International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:1, No:5, 2007

PIELG: A Protein Interaction Extraction System using a Link Grammar Parser from Biomedical Abstracts

Rania A. Abul Seoud, Nahed H. Solouma, Abou-Baker M. Youssef, and Yasser M. Kadah, *Senior Member, IEEE*

***Abstract*—**Due to the ever growing amount of publications about protein-protein interactions, information extraction from text is increasingly recognized as one of crucial technologies in bioinformatics. This paper presents a Protein Interaction Extraction System using a Link Grammar Parser from biomedical abstracts (PIELG). PIELG uses linkage given by the Link Grammar Parser to start a case based analysis of contents of various syntactic roles as well as their linguistically significant and meaningful combinations. The system uses phrasal-prepositional verbs patterns to overcome preposition combinations problems. The recall and precision are 74.4% and 62.65%, respectively. Experimental evaluations with two other state-of-the-art extraction systems indicate that PIELG system achieves better performance. For further evaluation, the system is augmented with a graphical package (Cytoscape) for extracting protein interaction information from sequence databases. The result shows that the performance is remarkably promising.

***Keywords*—**Link Grammar Parser, Interaction extraction, protein-protein interaction, Natural language processing.

I. INTRODUCTION

PROTEOMICS is aimed at understanding protein-protein interactions. The function of a protein can be characterized more precisely through knowledge of protein-protein interactions. Protein-protein interactions are important for many biological functions. They link many proteins in the cell into large connected interaction networks. Each protein can have one or more of many roles in the network. Moreover, networks of interacting proteins provide a first level of understanding the cellular mechanism.

Many tragic and costly problems in human health care to be solved need the support of continuous updated information about protein-protein interactions such as tissue loss or organ failure. Applications that repair or replace portions of or whole living tissues (e.g., bone, dentine, or bladder) using living cells is named *Tissue Engineering (TE)*. For example, dentine formation is the process of regenerating dental tissues by tissue engineering principles and technology. Dentine formation is governed by biological mediators or growth factors (protein) and interactions amongst different proteins. Dentine formation needs the support of continuous updated information about protein-protein interactions.

Researches in the last decade have resulted in the production of a large amount of information about protein functions involved in dentine formation process. That generated data is highly connected; hence, should be such data is made easily available. Scientists in that field are aided by many online databases covering different aspects of protein function, such as protein–protein interaction DIP [1] and BOND [2], CSNDB [3] and SPAD [4]. However, since they are dependent on human experts, they rarely store more than a few thousand of the best-known protein relationships and do not contain the most recently discovered facts and experimental details.

The information about protein – protein interactions involved in dentine formation process is scattered throughout numerous publications in scientific journals and/or abstracts. According to the U.S. National Library of Medicine [5], PubMed [6] includes over 17 million citations from MEDLINE and other life science journals for biomedical articles dating back to the 1950s, with about 40,000 being added each month. Hence, manual collection of database updating data from such resources becomes impractically time consuming. Thus, having a scalable, robust system for protein interaction involved in dentine formation process discovery provides a major information extraction tool for molecular biologists in *tissue engineering laboratories* to automatically extract and transfer updated biological data about protein-protein interactions from unstructured form, to a structured form to be used in their respective applications.

In this paper we present PIELG system. PIELG is a Protein Interaction Extraction System using a Link Grammar Parser from biomedical abstracts. PIELG is a fully automated extraction system to extract protein interactions in natural language texts. Our approach tags protein names with the help

Rania A. Abul Seoud is with the Department of Computer Engineering, Faculty of Engineering, Fayoum University, Fayoum, Egypt.
E-mail: r-abulseoud@k-space.org.
Nahed h. Solouma is with the Laser Institute, Helwan University, Giza, Egypt. E-mail: nsolouma@k-space.org.
Abou-Baker M. Youssef is with the Department of Biomedical Engineering, Faculty of Engineering, Cairo University, Giza, Egypt
E-mail: ABaker@k-space.org.
Yasser M. kadah is with the Department of Biomedical Engineering, Faculty of Engineering, Cairo University, Giza, Egypt
E-mail: ymk@k-space.org.

of protein names and linguistic ontologies. The system extracts complete interactions by analyzing the matching contents of syntactic roles and their linguistically significant combinations. The system uses phrasal-prepositional verbs patterns to overcome preposition combinations problems. PIELG is purely implemented with Perl under Linux platform. The recall and precision are 47.4% and 62. 65%. Experimental evaluations with two other state-of-the-art extraction systems indicate that PIELG system achieves better performance. For further evaluation, the PIELG system is augmented with a graphical package (Cytoscape) for extracting protein interaction information from sequence databases. The augmentation process evaluates the extracted interaction by drawing the pathways for the extracted interaction. Then compare those pathways with the stored pathways in Cytoscape. Our experimental results show that the PIELG system achieves better performance without the need of manual pattern creation (by user) which is required for other systems.

II. Related Work

The goal of relationship extraction is to detect occurrences of a pre-specified type of relationship between a pair of entities of given types. While the type of the entities is usually very specific (e.g. genes, proteins or drugs), the type of relationship may be very general (e.g. any biochemical association) or very specific (e.g. a regulatory relationship).

Many approaches have been proposed for information extraction (IE) from scientific publications, ranging from simple statistical methods to advanced natural language processing (NLP) systems. The first step done towards Information extraction was to recognize the names of proteins, genes, drugs and other molecules [7]. The next step was to recognize interaction events between such entities [8]. Basic information extraction approaches rely on the matching of pre-specified templates (patterns) or rules. A number of groups reported application of pattern-matching-based systems for protein-function information extraction [8], [9], [10]. The shortcoming of such systems is their inability to process correctly anything other than short, straightforward statements, which are quite rare in information-saturated PubMED abstracts.

In the last few years, Natural Language Processing (NLP) has become a rapidly-expanding field within bioinformatics, as the literature keeps growing exponentially [11] beyond the ability of human researchers to keep track of, at least without computer assistance. Natural language processing techniques rely on syntactic and semantic knowledge that is often manually encoded for a particular domain. Initially NLP is used for machine translation, speech recognition and also knowledge representation. Information Extraction (IE) researches use NLP techniques such as automated Part-of-Speech tagging to pre-process documents and to extract underlying information. NLP-based methods perform a substantial amount of sentence parsing to decompose the text into a structure from which relationships can be readily extracted. Many natural language processing approaches at various complexity levels have been used successfully to extract various classes of data from biological texts, including protein-protein interactions.

