in conjunction with other tags separated by operators like “or”, “,” etc., and the tags in the immediate vicinity of the relation do not match the tag given in the relation triplet.

Table 5. Precision and Recall Values of the *Biological Relation Extraction* Process

Relation Triplets	**Total # times a relation triplet is identified by the system**	**Total # times a relation triplet is correctly identified by the system**	**Total # times a relation triplet occurs correctly in test corpus**	**Precision (%)**	**Recall (%)**
<Protein_molecule, Activates, Protein_molecule>	7	7	8	100.00	87.50
<Protein_molecule, Expressed in, Cell_type>	15	12	15	80.00	80.00
<DNA_domain_or_region, Expressed_in, Cell_type>	5	5	7	100.00	71.43
<Protein_molecule, binds to, <DNA_domain_or_region>	7	7	7	100.00	100.00
<Protein_family_or_group, Interacts with, DNA_domain_or_region>	3	3	3	100.00	100.00
<Protein_molecule, Activated in, Protein_molecule>	2	2	2	100.00	100.00
<DNA_domain_or_region, Regulated by, Protein_family_or_group>	3	3	4	100.00	75.00
<DNA_family_or_group, Associated with, DNA_domain_or_region>	2	2	2	100.00	100.00
<Protein_molecule, Associated with, Protein_molecule>	5	5	6	100.00	83.33
<Protein_molecule, Induces, Protein_family_or_group>	4	3	4	75.00	75.00
Average				**95.50**	**87.23**

9.2. Evaluation of biological information extraction process

The performance of the biological information extraction process can be judged by comparing the information extraction accuracy against human evaluation. We present here a detailed performance analysis of the information extraction module, through analysis of several queries posed to the entire corpus. For each query, all generated answers are subjected to human judgment for accuracy analysis, where all possible relevant answers are located using the evaluation software discussed earlier.

Precision and recall values for the information extraction process are computed as follows:

$$precision = \frac{\#correct\ answers\ generated\ by\ the\ system\ (True\ Positives)}{\#total\ answers\ extracted\ by\ the\ system\ (True\ Positives + False\ Positives)}$$

$$recall = \frac{\#correct\ answers\ generated\ by\ the\ system\ \ (True\ Positives)}{\#total\ correct\ answers\ identified\ by\ human\ evaluator\ (True\ Positives + False\ Negatives)}$$

Now, we present some example queries and corresponding answers generated by the system. Each query template presented, is followed by a snapshot of the query interface showing a partial list of the sentences retrieved from the abstract database.

Query 1: *<*, Modulates, *>*

Medline No.	***Sentence No.***	***Sentence***
MEDLINE: 99377310	4	In the present study, by using an in vitro model, we examined whether stimulation with interleukin-6 (IL-6), an immunoregulatory, multipotential cytokine, modulates the expression and activities of the MSR in macrophages.
MEDLINE: 94240138	9	The absence in I kappa B gamma-1 and I kappa B gamma-2 of a protein kinase A site whose phosphorylation modulates p70I kappa B gamma inhibitory activity suggests that alternative RNA splicing may be used to generate I kappa B gamma isoforms that respond differently to intracellular signals.
MEDLINE: 96427516	1	Cytomegalovirus modulates interleukin-6 gene expression.
MEDLINE: 96195539	1	Interferon-gamma modulates the lipopolysaccharide-induced expression of AP-1 and NF-kappa B at the mRNA and protein level in human monocytes.
MEDLINE: 96195539	2	Interferon-gamma (IFN-gamma) modulates the expression of several cytokines by human monocytes at the transcriptional level.
MEDLINE: 96224055	1	A hydrophobic domain of Ca2+-modulating cyclophilin ligand modulates calcium influx signaling in T lymphocytes.

Query 1 is a generic query in which only the biological relation "*modulates*" is specified. All abstracts that contain information about two biological entities being related through the relation "*modulates"* are to be retrieved. Six out of fifteen answers generated by the system are shown in the above table. No answer was missed.

