Enhanced PIELG: A Protein Interaction Extraction System using a Link Grammar Parser from biomedical abstracts

R. A. Abul Seoud and Y. M. Kadah
[1]*Department of Computer Engineering, Faculty of Engineering, Fayoum University, Fayoum, Egypt*
.[2]*Department of Biomedical Engineering, Faculty of Engineering, Cairo University, Giza, Egypt*
r-abulseoud@k-space.org

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 investigates the effect of adding a new module - Complex Sentence Processor (CSP) – to the PIELG system. PIELG is 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 enhanced to 49.33 % and 65.16 % respectively. Experimental evaluations with two other state-of-the-art extraction systems indicate that enhanced PIELG system achieves better performance. The result shows that the performance is remarkably promising.**

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

I. INTRODUCTION

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. This information is scattered throughout numerous publications in scientific journals. Hence, manual collection of this updated data is time consuming. Thus, it is great to have an automated information extraction system to extract updated biological data about interacted protein involved in dentine formation process. This will provide molecular biologists in *tissue engineering laboratories* with information which is used in their respective applications.

Many approaches have been proposed for information extraction (IE) ranging from simple statistical methods to advanced natural language processing (NLP) systems. The first step done towards IE was to recognize the names of proteins, genes, drugs and other molecules [1]. The next step was to recognize interaction events between such entities [2]. A number of groups reported application of pattern-matching-based systems for protein-function information extraction [2], [3], [4]. In the last few years, NLP has become a rapidly-expanding field within bioinformatics [5]. Many NLP approaches have been used to extract data from biological texts. More advanced systems utilizing shallow parsing techniques have been described to extract protein interactions [6]. The MedLee system [7] used domain-specific *context-sensitive grammars*. The Pathway Assist system uses an NLP system, MedScan [8] for the bio-medical domain that tags the entities in text and produces a semantic tree. Recently, it has been extended as GeneWays [9], which also provides a Web interface that allows users to search and submit papers of interest for analysis. The BioRAT [10] system uses manually engineered templates that combine lexical and semantic information to identify protein interactions. *Machine learning approaches* have also been used to learn extraction rules from user tagged training data [11]. Recently, extraction systems have also used *Link Grammar* to identify interactions between proteins. Ding et al. proposed an interaction extraction method based on Link Grammar Parser [12]. The ProtExt system [13] extending the idea of Ding et al., 2003. Both The IntEx [14] system and BioPPIExtractor system [15] use a Link Grammar Parser to extracts complete interactions.

This paper presents an enhanced version of PIELG system. PIELG is a fully automated extraction system to extract information about protein-protein interactions in biomedical text using the Link Grammar Parser. PIELG is purely implemented with Perl under Linux platform. PIELG investigates and classifies forms which are needed to extract interacting protein pairs. The enhanced version is obtained by adding a new module to PIELG. This module is Complex Sentence Processor (CSP). This module enables PIELG to handle complex sentence structures and to extract multiple and nested interactions specified in a sentence. This approach is based on first splitting complex sentences into simple clausal structures made up of syntactic roles. Then, the system tags biological entities with the help of biomedical and linguistic ontologies. Finally, the system extracts complete interactions by analyzing the matching contents of syntactic roles and their linguistically significant combinations. The enhanced PIELG handles complex sentences and extracts multiple and nested interactions specified in a sentence. The recall and precision of PIELG are 47.4% and 62. 65%. While our experimental results show that the recall and precision of the enhanced PIELG are enhanced to 49.33 % and 65.16 % respectively. Experimental evaluations with two other state-of-the-art extraction systems indicate that enhanced PIELG achieves better performance.

II. METHODOLOGY

A. system overview

A typical session in using PIELG involves the user providing an initial search specification (keywords). 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. Then PIELG extracts interaction

978-1-4244-2695-9/08/$25.00 ©2008 IEEE

information from abstracts and titles of scientific papers, and presents the extracted information in textual forms. The architecture of the enhanced PIELG is shown in Fig. 1. The details of the main modules Sentence Segmentation and tokenization, Named Entity identification and conversion, Simple Filtering and Transformation, Preprocessor, Link Grammar Parser and Link Grammar and Interaction Word Tagger are explained in [16]. Here we will explain the details of two modules CSP and IE.