More advanced systems utilizing shallow parsing techniques have been described to extract protein interactions [12]. Shallow parsers perform partial decomposition of sentence structure. Unlike word-based pattern matchers; shallow parsers [13] perform partial decomposition of a sentence structure. They identify certain phrasal components and extract local dependencies between them without reconstructing the structure of an entire sentence. In some cases, shallow-parsers are used in combination with various heuristic and statistical methods [14]. The most promising candidates for a practical information extraction system are ones based on full-sentence parsing as they deal with the structure of an entire sentence and therefore are potentially more accurate. However, full parsers are significantly slower and require more memory. A problem of parsing ambiguity can be reduced by employment of domain-specific *context-sensitive grammars*. This approach has been implemented in a system called MedLee [15]. Another system is called GENIES [16] which utilizes a grammar based NLP engine for information extraction. *Context-free parsing systems*, on the other hand, are general enough to be applicable to any domain, but completely generic systems seem to be impractical and inefficient. The Pathway Assist system uses an NLP system, MedScan [17] for the bio-medical domain that tags the entities in text and produces a semantic tree. Slot filler type rules are engineered based on the semantic tree representation to extract relationships from text. Recently, it has been extended as GeneWays [18], which also provides a Web interface that allows users to search and submit papers of interest for analysis. The BioRAT [19] system uses manually engineered templates that combine lexical and semantic information to identify protein interactions. *Grammar engineering approaches*, on the other hand use manually generated specialized grammar rules that perform a deep parse of the sentences. *Machine learning approaches* have also been used to learn extraction rules from user tagged training data [20]. These approaches represent the rules learned in various formats such as decision trees or grammar rules.

Recently, extraction systems have also used *Link Grammar* to identify interactions between proteins. Link grammar itself is a robust and powerful framework. It can handle lots of irregularities and attempt to interpret sentences even when they are ungrammatical or contain some unknown words. Their approach relies on various linkage paths between named entities such as the gene and protein names. Ding et al. proposed an interaction extraction method based on Link Grammar Parser [21].They made a great leap in biomedical information extraction area. However, their work is limited to

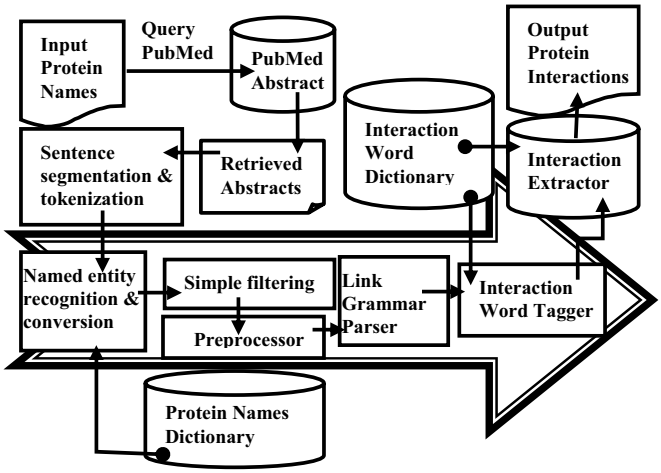

Fig. 1 System architecture

counting length of link paths only, neglecting the abundant grammatical information along the paths. In fact, the grammatical information is most valuable for interaction extraction. Basically, we cannot extract accurate information of interactions until the grammatical information is exhaustively exploited.

The ProtExt system [22] extending the idea of Ding et al., 2003. They proposed a novel template language (PETL) for extracting protein-protein interactions. Their system extracts protein-protein interactions embedded in sentences more accurately and customizable. Their information extraction approach relies on the matching of *pre-specified templates* (patterns) or rules. The underlying assumption is that sentences conforming exactly to a pattern or a rule express the predefined relationship(s) between the sentence entities. However, they need to consider a template optimizer to speed the matching which can pack numerous templates into one template using a more sophisticated data structure. Manually writing patterns for every verb is not practical for general purpose applications.

The IntEx [23] system splits complex sentences into simple clausal structures made up of syntactic roles. Their extraction system handles complex sentences and extracts multiple and nested interactions specified in a sentence. IntEx system achieves better performance without the labor intensive pattern engineering requirement. However, researchers are also interested in contextual information such as the location and agents for the interaction and the signaling pathways of which these interactions are a part. They don't extract the detailed contextual attributes (such as bio-chemical context or location) of interactions might give extra information to the biologist.

They don't identify the relationships among interactions extracted from a collection of sentences (such as one interaction stimulating or inhibiting another) to construct "Protein Interaction Pathways" from abstracts and full text articles. They didn't Attempt to improve the parse output of the Link Grammar System by augmenting the dictionaries of the Link Grammar Parser with medical terms with their linking requirements.

BioPPIExtractor system [24] applies Conditional Random Fields model to tag protein names in biomedical text, then uses a Link Grammar Parser to extracts complete interactions. Their main aim is to introduce CRFs-based protein name recognition method and evaluate its contribution to the overall protein – protein interaction performance. Their experiment results show that introduction of this method indeed helps to improve the PPI performance. However, the recall errors of BioPPIExtractor are due to the complicity of the protein interaction expression so they faced a difficulty to compile the appropriate extraction rules and, therefore, many interactions are missed out. The leading cause of precision errors of BioPPIExtractor is their not perfect extraction rules. Every system evaluates on a different test set, and so it is quite difficult to compare systems.

Although most of the previous mentioned biomedical information extraction systems focus on verbs which represent target events by themselves (i.e. "activate", "bind"), there are many cases that combinations of verbs, prepositions and certain nouns form proper IE forms. In this paper we present the PIELG system which investigates and classifies forms which are needed to extract interacting protein pairs. The PIELG system covers many linguistic variations of the interaction words in various contexts. Among the most frequently used forms is the nominalization form (converting a predicate to a noun phrase). Examples for nominalization are *interaction of*, *phosphorylation of*, *dephosphorylation of*, and so on. While many previous information extraction systems have concentrated only on the verbal forms of interactions, patterns for the nominal form in the case of 'phosphorylate' interactions is needed. The PIELG system covers nine classes based on constituents of the verbs as shown latter.

Also, the PIELG system success to extract of detailed contextual attributes of interactions by interpreting modifiers like: location/position modifiers (*in, at, on, into, up, over...),* agent/accompaniment modifiers (by, with…), purpose modifiers (*for*…), and theme/association modifiers (*of*...).

III. System Overview

A typical session in using PIELG involves the user providing an initial search specification (keywords). The keywords may be one protein name or pairs of protein names wanted to detect their interaction properties. Then PIELG downloads PubMed abstracts satisfying that specification. Each abstract is analyzed to identify sentences that mention interaction of proteins. These sentence clauses are then processed to obtain the interactions between proteins using syntactic roles of the sentence and their linguistically significant combinations. The *actor* and *patient* of each interaction are identified. These interaction evidence sentences are then grouped by actor and patient. Then PIELG extracts interaction information from abstracts and titles of scientific papers, and presents the extracted information in

textual forms. PIELG is purely implemented with Perl under Linux platform. The architecture of the PIELG system is shown in Fig. 1.The following sections briefly explain the workings of its modules.

IV. Sentence Segmentation and tokenization

Tokenization divides the input text into sentences and tokens and doing a lexical analysis. Each token represents the smallest linguistic unit; it can be a word (e.g. "run"), a numeric expression (e.g. "21st").

PIELG system for extracting interactions requires sentence segmentation since only the proteins within a sentence are considered when identifying interactions. This module identifies sentence and word boundaries. It splits the retrieved abstracts into sentences including titles of each paper. The title of a paper may include important information like the title of this paper: - *Dentin matrix protein-1 regulates dentin sialophosphoprotein gene transcription during early odontoblast differentiation.* This is done by using simple regular expressions, to identify sentence boundaries, assuming any period followed by a space and an uppercase letter is a sentence boundary. The word and sentence segmentation step is simplified. Tokens can be also tagged for other information.