Query 2: *<Interleukin-10, Inhibits, *>*

Medline No.	*Sentence No.*	*Sentence*
MEDLINE: 95238477	1	Interleukin (IL)-10 inhibits nuclear factor kappa B (NF kappa B) activation in human monocytes.
MEDLINE: 95238477	3	Our previous studies in human monocytes have demonstrated that interleukin (IL)-10 inhibits lipopolysaccharide (LPS)-stimulated production of inflammatory cytokines, IL-1 beta, IL-6, IL-8, and tumor necrosis factor (TNF)-alpha by blocking gene transcription.
MEDLINE: 95238477	4	Using electrophoretic mobility shift assays (EMSA), we now show that, in monocytes stimulated with LPS or TNF alpha, IL-10 inhibits nuclear stimulation of nuclear factor kappa B (NF kappa B), a transcription factor involved in the expression of inflammatory cytokine genes.
MEDLINE: 99155321	1	Interleukin-10 inhibits expression of both interferon alpha- and interferon gamma- induced genes by suppressing tyrosine phosphorylation of STAT1.
MEDLINE: 97335975	1	Interleukin-10 inhibits interferon-gamma-induced intercellular adhesion molecule-1 gene transcription in human monocytes.

In query 2 the right actor of relation is left unspecified. Five out of nine answers generated by the system are shown in the above table.

Query 3: *<Protein_molecule, Inhibits, *>*

Medline No.	*Sentence No.*	*Sentence*
MEDLINE: 95184007	2	I kappa B-alpha inhibits transcription factor NF-kappa B by retaining it in the cytoplasm.
MEDLINE: 95280909	3	Secreted from activated T cells and macrophages, bone marrow-derived MIP-1 alpha/GOS19 inhibits primitive hematopoietic stem cells and appears to be involved in the homeostatic control of stem cell proliferation.
MEDLINE: 95238477	1	Interleukin (IL)-10 inhibits nuclear factor kappa B (NF kappa B) activation in human monocytes.
MEDLINE: 95238477	3	Our previous studies in human monocytes have demonstrated that interleukin (IL)-10 inhibits lipopolysaccharide (LPS)-stimulated production of inflammatory cytokines, IL-1 beta, IL-6, IL-8, and tumor necrosis factor (TNF)-alpha by blocking gene transcription.
MEDLINE: 95238477	4	Using electrophoretic mobility shift assays (EMSA), we now show that, in monocytes stimulated with LPS or TNF alpha, IL-10 inhibits nuclear stimulation of nuclear factor kappa B (NF kappa B), a transcription factor involved in the expression of inflammatory cytokine genes.
MEDLINE: 99155321	1	Interleukin-10 inhibits expression of both interferon alpha- and interferon gamma- induced genes by suppressing tyrosine phosphorylation of STAT1.
MEDLINE: 97335975	1	Interleukin-10 inhibits interferon-gamma-induced intercellular adhesion molecule-1 gene transcription in human monocytes.

Query 3 is a generalized case of query 2 in which the left actor (protein_molecule) is a more generic concept than the earlier instance Interleukin 10. On reviewing the answers for the two queries, it is observed that the set of sentences retrieved in answer to query 2 is contained in the set of sentences retrieved for query 3, which is correct. It may also be noted from the answers generated that our entity recognizer is capable of recognizing IL-10 as a variation of Interleukin-10. Seven out of fifty six answers retrieved by the system for this query are shown in the above table.

Query 4: *<Protein_family_or_group, binds to, DNA_domain_or_region>*

Medline No.	***Sentence No.***	***Relevant Sentence***
MEDLINE: 95222739	2	Core binding factor (CBF), also known as polyomavirus enhancer-binding protein 2 and SL3 enhancer factor 1, is a mammalian transcription factor that binds to an element termed the core within the enhancers of the murine leukemia virus family of retroviruses.
MEDLINE: 96344715	5	NGFI-B/nur77 binds to the response element by monomer or heterodimer with retinoid X receptor (RXR).
MEDLINE: 97051821	6	An unidentified Ets family protein binds to the EBS overlapping the consensus GAS motif and appears to negatively regulate the human IL-2R alpha promoter.
MEDLINE: 96239482	3	ICSAT is structurally most closely related to the previously cloned ICSBP, a member of the IFN regulatory factor (IRF) family of proteins that binds to interferon consensus sequences (ICSs) found in many promoters of the IFN-regulated genes.
MEDLINE: 99102381	3	NF-Y is a ubiquitous and evolutionarily conserved transcription factor that binds specifically to the CCAAT motif present in the 5' promoter region of a wide variety of genes.
MEDLINE: 97419190	3	Studies of the mechanisms that enable EBV to infect nonactivated, noncycling B cells provide compelling evidence for a sequence of events in which EBV binding to CD21 on purified resting human B cells rapidly activates the NF-kappaB transcription factor, which, in turn, binds to and mediates transcriptional activation of Wp, the initial viral latent gene promoter.