C. Complex Sentence Processor (CSP)

CSP module takes text from the abstracts as the input and outputs a simple sentence for each abstract. It acts on the tagged and pre-processed text to produce simple sentences. The sub-system uses the Link Grammar Parser (LGP) to process complex sentences. The CSP splits the complex sentences into the internal clause format representing simple sentences using the link grammar parser. The LGP produces links between the words in a sentence that correspond to the syntactic structure of the sentence via subject, object, determiner etc. The link grammar parser is operated in union mode to get the combined linkage of all sub-linkages. The linkages from the parser are then sent to the CSP for analysis.

CSP identifies specific links from the linkage output of the LGP and follows the links to obtain the syntactic constructs such as Subject, Verb, Object and Verb modifying phrase. CSP follows a *verb-based approach* to extract the simple sentences. The assumption is that the verb represents the central idea of a clause. A sentence is identified as complex if there is more than one verb in the sentence. The links from the verbs are followed to get the subject, object and modifying phrase. The clause format used to represent simple sentences is: *Subject + Verb + Object + Modifying phrase to the verb.* The modifying phrases to the verb can be adverbial, prepositional or adjectival phrases. The components can be a single word or multi-word phrases. Each of the components, once identified is expanded to include multi-word phrases. The links used to expand the words to phrases are the determiners, adjectives or prepositional attachments. Then, the syntactic constructs are the representation of a simple sentence. A trace of the process for a sample sentence *"DGI is associated with mutations in DSPP and a gene encoding DSPP is processed into two proteins DSP and DPP.*" is given in Fig. 2. This example illustrates the working of the CSP. *DGI*, *gene encoding DSPP* and *DPP* are tagged as gene names by the Named entity recognition. The Named entity conversation module convert *gene encoding DSPP* to a personal noun which is familiar to the parser. Hence the sentence passes the preprocessor and simple filter and is sent to the link grammar parser. The link grammar parser is operated in union mode to get the combined linkage of all sub-linkages. The linkage obtained shows two clauses as part of the sentence, as indicated by the two S links from the verbs ' processed', ' associated' and 'is'. The subjects, objects and modifying phrases for each of these verbs are identified. The final output of CSP is a set of two clauses for the sentence as seen in the Fig. 2. The interaction word tagger identifies that each sentence has an interaction word (' processed',' associated') using interaction words dictionary. The interaction extractor module produced the output as shown in Fig. 2.The sentence is also in passive voice, and

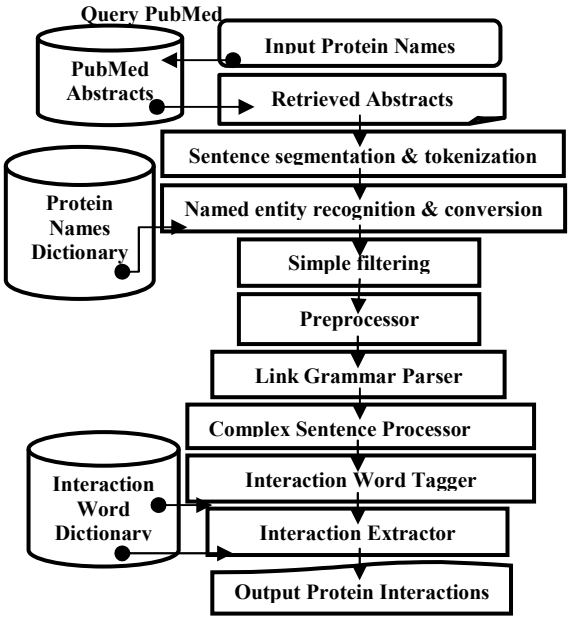

Fig. 1. System architecture

Coordinating conjunctions are also taken care of, as shown in the example.

a) Interaction Extractor (IE)

Interaction Extractor is the main component of PIELG**.** The enhanced PIELG aims to do deep analysis of the sentence to extract multiple and nested interactions from the sentence. The input to this module is simple clauses produced by the CSP module with marked interaction words. The output is the protein interactions founded in those simple clauses. PIELG uses a series of mapping rules to extract information about protein- protein interactions. Those mapping rules are built to first identify the main verb in the sentence. For that case, the system uses the procedure proposed in [17] for identifying the main verb. Then, the mapping rules are used