V. Named Entity identification and conversion

Named entity identification or Entity extraction is the process of identifying protein names in the text. The simplest and frequently used approach to entity identification is a dictionary matching approach. Entity names are compiled as a dictionary. A string match with an entry in the dictionary tags the words or phrases as protein names. A variety of publicly available databases provide the resources for entity names.

Some of the major current sources for gene-related terms: genome and proteome databases such as NCBI's LocusLink [25] and UniProt [26] contain many of the names and synonyms denoting known genes in various organisms. PIELG distills its dictionary of protein names from EXpaxy [27] and iHOP [28] databases. The dictionary of PIELG carries about 1000 entries. However, we do not do any synonym grouping or name clustering. Since our main goal is aimed at proposing a method for extracting protein-protein interactions, the current named entity recognizer is sufficient for this purpose.

Named entity conversion process is important for entity extraction. It is the process of converting each protein name into a *personal name*. Before conversion we need to make sure that each protein name has one identical representation. It is noticed that a protein name may have different appearances and lots of identical representations. For example, the protein name *Dentin matrix protein-1* may appear as *Dentin matrix protein 1*. Also, its abbreviation may appear in the text as *DMP-1* or *DMP 1*. This module tries to normalize protein names using a dictionary so that different names of the same protein are mapped to a standard name. The conversion process aimed to get the Link Grammar Parser handles texts with protein names of multiple words. This is done by converting each protein name into a *personal name.* This is necessary because link parser does not have an unbounded dictionary which may hold the vocabulary of all protein names. Common personal names are already known to the Link Grammar parser and doing this can prevent it from guessing the biochemical names. For example, *Bone morphogenetic proteins* will be replaced by *BMPs* and *Electron probe micro-analysist* will be replaced by *EPMA*. If we do not do the conversion, then perhaps few sentences can be well parsed by the parser. Besides, doing this usually can reduce the number of words in sentences, which is helpful to processing. This will reduce the total processing time of the total system. If we take the following sentence as an example *Dentin matrix protein-1 is verified by real-time reverse transcruption-polymerase chain reaction* it will be converted to *DMP-1 is verified by real-time RT-PCR.*

VI. Simple Filtering and Transformation

Simple filtering is the process used to reduce the processing time for an abstract. It filters out sentences that do not contain any interactions. Sentences are again searched for the protein pairs. The sentences that contain at least two protein names are alone chosen for processing.

The *Transformation* process is needed to make the Link parser able to handle text with some expressions including protein names. The expressions of multiple words properly would have required a wrapper around the parser. This wrapper is the transformer that will transfer those expressions into *personal names* from the text before passing it to the parser. The re-transformer is then inserted in the name back after parsing, for example, *gene expression of Alp* and *expression of the transcription factor RUNX2*. Besides, doing this usually can reduce the number of words in sentences, which is helpful to processing. This will reduce the total processing time of the total system.

VII. Preprocessor

Preprocessor allows removing of numerous structure ambiguities, which clearly benefits the parsing quality and execution time. The tagged sentences need to be pre-processed to replace syntactic constructs, such as parenthesized nouns and domain specific terminology that cause the parser to produce an incorrect output. This problem is overcome by replacing such elements with alternative formats that is recognizable by the parser.

The preprocessor forces the Link Grammar parser to recognize the biological names as noun forms. Since the parser recognizes words that start with an uppercase letter as a noun therefore, the pre-processor converts each protein personal name to a word starting with an uppercase letter. The words in the parentheses are removed to improve the parse output as they provide no additional information in many sentences. However, there is some loss of information regarding the interactions due to this process which bring

down the recall of the extraction system. The pre-processor performs minor punctuation corrections on the spacing of commas and semi-

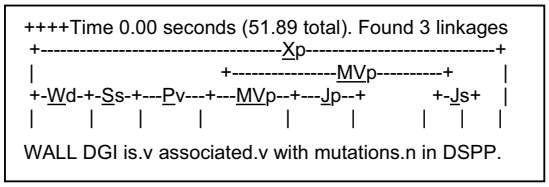

Fig. 2 A sample parser outpuft with links

colons in the text. It filters out some adverbs such as *however*, *hence*, *also*, *furthermore* etc. The preprocessor removes some information that is unrelated to biochemical interactions, such as a window of time: (1994-2008), probabilities, mathematical notations: (p _ 0.03), special characters, and so forth. The rationale of doing this is that it can save some computational effort during parsing without losing crucial information related to interactions and make sentences more understandable to Link Grammar Parser.

VIII. Link Grammar Parser and Link Grammar

Link grammar (LG) is an original theory of English syntax. It was written by Davy Temperley, Daniel Sleator, and John Lafferty of Carnegie Mellon University [29] to simplify English grammar with a context-free grammar. Link grammar [30] is a theory of syntax which builds relations between pairs of words, rather than constructing constituents in a tree-like hierarchy. It is based on a model that words within a text form links with one another. It thinks of words as blocks with connectors coming out. There are different types of connectors; connectors may also point to the right or to the left. A left-pointing connector connects with a right-pointing connector of the same type on another word. The two connectors together form a link.

In Link Grammar, a linkage is a single successful parse of a sentence: a set of links in which none of the connecting arcs cross. The words of a syntactic structure are connected in such a way, that the links satisfy the linking requirements for each word of the sentence (satisfaction) that the links do not cross (planarity) and that all words form a connected graph (connectivity). The linking requirements of each word are contained in a dictionary. If you draw arcs between related words in a sentence (for instance, between an adjective and the noun it modifies), your sentence is ungrammatical if arcs cross one another. It is a grammatical if they don't.

A valid sentence may have more than one complete linkage, just as a sentence may have several meanings. The Link Grammar Parser (LGP) [31] is a syntactic parser of English, based on link grammar. Given a sentence, Link Grammar Parser assigns to it a syntactic structure, which consists of a set of labeled links connecting pairs of words. These links are used not only to identify parts of speech (nouns, verbs, and so on), but also to describe in detail the *function* of that word within the sentence The Link Grammar Parser also produces a *constituent* representation of a sentence (showing noun phrases, verb phrases, etc.). For example, in a Subject-Verb-Object (S-V-O) language like English, the verb would look left to form a subject link, and right to form an object link. Nouns would look right to complete the subject link, or left to complete the object link. A sample parser output is depicted in Fig. 2 for the sentence; *DGI is associated with mutations in DSPP*. The primary parts of speech are labeled with *.n* and *.v* to indicate that these words are nouns and verbs, respectively. The labels of the links between words indicate the type of link. For example, the *Mv* connector in this sentence indicates a connection between the verb and its modifying phrase. In this case, the verb *associated* is connected to *with mutations*, identifying a modifying phrase.