Query 4 is very restrictive in which both the left and right actors as well as the relation are specified. For this query there are 16 answers retrieved, out of which 6 are shown.

Query 5: *<Cell_type, producers of, protein_molecule>*

Medline No.	***Sentence No.***	***Sentence***
MEDLINE: 95394020	2	The regulation of interleukin (IL)-2 gene expression has been investigated mainly in T lymphocytes, the predominant producers of IL-2.

This query generated only one answer and there were none missed.

Query 6: *<*, regulated by, Protein_family_or_group>*

Medline No.	***Sentence No.***	***Sentence***
MEDLINE: 95015010	5	We present data to show that the expression of TNF alpha is regulated by the transcription factor C/EBP beta (NF-IL6).
MEDLINE: 95015876	1	Inducible binding to the c-fos serum response element during T cell activation is regulated by a phosphotyrosine-containing protein.

MEDLINE: 96146856	4	The expression of the QR gene is regulated by the transcription factor AP-1.
MEDLINE: 97074532	5	PML appears to be transcriptionally regulated by class I and II interferons, which raises the possibility that interferons modulate the function and growth…
MEDLINE: 96315681	2	The lymphocyte-specific immunoglobulin mu heavy-chain gene intronic enhancer is regulated by multiple nuclear factors.
MEDLINE: 96278973	1	Tissue-specific activity of the gammac chain gene promoter depends upon an Ets binding site and is regulated by GA-binding protein.

Query 6 is also a generic query in which there is no constraint on the Left actor. This query finds all possible biological substances that are regulated by protein family or group. Six out of twenty eight answers generated by the system are shown in the above table.

Query 7: *<DNA_domain_or_region, regulated by, GATA-1>*

Medline No.	***Sentence No.***	***Sentence***
MEDLINE: 95280913	8	ER-mediated repression of GATA-1 activity occurs on an artificial promoter containing a single GATA-binding site, as well as in the context of an intact promoter which is normally regulated by GATA-1.

Query 7 is a specific query in which a user is interested to find documents which contain information about the generic concept *DNA_domain_or_region* being regulated by GATA-1, which is a specific protein molecule. This query also generated only one answer and there were none missed.

While the above queries were correctly answered with high precision and recall values, following are some queries, for which we have analyzed why certain answers were missed.

Query 8: *<Protein_molecule, activated by, Protein_molecule>*

Medline No.	***Sentence No.***	***Sentence***
MEDLINE: 99038208	6	The gene encoding the retinoic acid-synthesizing enzyme aldehyde dehydrogenase 1 (Aldh1), initially called Hdg-1, was found to be ectopically activated by HOX11 in this system.
MEDLINE: 99032541	7	beta-casein, a cytokine-inducible SH2-containing protein (CIS), and oncostatin M (OSM), suggesting that STAT6 activated by IL-4 substitutes for the function of STAT5 in T cells.
MEDLINE: 95190988	1	LMP-1 activates NF-kappa B by targeting the inhibitory molecule I kappa B alpha.
MEDLINE: 95190988	11	These results indicate that LMP-1 activates NF-kappa B in B-cell lines by targeting I kappa B alpha.
MEDLINE: 96329553	1	Granulocyte-macrophage colony-stimulating factor preferentially activates the 94-kD STAT5A and an 80-kD STAT5A isoform in human peripheral blood monocytes.