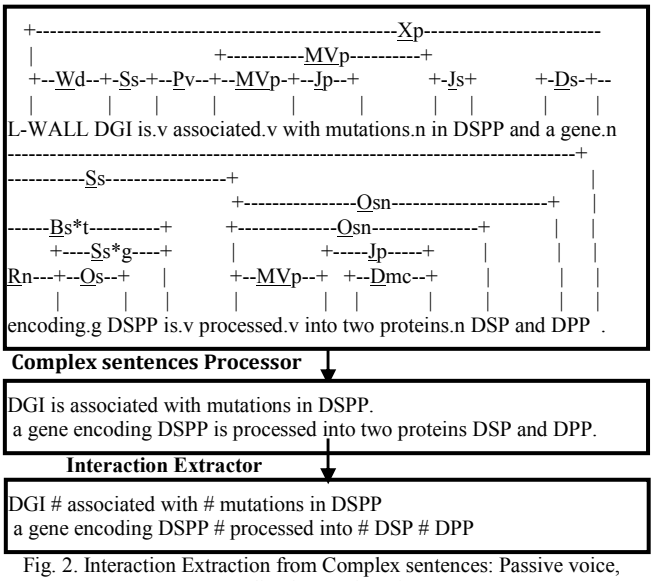

Fig. 2. Interaction Extraction from Complex sentences: Passive voice, coordinating conjunctions

to determine if those main verbs are truly representing the interaction in the text or not. If the main verb is not an interaction word then PIELG uses the designed algorithm to detect all verbs in the sentence until detecting 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. After identifying all interaction words in the sentence, the system applied another set of mapping rules to predict the subject, object of the verb [17], as well as the modifiers of all verbs and nouns. If subject, object or modifying phrase role itself is a protein name, then the system will extract interaction from the combination of subject-verb-object (*S-V-O) or* subject-verb- modifying phrase *(S-V-M).* So, almost all information about that instance of the protein - protein interactions can be extracted from the text, for both active and passive voices. For example, the sentence: *DMP-1 regulates DSPP during odontoblast differentiation.* The output will be in the form: [DMP-1, regulate, DSPP].

PIELG has taken various possible cases in which interaction words can occur in a sentence such as the nominalization form (e.g. converting an interaction word to a noun phrase). For example the sentence: "The Up-regulation of *Entity1* by *Entity2* ". The system treats the cases whether the theme appears before the *nominzlized interaction word* or after it. The system is also able to identify *dephosphorylation* relations. PIELG treats the problem of prepositions combinations such as *by-of*, *from-to* etc. For example the sentence:" *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, *LTA*-subject of the passive voice and *odontoblasts*-modifier of the main verb. So, the system uses *phrasal-prepositional verbs patterns* to find agent, predicate, theme and action to extract all information about the interaction. Finally, PIELG covered nine classes based on the syntactical variation of the interaction words in various contexts. These nine classes are:- interaction words is in active and passive voice, Modifying phrases of interaction words, interaction words After an Auxiliary Verb, interaction words in the past particle form , interaction words in the Infinitive, Nominalization form of interaction words, Preposition-based Patterns , Nested interactions.

III. RESULTS

a) The evaluation process of PIELG system

We conducted experiments using corpus that is limited to abstracts describing human protein function having roles in *dentine formation* process and involved in *dentinogenesis*. The corpus of the PIELG is selected in order to evaluate the protein-protein interaction method. The selected corpus to evaluate PIELG was consisted of 229 abstracts out of 1000 sentences, including abstract titles. The extracted interactions correspond to those 229 abstracts from the PubMed. Using abstracts ID's (PubMed ID's) of these 229 abstracts; we downloaded 527 records form BioGRID[1] database those interactions represented in the 229 abstracts. BioGRID database entries were downloaded as a flat file form. PIELG extracted 399 interactions from these 229 abstracts. For fair comparison, we have also limited our protein name dictionary used for tagging genes to the iHOP[2] entries. The evaluation process for PIELG was divided into two phases. The *first phase* was 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 were compared with BioGRID[3] entries manually. Tables 1 and 2 present the evaluation results as compared with the BioRAT and IntEx systems. The *second phase* of the evaluation process for PIELG was done by augmenting PIELG with a graphical package for drawing the extracted interactions. We used Cytoscape[4] which is a good tool for drawing directed graphs that can be adapted for extracting protein interaction information from sequence databases. We compared the extracted interactions from PIELG with the stored interactions in Cytoscape. The visualization process (Drawing Pathway Diagram) for a *specific protein* using Cytoscape composed of three stages: *Editing a New Network*, *Importing Fixed-Format Network Files, and Importing Networks from Web Services.* The details of the evaluation process are represented in [18]. By comparing the previous three stages, we could notice PIELG misses out some interactions. That is due to both BioGRID and NCBI Entrez Gene contains protein interactions from both abstracts and full text. PIELG 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.