The parser uses a dictionary that contains the linking requirements of each word and the possible part of speech assignments for the entries. It has a dictionary of about 60000 word forms. Also, it has coverage of a wide variety of syntactic constructions. The parser is robust; it is able to skip over portions of the sentence that it cannot understand, and assign some structure to the rest of the sentence. It is able to handle unknown vocabulary, and make intelligent guesses from context and spelling about the syntactic categories of unknown words. It has knowledge of capitalization, numerical expressions, and a variety of punctuation symbols. The parser has an internal timer. If the timer runs down before a complete or partial linkage has been found, the parser will output whatever it has found so far (termed a fragmented linkage). Link Grammar Parser has many Applications such like: - AbiWord [32] checks, information extraction of biomedical texts and events described in news articles, as well as experimental machine translation.

The Link Grammar Parser itself is a complex piece of software implementing a complex theory of language. The PIELG system uses the Perl module Lingua::LinkParser [33]. It is a Perl module implementing the Link Grammar Parser under Linux platform. This module is available at CPAN [34] directly, embeds the parser. This module provides access to the parser Application Program Interface (API) [35] using Perl objects to easily analyze linkages. The API makes it easy to incorporate the parser into other applications. The API provides a set of basic data structures and function calls that allow the programmer to easily design a customized parser. The module organizes data returned from the parser API into an object hierarchy consisting of, in order, sentence, linkage, sub-linkage, and link.

The word dictionaries of the Link Grammar Parser are from conversational English which do not include the biological named entities. The LG parsers' lexicon can be easily enhanced to produce better parses for biomedical text [36]. We use two methods to extending the lexicon of the Link Grammar Parser. The first method is to use the LinkGrammar-WN [37] which aims to import lexical information from WordNet. WordNet [38] is an online lexical reference system that in recent years has become a popular tool for Artificial Intelligence (AI) researchers. The LinkGrammar-WN v1.0

release contains 14,392 noun word forms not available within the original LGP lexicon, thus increasing the size of the LGP

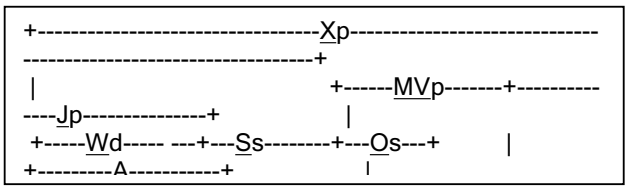

Fig. 3 The linkage given by the Link Grammar Parser for the sentence "DMP-1 regulates DSPP during odontoblast differentiation."

lexicon by 25%. The second extension method is to use the extended Link Grammar Parser [39] where the lexicon is extended by the lexicon from UMLS' [40] Specialist lexicon enabled to general-purpose language processing tools.

That enables Link Grammar Parser to manipulate medical text. The typically non-technical vocabularies must be augmented with a large medical lexicon. It applies a heuristic method to import lexical definitions of about 200,000 word senses into the LG dictionary, more than tripling its size from the UMLS's Specialist lexicon. This extension of Link Grammar's dictionary [41] effects on its performance. This extension can significantly improve efficiency, parsing performance and significantly reduced ambiguity. The extended parser manipulates biomedical text well.

IX. Interaction Word Tagger

Once protein names have been found, the relationships between them need to be ascertained. The words that convey a biologically significant action between two protein names are labeled as *interaction words*.

For example in sentence *DMP-1 regulates DSPP during early odontoblast differentiation*, the main verb *regulate*, describes the action performed by *DMP-1* on *DSPP*, is an example of interaction word. Some other example of interaction words are *bind*, *down-regulation*, *phosphrylation*, *bind*, *associate* and *complex* etc.

This can be done in a number of ways depending on the Information Extraction (IE) task. The system uses dictionary look-up method to identify interaction words in the sentences. We use a category/keyword dictionary for identifying terms describing interactions. The category/keyword dictionary is adapted from [16] with additional categories and keywords found to be prevalent in our corpus. A list of interaction words, which consists of 45 noun and 53 verb roots, was compiled from the literature. In order to broaden the list of potential interaction words, all inflected variants of known interaction words are also considered. Further, also all predictable spelling and derivational variants are considered. The dictionary is enriched manually with additional verbs that are known to refer to interactions. The *direct* and *indirect* physical interaction words are split into as shown in Table 1. *Example:* if the word *labeled* appears in the corpus as an interaction word, we also consider the words *label, labels, labeling, labeled* to be potential interaction words. Similarly, for the word *rebinds* we also consider the words *re-binds, rebind, re-bind, rebound, re-bound, rebinding, rebinding.*

TABLE 1
Direct and Indirect Interaction Words

Direct interaction verbs	Indirect interaction verbs
bind (bound)	induc(-es,-ed)
interact (-s,-ed)	trigger(-s,-ed)
stabilize (-s,-d)	block(s),
phosphorylate(-s,-d)	enhance(s)
ubiquinate(-s,-d)	synergize(s)
sumoylate(-s,-d)	cooperate(s)
degrade(-s,-d)	localizes
block(s).	regul(-ates,-ion)
	activate(s)
	inhibit(s)
	control(s)
	translocate(s)
	antagonize(s)
	amplif(-y,-ies)
	transduce(s)
	degrade(s)
	trigger(s).

X. Interaction Extractor (IE)

Interaction Extractor is the main component of the PIELG system**.** Its aim is to do deep analysis of the sentence to extract multiple and nested interactions from the sentence. It uses a series of mapping rules to extract information about protein-protein interactions. Those mapping rules could be applied to first identify the main verb in the sentence. Then, determining if those verbs are truly representing the interaction between two protein names (interaction words), in the text or not. If the main verb is not an interaction word then the algorithm detects all verbs in the sentence until detecting an interaction word.

Then, it uses the deep parse tree structure presented by the Link Grammar. It considers a thorough case based analysis of contents of various syntactic roles of the sentences like their subjects (S), verbs (V), objects (O) and modifying phrases (M) as well as their linguistically significant and meaningful combinations like S-V-O or S-V-M. Then finding and extracting protein–protein interactions only if a syntactic role (or meaningful combination) has at least two protein names and an interaction word.

A. The main verb is an interaction word

If the main verb is an interaction word, the system applied a set of rules to predict the subject for each of these. The scheme also helps to find out the object of the verb, when present, as well as the modifiers of all verbs and nouns. The prediction scheme begins once the sentence has been passed through the link parser and the linkage for that sentence has been obtained. The system uses the procedure proposed in [42] for identifying the main verb. After identifying the main verb, if it is marked as an interaction word from the interaction word tagger the system will continue to predict the subject and the object. After all the main verbs have been

identified, the subject, the object (if it exists) and the modifying phrases of both the verb

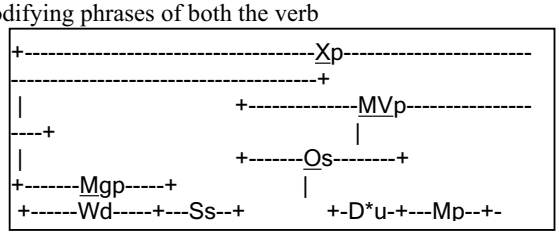

Fig. 4 The linkage (parse) given by the link grammar parser

and the object will also have to be predicted based on the rules presented in [42]. The rules are applied in hierarchical to identify the subjects (S), and objects (O) as well as all available modifying phrases (M) of the sentences. For example, if the input of the system is a clausal structure of the sentence: *DMP-1 regulates DSPP during odontoblast differentiation.* The sentence is parsed by the Link Grammar Parser (LGP) as in Fig. 3. The output will be in the form: (PROTEIN1, Interaction word, PROTEIN2) as explained in the following algorithm for Interaction Extractor.