Five out of twenty four answers generated by the system are shown above. In this query, since the left and right actors are same, the relations *activates* and *activated by* play equivalent roles. This is taken care of by the system and as shown in the above table, all sentences containing phrases either of the form *protein_molecule activates protein molecule*, or *protein_molecule activated by protein_molecule* are extracted from the abstract database. Although, all the extracted instances are found to be relevant for the user query, we illustrate an example sentence that could not be retrieved by our system due to the complexity of the underlying natural language construction. The following sentence is the fourth one in MEDLINE: 99049827.

```
<cons sem="G#protein_molecule">IFN regulatory factor 1</cons> (<cons
sem="G#protein _molecule">IRF-1</cons>) is a <cons sem="G#protein_family_
or_group">transcription factor </cons> activated by either <cons sem="G#
protein_molecule">CD40</cons> or <cons sem= "G#protein_family_or_group
">cytokines</cons>.
```

This sentence presents information of the form “E1 is a E2 which is *activated* by either E3 or E4” where E1, E2, E3 and E4 are entities. E1 and E3 are *protein_molecules* whereas E2 and E4 are elements of *protein_family_or_group*. Even though, this sentence contains a correct answer of the form “E1 is *activated* by E3”, the system misses it. This is due to the fact that the system wrongly recognizes E2 and E3 as the left and right actors respectively of the relation *activated by* and since E2 belongs to *protein_family_or_group*, this sentence could not be extracted as a relevant one.

We have observed that some failures occur when the relevant relation is a part of a tagged entity. For example consider the following sentence, which is the seventh sentence in MEDLINE: 97216066

```
Both <cons sem="G#protein_molecule">FMLP</cons> and <cons sem="G#protein
_molecule"><cons sem="G#protein_molecule">PAF</cons> activated MAP kinase
kinase-3 </cons> (<cons sem="G#protein_molecule">MKK3</cons>), a known
activator of <cons sem= "G#protein_molecule">p38 MAPk</cons>.
```

In this sentence, *PAF* is tagged as a *protein_molecule* and *PAF activated MAP kinase kinase3* is tagged as another *protein_molecule* due to which *activated* was not considered as a relation by the system.

Query 9: *<cell_type, stimulated with, protein_molecule>*

Medline No.	***Sentence No.***	***Sentence***
MEDLINE: 95256455	3	Adhesion of human monocytes to P-selectin, the most rapidly expressed endothelial tethering factor, increased the secretion of monocyte chemotactic protein-1 (MCP-1) and tumor necrosis factor-alpha (TNF-alpha) by the leukocytes when they were stimulated with platelet-activating factor.
MEDLINE: 95325614	4	Eosinophils were purified from peripheral blood by discontinuous Percoll gradients and stimulated with IL-5.
MEDLINE: 94312466	4	IL-4 down-regulated mRNA accumulation of the proinflammatory cytokines IL-1 beta, IL-8, and TNF-alpha in monocytes stimulated with IL-2, IL-3, and GM-CSF.
MEDLINE: 96310948	7	Indeed, monocytes pretreated with IL-10 are able so inhibit NF-kappa B nuclear activity in purified T lymphocytes stimulated with OKT3.
MEDLINE: 99225047	10	Moreover, RA induced apoptosis of CD34+ cells and CD34+CD71+ cells stimulated with erythropoietin.
MEDLINE: 99242525	5	Purified peripheral eosinophils were stimulated with IFN-gamma at 37 degrees C for 1-60 min.

Query 9 models a user query, which tries to identify information about cell types being stimulated with any kind of protein molecule. Six out of twelve answers generated by the system are shown in the above table. One of the relevant answers missed for this query is a sentence from MEDLINE: 96205960, which reads as follows:

```
<cons sem="G#cell_type">Peripheral blood T cells</cons> stimulated with <cons
sem="G#   other_organic_compound">PMA</cons>/<cons   sem="G#protein_molecule">
alphaCD28</cons > produced <cons sem="G#protein_molecule">IL-2</cons> in the
presence of <cons sem="G#other_organic_compound">CsA</cons>.
```

In this sentence, the system fails to recognize that both *PMA* and *alphaCD28* are candidates for right actor. However, since PMA is the closer of the two but is tagged separately from *alphaCD28* as an organic compound, our system fails to recognize *alphaCD28* as a relevant right actor.