TABLE 3: RECALL COMPARISON

Recall Results	PIELG Cases	PIELG %	Enhanced PIELG Cases	Enhanced PIELG %	IntEx Cases	IntEx %	BioRAT Cases	BioRAT %
Match	250	47.4	262	49.7	142	26.9	79	20.3
No Match	277	52.5	265	50.2	385	73.0	310	79.6
Totals	527	100	527	100	527	100	389	100

TABLE 4: PRECISION COMPARISON

Precision Results	PIELG Cases	PIELG %	Enhanced PIELG Cases	Enhanced PIELG %	IntEx Cases	IntEx %	BioRAT Cases	BioRAT %
Correct	250	62. 6	262	65.6	262	65.6	239	55.0
Incorrect	149	47.4	137	34.3	137	34.3	195	44.9
Totals	399	100	399	100	399	100	434	100

b) The evaluation process of the enhanced PIELG system

The evaluation of the information extraction performance is done 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 were compared with BioGRID[5] entries manuallyas shown in Tables 1 and 2. Table 1 shows the recall from these abstracts by the enhanced PIELG, namely 49.33%, which is much higher than PIELG (47.43%), BioRAT (20.31%) and IntEx (26.94%). Table 2 shows the precision from these abstracts by the enhanced PIELG, 65.63%, which is a bit higher than PIELG, 62. 65%, and BioRAT (55.07%) and but the same as IntEx (65.66%).

IV. 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.

[1] http://www.thebiogrid.org/
[2] http://www.ihop-net.org
[3] http://www.thebiogrid.org/
[4] http://www.cytoscape.org/
[5] http://www.thebiogrid.org/

978-1-4244-2695-9/08/$25.00 ©2008 IEEE

The heart of the system lies in the working of the rules for prediction of subject, object and their modifiers. The rules for PIELG are derived by running the link parser on abstracts of scientific papers including abstract and titles. Most missed interactions are caused by semantic problems because LGP is a syntactic parser. Other systems which use semantic parser rather than syntactic 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. 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 [19]**.** Also, 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. Also, PIELG successes to extract detailed contextual attributes of interactions by interpreting modifiers like. Also, after adding the new module *CSP* to PIELG the Precision and Recall are enhanced. Where, CSP uses Link Grammar to simplify complex sentences by splitting complex sentences into a collection of simple sentence.

V. 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. PIELG 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 PIELG 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] 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.

[2] 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.

[3] 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.

[4] 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.

[5] A. Clegg, and A. Shepherd “Benchmarking Natural-Language Parsers for biological Applications using dependency Graphs,” *J. BMC Bioinformatics,* vol.8- pp. 24, Jan 2007.

[6] 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.

[7] C. Friedman, “MedLEE - A Medical Language Extraction and Encoding System,” *Columbia University, and Queens College of CUNY,* http://lucid.cpmc.columbia.edu/medlee. 1995.

[8] C. Friedman, “MedScan - A Medical Language Extraction and Encoding System,” *Columbia University, and Queens College of CUNY,* http://www.ariadnegenomics.com/products/medscan. 1995.

[9] A. Rzhetsky, “Geneways: A search engine and information extraction tool for biological research,” *Columbia Genome Center,* http://geneways.genomecenter.columbia.edu. 2005.

[10] D. Corney, D. Jones and B. Buxton, “BioRAT System,” *Columbia Genome Center,* http://bioinf.cs.ucl. Ac.uk/biorat. 2005.

[11] 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.

[12] 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.

[13] 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.

[14] 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: Mining biological semantics,* pp. 54-61, 2005.

[15] 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.

[16] Rania Abulseoud, Abou-Bakr Youssef and Yasser M. Kadah, “Extraction of protein interaction information from unstructured text using a link grammar parser,” Proc. of the 2007 International Conference on Computer Engineering & Systems (ICCES'07), Cairo, Egypt, November 2007.

[17] V. Harsha, Madhyastha, N. Balakrishnan, K.R. Ramakrishnan “Event Information Extraction Using Link Grammar,” *Inter. Workshop Research Issues in Data Eng.: (RIDE'03),* pp. 16- 22, 2003.

[18] Rania A. Abul Seoud, Nahed H. Solouma, Abou-Baker M. Youssef, Yasser M. Kadah, "PIELG: A Protein Interaction Extraction System using a Link Grammar Parser from Biomedical Abstracts", *International Journal of Biomedical Sciences*, Vol. 3, No.3., pp. 169-179, Summer 2008.

[19] 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.