1. The main verb *regulate* is identified.
2. The algorithm uses the links given by the LG parser to predict and obtains subject, object and modifying phrase as shown: Subject (S): *DMP-1* Object (O): *DSPP* which are both protein names. Modifying Phrase (MV): *odontoblast differentiation.*
3. The main verb is an interaction word.
4. The system tries to extract interaction between subject, verb and object combination (S-V-O).
5. Since the main verb is tagged as an interaction word, the interaction extractor uses the S-V-O composite role from to find and extract the following complete interaction: [DMP-1, regulate, DSPP]

B. The main verb is not an interaction word

If the main verb is not an interaction word each occurrence of the interaction word or one of *its* synonyms and hyponyms is to be one occurrence of the required interaction. *So,* by finding the subject, object as well as all available modifiers, almost all information about that instance of the event can be extracted from the document. For example, if the input of the system is a clausal structure of the sentence: "BMP enhances the expression of DSPP by directly stimulating DGI". The LG parser gives the output in the form of links between words as shown in Fig. 4. There are two interaction words the output of the system will be as follows [BMP, enhance, DMP-1] and [BMP, stimulate, DGI] as explained in the following algorithm for Interaction Extractor.

C. Phrasal-prepositional verbs patterns

Phrasal-prepositional verbs are made of: Verb + Particle + Preposition Combinations (Phrasal Verbs + Prepositions)**.** In this part we are interested in the case of preposition combinations. There are a small number of preposition combinations, such as *by-of*, *from-to* etc., which occur frequently within the clauses.

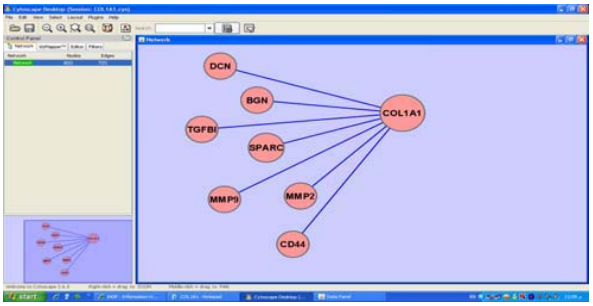

Fig. 5 COL1A1 Network generated by creating network manually

Those prepositional combinations are used to distinguish the agent, the predicate and the theme of the interactions. This is an example of prepositional combinations in phrasal-prepositional verbs *Gene expression **of** TGF-beta1 was sharply down-regulated **by** LTA **in** odontoblasts*. In this example, there is a preposition combination between *by* and *in*. There are two modifier phrases. The first one is *LTA* which is the subject of the passive voice. The second one is *odontoblasts* which is the modifier of the main verb. To solve this problem, the system uses *phrasal-prepositional verbs patterns* to find agent, predicate, theme and action to extract the interaction, for both active and passive voices. For preposition based deep extraction the system uses a pseudo code. The algorithm is repeated for each sentence of the text. This code starts by finding pattern corresponding to the prepositional combinations in the string. If the prepositional combinations exists the pattern (predefined patterns), then extract protein - protein interactions using the pattern. The predefined patterns for the previous sentence is the *by-in* pattern [(PROTEIN1 (predicate)) (is/are) or (was/were) (Interaction-Word (action)) by ... (PROTEIN2 (agent)) ... in... (Theme) ...]. The interaction extractor is able to extract the correct interaction (LTA, down-regulate, TGF-beta1, in, odontoblasts). The final step in the interaction extraction module is re-transformation. The main job of the re-transformer is to insert multiple words of protein names back after manipulation.

XI. Experiment

A. Corpus

We conducted experiments using corpus that is limited to abstracts describing human protein function. This corpus is selected to be about proteins currently considered to have roles in *dentine formation* process and involved in *dentinogenesis*. The selected corpus consists of 229 abstracts out of 1000 sentences, including abstract titles, with annotated proteins and interactions. Those 1000 sentences are sentences which contain one pair of proteins and one interaction word. If a sentence includes more than one interaction, all interactions are counted as answers. The extracted interactions correspond to 229 abstracts from the PubMed. Using abstracts ID's (PubMed ID's) of these 229 abstracts; we downloaded 527

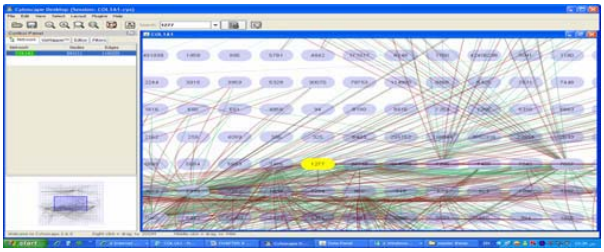

Fig. 6 COL1A1 Network generated by Importing Fixed-Format Network Files

records form BioGRID[1] database those interactions represented in the 229 abstracts. BioGRID database entries were downloaded as a flat file form. PIELG system extracted 399 interactions from these 229 abstracts.

B. Classification of treated forms

There are many types of surface variations that express the same information regardless of users' perspectives. The PIELG system covered nine classes based on the syntactical variation of the interaction words in various contexts as shown in Table 2.

XII. Evaluation

The evaluation process for the PIELG system is divided into two phases. The *first phase* is the evaluation of the information extraction performance by measuring the metrics Precision and Recall. And so, perform experimental evaluations with two other state-of-the-art extraction systems – the BioRAT and IntEx. The extracted results are compared with BioGRID[2] entries manually. If an interaction extracted by PIELG is not found in BioGRID, it could be that (a) it is a false-positive example, reducing the precision of PIELG; or (b) the interaction is missing from BioGRID. The latter case consists of interactions that are mentioned in papers, but have not been added to BioGRID. Like BioRAT and IntEx, We manually re-analyzed these records with no reference to BioGRID but instead we counted how many of PIELG's predictions were correctly extracted from the text. Table 3 shows the recall from these abstracts by PIELG, namely 47.4%, which is much higher than BioRAT (20.31%) and IntEx (26.94%). Table 4 shows the precision from these abstracts by PIELG, 62. 65%, which is a bit higher than BioRAT (55.07%) and but lower than IntEx (65.66%).