Another such miss occurs for second line of MEDLINE: 99069282

```
<cons   sem="G#cell_type">Human   monocytic   cells</cons>express<cons   sem="G#
protein_molecule">interleukin-1beta</cons>   (<cons   sem="G#protein_molecule">
IL-1beta</cons>)  when  stimulated  with  the  <cons  sem="G#protein_molecule">
extracellular  matrix  glycoprotein  </cons>,  <cons  sem="G#protein_molecule">
fibronectin</cons> (<cons sem="G#protein_ molecule">FN</cons>).
```

In this sentence there are two relational verbs *express* and *stimulated with* and the subject of first relation is also a subject for the second relation. But, our system is incapable of recognizing this, since we have dealt with strict binary relations only.

Query 10: *<protein_molecule, expressed in, T cells>*

Medline No.	*Sentence No.*	*Sentence*
MEDLINE: 97069862	10	These data indicate that TAL-1, expressed in T cells, is per se a potent oncogene, which may exert a key leukemogenetic role in the majority of T-cell acute lymphoblastic leukemias.
MEDLINE: 97154704	12	In contrast, when RBTN-2 is inappropriately expressed in T cells, RBTN-2 would interact predominantly with elf-2b; this interaction may lead to T cell proliferation.
MEDLINE: 93204999	5	Surprisingly, the levels of SRF constitutively expressed in T cells are consistently higher than in other cell types.

For query 10 there are 3 generated answers all of which are shown in the above table. For this query, the system could not extract the following sentence due to non-assignment of a tag to the term T Cells.

```
<cons    sem="G#protein_molecule">Granulocyte-macrophage    colony-stimulating
factor</cons> (<cons sem="G#protein_molecule">GM-CSF</cons>) is a hemopoietic
growth   factor   that   is   expressed   in   activated   T   cells,   fibroblasts,
macrophages, and endothelial cells.
```

Another sentence which is relevant to the above query, but is missed by the system since the left actor could not be identified properly, is shown below:

```
<cons   sem="G#protein_molecule">ETS1</cons>   is   a   <cons   sem="G#protein_
family_or_group">transcription  factor</cons>  of   the  <cons  sem="G#protein_
family_or_group">ETS   family   </cons>   that   is   expressed   in   <cons   sem=
"G#cell_type">T cells</cons>.
```

On analysis of a large number of queries, it was found that all misses occur due to one of the following: (i) co-occurrence of the desired tag with another tag coming between it and the relation (ii) verb itself is a part of the tag (iii) non-assignment of tags or inconsistent tagging (iv) more than one relational verb is associated with one entity.

Table 6 summarizes the performance of the system in terms of precision and recall, as defined earlier, for the 10 queries presented above. It also presents the average performance of the system. Analysis shows that the performance of the system in terms of relevance of answers extracted is quite high. However, as discussed earlier, binary relations are not capable of handling all natural language nuances. Hence some answers are missed thereby reducing the recall value of the system.

Table 6. Precision and Recall values of query answering for *Biological Information Extraction*

Query#	**# Answers Extracted by the System**			**# Missed answers (CNE)**	**Precision (%)**	**Recall (%)**	**Avg. Precision (%)**	**Avg. Recall (%)**
	Correct (CE)	***Incorrect (IE)***	***Total (CE+IE)***					
1	15	0		0	100.00	100.00	98.87	84.68
2	9	0		1	100.00	90.00		
3	54	2		7	96.43	88.52		
4	16	0		6	100.00	72.73		
5	1	0		0	100.00	100.00		
6	27	1		10	96.43	72.97		
7	1	0		0	100.00	100.00		
8	23	1		2	95.83	92.00		
9	12	0		5	100.00	70.59		
10	3	0		2	100.00	60.00		

10. Conclusions and future work

In this paper we have presented a system for ontology-based biological information extraction and query answering. The unique aspect of our system lies in its capability to extract information about generic biological relations from tagged biological documents. The extracted information is used to retrieve relevant sentences from biological documents in answer to a user query.

The proposed system, BIEQA, accepts as input a biological ontology and text documents which are tagged according to this ontology. In this implementation, we have considered the GENIA corpus which 2000 MEDLINE abstracts manually tagged according to the GENIA ontology of Molecular Biology. BIEQA employs deep text mining to extract information about the likelihood of various entity-relation occurrences within these documents. The set of relations mined from a specific collection is stored as feasible fuzzy biological relations for that collection. The fuzzy membership of a relation reflects its degree of association with specific biological substances or locations.

The fuzzy relations extracted from a corpus are stored in a structured database, to provide an efficient mechanism for extracting biological information from text documents. The system employs novel indexing structures to access the stored information efficiently. User interaction with the system is provided through an ontology-guided interface, which enables the user to formulate queries at various levels of specificity including combinations of specific biological entities, tags and relations. The system guides users to formulate feasible queries. User queries are analyzed on the basis of relations and tags or entities present in them.

Currently, this work is being extended to incorporate rules to handle complex subjects and relations which are the causes for some kinds of misses as we have analyzed earlier. Incorporation of more natural-language handling techniques can definitely make the query processor more user-friendly. Specifically, mechanisms to handle negation, conjunction and disjunction of concepts expressed in natural language are required to handle more natural language queries.

Since this work presents a method to extract feasible relations from tagged corpora, the extracted relations can be successfully used to enrich the underlying ontology appropriately. Presently, we are working towards generating a fuzzy ontology structure, in which, biological relations between various biological concepts can be stored along with their strengths. Strength of relations can thereby play a role in determining relevance of a document to a user query.

The mined information about relations can also be successfully employed to analyze and learn the underlying principle of tagging the documents. This can help in tagging future documents automatically.

References

[1] A. Yakushiji, Y. Teteisi, Y. Miyao, J. Tsujii, Event Extraction from Biomedical Papers Using a Full Parser, Pac Symp Biocomput, 2001, pp. 408-19.

[2] B. J. Stapley and G. Benoit, Bibliometrics: Information Retrieval and Visualization from Co-occurrence of Gene Names in Medline Abstracts, in: Proceedings of the Pacific Symposium on Biocomputing, Oahu, Hawaii, 2000, pp. 529-540.

[3] C. Friedman, P. Kra, H. Yu, M. Krauthammer, A. Rzhetsky, GENIES: A Natural-Language Processing System for the Extraction of Molecular Pathways from Journal Articles, Bioinformatics, vol. 17, Suppl. 1, 2001, pp. s74-s82.

[4] C. Nobata, N. Collier, J. Tsujii, Automatic Term Identification and Classification in Biology Texts, in: Proceedings of the Natural language Pacific Rim Symposium, 1999, pp. 369-375.

[5] D. Gavrilis, E. Dermatas, Automatic Extraction of Information from Molecular Biology Scientific Abstracts, Specom 2003.

[6] D. Hanisch, J. Fluck, H. T. Mevissen, R. Zimmer, Playing Biology's Name Game: Identifying Protein Names in Scientific Text, Pacific Symposium on Biocomputing 8, 2003, pp. 403-414.

[7] D. Proux, F. Rechenmann, L. Julliard, V. Pillet, B. Jacq, Detecting Gene Symbols and Names in Biological Texts: A First Step Toward Pertinent Information Extraction, in: Genome Inform Ser Workshop Genome Inform 9, 1998, pp. 72-80.

[8] F. Rinaldi, G. Scheider, C. Andronis, A. Persidis, O. Konstani, Mining Relations in the GENIA Corpus, in: Proceedings of the 2nd European Workshop on Data Mining and Text Mining for Bioinformatics, Pisa, Italy, 24 September 2004.

[9] H. M. Muller, E. E. Kenny, P. W. Strenber, Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature, PloS Biology 2(11):e309, 2004, http://www.plosbiology.org.

[10] J.-D. Kim, T. Ohta, Y. Tateisi, J. Tsujii, GENIA Corpus – A Semantically Annotated Corpus for Bio-Textmining, Bioinformatics, Vol. 19, Suppl. 1, 2003, pp. i180-i182.

[11] J. Kazama, T. Makino, Y. Ohta, Y. Tsujii, Tuning Support Vector Machines for Biomedical Named Entity Recognition, in: Proceedings of the ACL Workshop of the Natural Language Processing in the Biomedical Domain, Philadelphia, PA, USA, July 2002, pp. 1-8.