The *second phase* of the evaluation process for PIELG system was done by augmenting PIELG with a graphical package for drawing the extracted interactions. We used Cytoscape[3] which is a good tool for drawing directed graphs that can be adapted for extracting protein interaction information from sequence databases. We compare the extracted interactions from the PIELG system with the stored interactions in Cytoscape. The

[1] http://www.thebiogrid.org/
[2] http://www.thebiogrid.org/
[3] http://www.cytoscape.org/

TABLE II
Linguistic Variations of the Interaction Words in Various Contexts

Class (1):- Active main verb
• Entity 1 recognizes and **activates** Entity 2. • Our results indicate that Entity 1 **inhibits** the activated Entity 2.
Class (2):- Passive
• Entity 2 is activated by Entity 1. • The expression of Entity 1 is induced by Entity 2 in primary cultured dental pulp cells not in calvaria osteoblasts.
Class (3):- Modifying phrases of verbs
• Both *Entity1* and *Entity 2* interact with cell surface *Entity 3* through their amino termini. • *Entity 1* associates with the *Entity 2*.
Class (4):- After an Auxiliary Verb
• *Entity 1* may bind large amount of *Entity 2*.
Class (5):- Past particle
• *Entity 1* activated *Entity 2*.
Class (6):- Infinitive
• *Entity 1* is able to inhibit *Entity 2*.
Class (7):- Nominalization
• The Up-regulation of *Entity1* by *Entity2* in *Entity3* was activated. • *Entity1* up-regulation by *Entity2* in *Entity 3* was activated. • Dephosphorylation of *Entity1* by *Entity2* was carried out. • The phosphophoryn activation of *Entity1* implies this is a direct effect .
Class (8):- Preposition-based Patterns
• *Entity 1* was expressed by *Entity2* throughout *Entity 3* in *Entity4* . • *Entity 1* is cleaved into *Entity2* and *Entity3* in *Entity4.* • *Entity 1* is associated with mutations in *Entity2.* • *Entity1* is probably regulated by *Entity2* during dentinogenesis .
lass (9):- Nested interactions
• *Entity1* signals *Entity 2* by directly stimulating *Entity 3.* • *Entity1* prevents the decrease of *Entity2* and inhibits *Entity3.*

visualization process (Drawing Pathway Diagram) for a *specific protein* using Cytoscape composed of three stages:

1. *Edit a New Network*: creating an empty network in Cytoscape and manually add nodes and edges to draw the extracted interactions from the PIELG system. We gathered the extracted interaction prosperities from the PIELG system for Collagen, type I (COL1A1) (as an example). Then, a network for the extracted interactions is drawn using Cytoscape as shown in Fig. 5.

2. *Import Fixed-Format Network Files:* We retrieve the interaction prosperities of Collagen, type I (COL1A1) from BioGRID database. The interactions of Collagen, type I (COL1A1) are downloaded as a flat file from BioGRID database. Then we use Cytoscape to create networks by importing pre-existing, formatted network files as shown in Fig. 6.

3. *Import Networks from Web Services:* We use Cytoscape to create networks by importing networks from Web Service.

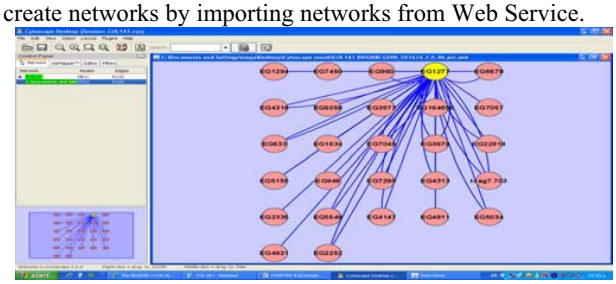

Fig. 7 COL1A1 Network generated by Entrez Gene data

We will retrieve protein-protein interaction networks from NCBI Entrez Gene. NCBI web service client uses this section to build networks. The network in Fig. 7 is generated from interaction data matching the keyword *Homo sapiens*. By comparing the previous three stages, we could notice that PIELG system misses out some interactions. That is due to both BioGRID and NCBI Entrez Gene contains protein interactions from both abstracts and full text. PIELG system is tested only on the abstracts. So it misses out some interactions that are only present in the full text. If those interactions are excluded, PIELG can have a higher recall.

XIII. Discussion

The highly technical terminology and the complex grammatical constructs that are present in the biomedical abstracts make the extraction task difficult, Even a simple sentence with a single verb can contain multiple and/or nested interactions. That's why PIELG is based on a deep parse tree structure presented by the Link Grammar and it considers a thorough case based analysis of contents of various syntactic roles of the sentences as well as their linguistically significant and meaningful combinations.

The heart of the system lies in the working of the rules for prediction of subject, object and their modifiers. The rules for the PIELG system are derived by running the link parser on abstracts of scientific papers including abstract and titles. Most missed interactions are caused by semantic problems. Currently it is not necessarily the case that more powerful grammars lead to better biochemical interaction extraction. Until recently, most information extraction systems for mining semantic relationships from texts of technical sublanguages avoided full parsing [43].

Semantic parsers for English language will be more useful and meaningful for the extraction tasks compared to syntactic parsers. But constructing semantic parser is a difficult task and this parser will be more domains dependent.

It is important to note, that using the Link Grammar in the proposed information extraction system makes it applicable to a large number of areas ranging from pathway analysis to clinical information and protein structure-function relationships. The time took for full parsing is also a problem for Information Extraction systems.

TABLEIII
Recall Comparison of IntEx and BioRAT from 229 Abstracts When Compared with BioGRID Database

Recall Results	PIELG		IntEx		BioRAT	
	Cases	Percent %	Cases	Percent %	Cases	Percent %
Match	250	47.4	142	26.94	79	20.31
No match	277	52.56	385	73.06	310	79.69
Totals	527	100	527	100	389	100

TABLEIV
Precision Comparison of IntEx and BioRAT from 229 Abstracts When Compared with BioGRID Database

Precision Results	PIELG		IntEx		BioRAT	
	Cases	Percent %	Cases	Percent %	Cases	Percent %
Correct	250	62. 65	262	65.66	239	55.07
Incorrect	149	47.45	137	34.34	195	44.93
Totals	399	100	399	100	434	100

The PIELG system success to extract detailed contextual attributes of interactions by interpreting modifiers like: location/position modifiers (*in, at, on, into, up, over...),* agent/accompaniment modifiers (by, with…), purpose modifiers (*for*…), and theme/association modifiers (*of*...). Finally, several issues make extracting interactions and relationships difficult since:

1. The task involves free text – hence there are many ways of stating the same fact.
2. The genre of text is not grammatically simple.
3. The text includes a lot of technical terminology unfamiliar to existing natural language processing systems.
4. Information may need to be combined across several sentences.
5. There are many sentences from which nothing should be extracted.
6. The abstracts of some papers are also used to take into consideration *technical style of writing*.

XIV. Conclusion

This paper presents a protein–protein interaction extraction system specially designed to process biomedical literature– PIELG. PIELG is based on a deep parse tree structure presented by the Link Grammar and it considers a thorough case based analysis of contents of various syntactic roles of the sentences as well as their linguistically significant and meaningful combinations. The PIELG system covers many linguistic variations of the interaction words in various contexts. It covers nine classes based on constituents of the verbs. It succeeded to extract detailed contextual attributes of interactions by interpreting modifiers. However, we have developed and evaluated PIELG, for analysis of biomedical literature. Experimental evaluations of the PIELG system with the-state-of-the-art systems – the BioRAT and IntEx indicate that PIELG's performance is better. From the results of the PIELG evaluation process, we can conclude that its performance is satisfactory for the real-time PubMed processing. The results also shows that syntactic role-based

approach compounded with linguistically sound interpretation rules applied on the full sentence's parse can achieve better performance than existing systems which are based on manually engineered patterns. Those systems are both costly to develop and are not as scalable as the automated mechanisms presented in this paper. The high precision of the PIELG stems from its full-sentence parsing approach and presently comes at the price of a lower recall rate. However, the volume of data can be increased several times by implementing a reasonable set of improvements to the system, extending the protein names dictionary towards the description of experimental data. We estimate that the current PIELG**'s** coverage rate could be enhanced by increasing the lexicon size of the Link Grammar Parser, improving its quality, and by slightly improving its grammar. In addition, even with its coverage PIELG is still immediately applicable for an information extraction task. Also, utilization of protein names dictionary provides an ability to change the scope of extracted information, making entire system more flexible, and along with high performance, favorably differentiates.

REFERENCES

[1] D. Eisenberg, "DIP - Database of interacting Proteins," *University of California,* http://dip.doe-mbi.ucla.edu. 1999.
[2] "BOND - Biomolecular Object network databank," *Thomson Scientific,* http://www.bind.ca. 1999.
[3] T. Igarashi, and H. Kaminuma, "CSNDB - Cell Signaling Networks Database," *National Institute of Health Sciences, Japan,* http://geo.nihs.go.jp/csndb. 1998.
[4] H. Higashi-ku, and Fukuoka, "Signaling PAthway Database (SPAD)," *Kyushu University,* http: //www.grt.Kyushu-u.ac.jp/ eny-doc. 1998.
[5] "MEDLINE - National Library of Medicine (NLM)," *National Institutes of Health (NIH),* http://www.nlm.nih.gov. 1993.
[6] "PubMed Centeral," *(NCBI),* http://www.ncbi.nlm.nih.gov /sites/entrez/. 1988.
[7] K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi., "Toward Information Extraction: Identifying protein names from biological papers," *Proc. Pacific Symp. Biocomputing*, pp. 707-718, 1998.
[8] C. Blaschke, M. Andrade, C. Ouzounis, and A. Valencia, "Automatic extraction of biological information from scientific Text: Protein-Protein interactions," *Proc. AAAI Conf. Intelligence sys. in Molecular biology*, pp. 60-67, 1999.
[9] T. Sekimizu, H.S. Park, and J. Tsujii, "Identifying the Interaction between Genes and gens Products based on Frequently Seen Verbs in MEDLINE Abstracts," *Genome inform Ser Workshop Genome inform.,* pp. 62–71, 1998.
[10] N.S. Kiong, M. Wong, "Toward Routine Automatic pathway Discovery from on-line scientific text Abstractsd," *Proc. Tenth Inter. Workshop Genome inform*., pp. 104-112, 1999.
[11] A. Clegg, and A. Shepherd "Benchmarking Natural-Language Parsers for biological Applications using dependency Graphs," *J. BMC Bioinformatics,* vol.8- pp. 24, Jan 2007.
[12] J. Thomas, "D. Milward, C.A. Ouzounis, S. Pulman, and M. Caroll, "Automatic Extraction of Protein Interactions from Scientific Abstracts", *Pacific Symp. Biocomputing*, pp. 541-552, 2000.
[13] L. Gondy, C. Hsinchun, and D. Jesse, "A Shallow Parser Based on Closed-Class Words to Capture Relations in Biomedical Text," *J. Biomedical Informatics,* vol.36, pp. 145-158, August 2004.
[14] G. Claudio, L. Alberto, and Lorenza Romano, "Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature," *Proc. 11th Conf. the European Chapter of the Association for Computational Linguistics (EACL 2006),* 2006.
[15] C. Friedman, "MedLEE - A Medical Language Extraction and Encoding System," *Columbia University, and Queens College of CUNY,* http://lucid.cpmc.columbia.edu/medlee. 1995.
[16] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky, "GENIES: A Natural-Language Processing System for the Extraction of Molecular Pathways from Journal Articles," *J. Bioinformatics*, vol. 17, pp. 74-82(9), June 2001.
[17] C. Friedman, "MedScan - A Medical Language Extraction and Encoding System," *Columbia University, and Queens College of CUNY,* http://www.ariadnegenomics.com/products/medscan. 1995.
[18] A. Rzhetsky, "Geneways: A search engine and information extraction tool for biological research," *Columbia Genome Center,* http://geneways.genomecenter.columbia.edu. 2005.
[19] D. Corney, D. Jones and B. Buxton, "BioRAT System," *Columbia Genome Center,* http://bioinf.cs.ucl. Ac.uk/biorat. 2005.
[20] J. Xiao, J. Su, G. Zhou and C. Tan, "Protein-Protein Interaction Extraction: A Supervised Learning Approach," *Proc. first Inter. Symp. Semantic mining in Biomedicine (SMBM 2005)*, pp. 51-59, 2005.
[21] J. Ding, D. Berleant, J. Xu, and A.W. Fulmer, "Extracting Biochemical Interactions from MEDLINE Using a Link Grammar Parser," *Proc. 15th IEEE Inter. Conf. Tools with Artificial Intelligence (ICTAI'03)*, pp. 467-471, 2003.
[22] Y.C. Lin, C.L. Peng, C.Y. Kao, H.F. Juan,H. C. Huang, "ProtExt: A system for protein-protein interactionextraction from PubMed abstracts" , *Proc. 12th Inter. Conf. Intelligent Systems for Molecular Biology (ISMB) and Conf. Computational Biology (ECCB),* 2005.
[23] S.T. Ahmed, D. Chidambaram, H. Davulcu, and C. Baral, "IntEx: A Syntactic Role Driven Protein-Protein Interaction Extractor for Bio-Medical Text," *Proc. ACL-ISMB workshop linking biological literature, ontologies and databases: Mining biological semantics,* pp. 54-61, 2005.
[24] Z. Yang, H. Lin, and B. Wu, "BioPPIExtractor: A Protein–Protein Interaction Extraction System for PubMed Abstracts," *J. Expert Systems with Applications*, Article in press, doi: 10.1016 /j.eswa.2007.12.014. 23 Dec. 2007.
[25] "LocusLink - Database of genes," *(NCBI),* http://www.ncbi.nlm.nih.gov/sites/ entrez?db=gene. 1988.
[26] "Universal Protein Resource (UniProt)," *European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR),* http://beta.uniprot.org. 2002.
[27] "ExPASy Proteomics Server*," Swiss Institute of Bioinformatics (SIB*), http://www.expasy.ch. 2003.
[28] R. Hoffmann and A. Valencia, "A Gene Network for Navigating the Literature - iHOP," *Nature Genetics*, http://www.ihop-net.org. 2004.
[29] D. Temperley, D. Sleator, and J. Lafferty, "Link Grammar," *Carnegie Mellon University*, http://www.link.cs.cmu.Edu/link. 1998.
[30] D. Sleator, and D. Temperley, "Parsing English with a Link Grammar," *Third International Workshop on Parsing Technologies*, pp. 277-292, 1993.
[31] D. Grinberg, J. Lafferty, and D. Sleator, "A Robust Parsing Algorithm for Link Grammars," *Proc. second inter. colloquium on grammatical inference and applications,* vol. 862, pp. 78–92,1995.
[32] D. Temperley, D. Sleator, and J. Lafferty, "Abiword- word processor for everyone," *Carnegie Mellon University*, http://www.abisource.com. 1998.
[33] D. Brian "Lingua::LinkParser- Perl module implementing the Link Grammar Parser," *Carnegie Mellon University*, http://search.cpan.org/~dbrian/Lingua- LinkParser 1.08. 2004.
[34] "CPAN - Comprehensive Perl Archive Network," http:// www.cpan.org. 1995
[35] D. Temperley, D. Sleator, and J. Lafferty, "The parser Application Program Interface (API)," *Carnegie Mellon University*, http://www.abisource.com/projects/link-grammar/api/index.html. 1998.
[36] S. Pyysalo, T. Salakoski, S. Aubin and A. Nazarenko, "Lexical Adaptation of Link Grammar to the Biomedical Sublanguage: a Comparative Evaluation of Three Approaches," *J. BMC Bioinformatics*, vol. *7,* pp. 60-67, November 2006.
[37] E. Turner, "The LinkGrammar-WN," http://www.eturner.net/ linkgrammar-wn.2007
[38] "WordNet-a lexical database for the English language," *Princeton University*, http://wordnet.princeton.edu. 2006
[39] P. Szolovits, "Adding a Medical Lexicon to an English Parser," *Proc. AMIA 2003 Annual Symposium.* pp. 639-643 ,2003
[40] "UMLS'-Unified Medical Language System," *U.S. National Library of Medicine*, http://umlsinfo.nlm.nih.gov.1999.
[41] S. Pyysalo, F. Ginter, T. Pahikkala, J. Boberg, J. J¨arvinen, and T. Salakoski, "Evaluation of Two Dependency Parsers on Biomedical Corpus Targeted at Protein—Protein interactions," *J. Inter. Medical Informatics*, Vol. 75, Issue 6, pp. 430-442, June 2005.