[12] J. Thomas, D. Milward, C. Ouzounis, S. Pulman, M. Carroll, Automatic Extraction of Protein Interactions from Scientific Abstracts, in: Pacific Symposium on Biocomputing, 2000, pp. 538-549.

[13] K. Fukuda, T. Tsunoda, A. Tamura, T. Takagi, Toward Information Extraction: Identifying Protein Names from Biological papers, in: Proceedings of the Pacific Symposium on Biocomputing, Hawaii, 1998, pp. 707-718.

[14] K. Su, M. Wu, J. Chang, A Corpus-based Approach to Automatic Compound Extraction, in: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL'94).

[15] M. A. Andrade, A. Valencia, Automatic Extraction of Keywords from Scientific Text: Application to the Knowledge Domain of Protein Families, Bioinformatics 14(7), 1998, pp. 600-607.

[16] M. Abulaish, L. Dey, An Ontology-based Pattern Mining System for Extracting Information from Biological Texts, Lecture Notes in Artificial Intelligence 3642, Part II, Springer, 2005, pp. 420-429.

[17] M. Abulaish, L. Dey, Biological Ontology Enhancements: A Text Mining Framework, in: Proceedings of the 2005 IEEE/WIC/ACM Int'l Conference on Web Intelligence, France, 2005, pp. 379-385.

[18] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, *et al.*, Gene Ontology: Tool for the Unification of Biology, The Gene Ontology Consortium, Nat. Genet 25, 2000, pp. 25-29.

[19] M. Craven, J. Kumlien, Constructing Biological Knowledge Bases by Extracting Information from Text Sources, in: Proceedings of the 8th Int'l Conference on Intelligent Systems for Molecular Biology (ISMB'99), 1999, pp. 77-86.

[20] N. Collier, C. Nobata, J. Tsujii, Extracting the Names of Genes and Gene Products with a Hidden Markov Model, in: Proceedings of the 18th Int'l Conference on Computational Linguistics (COLING'2000), 2000, pp. 201-207.

[21] N. Uramoto, H. Matsuzawa, T. Nagano, A. Murakami, H. Takeuchi, K. Takeda, A Text-Mining System for Knowledge Discovery from Biomedical Documents, IBM Systems Journal 43(3), July 2004, pp. 516-533.

[22] P. G. Baker, C. A. Goble, S. Bechhofer, N. W. Paton, R. Stevens, A. Brass, An Ontology for Bioinformatics Applications, Bioinformatics 15, 1999, pp. 510-520.

[23] R. Gaizauskas, G. Demetriou, P. J. Artymiuk, P. Willett, Protein Structures and Information Extraction from Biological Texts: the PASTA System, Bioinformatics 19(1), 2003, pp. 135-143.

[24] T. C. Rindflesch, L. Hunter, A. R. Aronson, Mining Molecular Binding Terminology from Biomedical Text, in: Proceedings of the AMIA Symposium, 1999, pp. 127-131.

[25] T. -K. Jenssen, A. Laegreid, J. Komorowski, E. Hovig, A Literature Network of Human Genes for High-Throughput Analysis of Gene Expression, Nat. Genet 2001, pp. 28: 21-28.

[26] T. Ono, H. Hishigaki, A. Tanigami, T. Takagi, Automated Extraction of Information on Protein-Protein interactions from the Biological Literature, Bioinformatics 17(2), 2001, pp. 155-161.

[27] T. Sekimizu, H. S. Park, J. Tsujii, Identifying the Interactions Between Genes and Genes Products based on Frequently Seen Verbs in Medline Abstract, Genome Informatics 9, 1998, pp. 62–71.

[28] T. T. Quan, S. C. Hui, T. H. Cao, FOGA: A Fuzzy Ontology Generation Framework for Scholarly Semantic Web, in: Proceedings of the 2004 Knowledge Discovery and Ontologies Workshop (KDO'04), Pisa, Italy, 24 September 2004.

[29] Y. Tateisi, T. Ohta, N. Collier, C. Nobata, J. Tsujii, Building Annotated Corpus in the Molecular-Biology Domain, in: Proceedings of the COLING 2000 Workshop on Semantic Annotation and Intelligent Content, 2000, pp. 28-34.