[42] V. Harsha, Madhyastha, N. Balakrishnan, K.R. Ramakrishnan "Event Information Extraction Using Link Grammar," *Inter. Workshop Research Issues in Data Eng.: Multi-lingual Information Management (RIDE'03),* pp. 16- 22, 2003.
[43] L. Hirschman, J.C. Park, J. Tsujii, L. Wong, and C.H. Wu5, "Accomplishments and Challenges in Literature Data Mining for Biology." *J. Bioinformatics,* vol. 18, pp. 1553-1561, June 2002.

Rania Ahmed Abul Seoud received the B.S. degrees in Electrical Engineering- Communications and Electronics Department at Cairo University – Fayoum Branch in 1998 and M.S.E. degrees in Computer Engineering at Cairo University in 2005. She is currently a Doctoral student at Cairo University. She worked as a Demonstrator and a Teaching Assistant in Electrical Engineering Department of Miser University for Science and Technology, Egypt since 1998. Currently, she is a Teaching Assistance in Electrical Engineering Department of Fayoum University, Egypt. She has a two published paper in that field. One is in the Third Cairo Int'l Conference on Biomedical Engineering 2006, (CIBEC06), Cairo, Egypt. The second is in The 2007 International Conference on Computer Engineering & Systems (ICCES'07), Cairo, Egypt. Her areas of interest in research are Artificial Intelligence, Natural Language Processing and Application of Artificial Intelligence to computational biology and bioinformatics.

Dr. Nahed H. Solouma received her B.Sc. and M.Sc. degrees from the Biomedical Engineering Department at Cairo University in 1990 and 1996 respectively. She received another M.Sc. in Laser applications in medicine and biology from the National Institute of Laser Enhanced Sciences, Cairo University in 1997. Her Ph.D. degree was from the Biomedical Engineering department, Cairo University in 2001. She worked as a Lecturer and then an Associate Professor in NILES, Cairo University since 2001 till 2007. She is currently in sabbatical leave in Imperial College as an academic visitor in the Biomedical Engineering Institute. Her research interest is in biological modeling and medical data processing.

Abou-Bakr M. Youssef graduated from Cairo University — Faculty of Engineering — Electrical Engineering & Communication Department with Distinction with Honor in 1974. He joined the Faculty of Medicine and worked in parallel on engineering research and received M.S.E.E. degree in 1978, and the Ph.D. degree in biomedical engineering (MIT fellow) in 1982—Cairo University. While continuing his engineering career, he was also able to graduate from the Medical School, Cairo University 1980, and earned the MBBH, Distinction with Honors, and the foreign equivalent, ECFMG, in 1981. In 1984, he received the M.S. degree in radiology from Cairo University, and the MD degree from Heildelberg University, Heildelberg, Germany, in 1989. He conducted medical research in the German cancer research center (DKFZ) from 1987–1993. He contributed to the establishment of the Biomedical Engineering Department at Cairo University in 1977, developed courses of biomedical equipment and clinical engineering as well as medical imaging, and joined in the establishment of the Technology Transfer Focus in Medical School-Cairo University, in cooperation with the National Science Foundation in 1978–1980. He gained clinical experience in his research and educational life span by working as the Head of the Ultrasound Department in major hospitals in Cairo as well as within private clinics. In 1990–1992, he participated in the Health Care Development Plan, in cooperation with the Egyptian Scientific Research Academy, Cairo. Many of his researches were published in international and local scientific magazines. Dr. Youssef was registered as a Diagnostic Medical Sonographer RDMS in 1995. As a member of the American Institute of Ultrasonography in Medicine (AIUM), in 1997 he was elected as a Senior Member. He established the Biomedical Division of International Electronics, as a practical and technical service provider for medical industry and health care community in 1991. He was appointed CEO for International Electronic Company Biomedical group, Cairo. He participated with the Information & Decision Support Center—Egyptian Council of Ministers, in the release of medical software and expert systems in 1994. He was awarded from the government twice for the scientific research—Cairo University 1998 and the national award for advanced technological sciences—2001.

Dr. Yasser M. Kadah received his B.Sc. and M.Sc. degrees from the Biomedical Engineering Department at Cairo University in 1989 and 1992 respectively. He received his Ph.D. in Biomedical Engineering from the University of Minnesota in 1997. He worked as a research assistant with the Department of Radiology, University of Minnesota Medical School in 1996-1997 and as a post-doctoral fellow at the Center of Magnetic Resonance Research at the University of Minnesota in 1998. He has been with the Biomedical Engineering department at Cairo University since 1998 where he is currently an Associate Professor. He worked between 1998 and 2002 as the director of research and development at IBE Technologies, Egypt. He was on leave of absence between Dec. 2002 and July 2004 to work at the Biomedical Imaging Technology Center of the W.H. Coulter department of Biomedical Engineering at Emory University and Georgia Institute of Technology. He is currently back teaching at Cairo University and directing the research activities at International Biomedical Engineering Technologies (IBE Tech) of Giza, Egypt. Along his career, he received several awards and recognitions including the record for highest undergraduate GPA in the department of Biomedical Engineering at Cairo University, the IDB Merit Scholarship (1993-1996), his biography was selected to appear in Marquis Who's Who in the World (2002), as well as the E.K. Zaviosky Stipend from the International Society of Magnetic Resonance In imaging (2002). He is an active member of the IEEE, ISMRM, and SPIE. His research interests include medical imaging and in particular MRI and ultrasound, and multi-dimensional signal processing for biomedical applications.