Research Journal of Applied Sciences, Engineering and Technology 8(22): 2282-2289, 2014
DOI:10.19026/rjaset.8.1230
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: September 13, 2014 Accepted: September 20, 2014 Published: December 15, 2014

Research Article

Dictionary and Gene Ontology Based Similarity for Named Entity Relationship Protein-protein Interaction Prediction from Biotext Corpus

[1]Smt K. Prabavathy and [2]P. Sumathi
[1]Department of Computer Science, Manonmanium Sundaranar University, Tirunelveli,
[2]Department of Computer Science, Government Arts College, Coimbatore, Tamil Nadu 627012, India

Abstract: Protein-protein interactions functions as a significant key role in several biological systems. These involves in complex formation and many pathways which are used to perform biological processes. By accurate identification of the set of interacting proteins can get rid of new light on the functional role of various proteins in the complex surroundings of the cell. The ability to construct biologically consequential gene networks and identification of the exact relationship in the gene network is critical for present-day systems biology. In earlier research, the power of presented gene modules to shed light on the functioning of complex biological systems is studied. Most of modules in these networks have shown small link with meaningful biological function, because these methods doesn't exactly calculate the semantic relationship between the entities. In order to overcome these problems and improve the PPI results in the biotext corpus a new method is proposed in this research. The proposed method which directly incorporates Gene Ontology (GO) annotation in construction of gene modules and Dictionary-based text is proposed to extract biotext information. Dictionary-Based Text and Gene Ontology (DBTGO) approach that integrates with various gene-gene pairwise similarity values, protein-protein interaction relationship obtained from gene expression, in order to gain better biotext information retrieval result. A result analysis has been carried out on Biotext Project at UC Berkley. Testing the DBTGO algorithm indicates that it is able to improve PPI relationship identification result with all previously suggested methods in terms of the precision, recall, F measure and Normalized Discounted Cumulative Gain (NDCG). The proposed DBTGO algorithm can facilitate comprehensive and in-depth analysis of high throughput experimental data at the gene network level.

Keywords: Biotext corpus, gene network, gene ontology, Information Extraction (IE), Named Entity Relationship (NER), preprocessing, Protein-Protein Interaction (PPI), word-sense disambiguator

INTRODUCTION

Recent technical progression in high-throughput research has been effectively bringing about a revolution regarding modern biological and biomedical studies. By means of microarrays, expression plane of thousands of genes can be measured at the same time (Schulze and Downward, 2001). Among yeast two-hybrid assays, pairwise interactions between thousands of proteins can be detected analytically (Ito *et al*., 2001; Uetz *et al*., 2000). With tandem mass spectrometry, a bulky number of proteins can be sequenced and distinguished rapidly (Aebersold and Mann, 2003). Each type of data explains the biological system under investigation from a specific point of view.

Recent advances in biomedical research methods have greatly accelerated the rate at which new information is published. Several research articles for biomedical text mining have been published in earlier years (Winnenburg *et al*., 2008; Zweigenbaum *et al*., 2007; Ananiadou *et al*., 2010). The concentration of the Biotext community has recently focused on Information Extraction (IE), especially on the growth of IE systems for mining protein-protein interactions. Information extraction systems identify entities and their relationships from free text without human intervention, producing a structured representation of the relevant information stated in the input text. For example, support researchers in the background searches and provide as the basis for the inference of semantic relationships, such as candidate pathways, stated across several publications. An annotated corpus is a group of texts that have been improved with markup indicating linguistic and domain information such as syntactic structure, named entity identification and entity relationships.

The IE systems are mostly used to extract the information from the text are named entity recognition, parsing and domain analysis. The named entity recognition finds the entities whose relationships are to be found is named as Named Entity Relationship

Corresponding Author: Smt K. Prabavathy, Department of Computer Science, Manonmanium Sundaranar University, Tirunelveli, Tamil Nadu 627012, India

(NER), parsing recovers the syntactic structure of the text and domain analysis extracts the relationships among the named entities by means of the information from the other processing steps. Conventional relationship extraction is concentrated on studying about biomedical relation extraction (e.g., protein-protein interaction and gene-disease relation) from biomedical terms (e.g., genes, proteins, diseases, or drugs) (Abacha and Zweigenbaum, 2011).

Abacha and Zweigenbaum (2011) is able to detect the specified semantic relationship between each pair of entities through MetaMap (Aronson and Lang, 2010) to identify medical substances whereas a linguistic patterns approach find out the semantic relationship between each pair. Chun *et al*. (2006) could propose a system to extract gene-disease relations from Medline. They used a machine learning-based named entity recognition system to eliminate inaccurate disease and gene names caused by dictionary matching-based term recognition. They examined that by developing the terms recognition performance would also improve the relationship extraction precision.

Also, several other Information extraction systems in biology make use of pattern matching approaches (Huang *et al*., 2004), which sometimes have partial generalization. Moreover, the closer analysis the text, the more patterns is required to take account of the large amount of grammatical variation in texts. The major disadvantage is that some measure of semantic processing away from pattern matching is necessary that is superior to either text strings connected with surface analyses.

The major aim of this study is to develop an efficient information extraction method for bio text corpus or biotext. The proposed dictionary-based text mining approach extracts information from the biotext and then preprocessing is carried out to remove noise and irrelevant text in the biotext to further improve the PPI Named Entity Relationship (NER) identification between the genes and protein gene-protein relationship for specific entity. The preprocessing step consists of the following steps such as stemming, tokenization, stop word removal, Morphological analysis, Word-Sense Disambiguator (WSD). These preprocessing steps are carried out as the biomedical text includes not only English characters but many special terms such as the names of genes, proteins and chemicals. Named Entity Relationship (NER) is carried out by creation of the gene network and gene-gene relationship measurement. This proposed approach gives more types of relationships and increases the number of biomedical entities. These resources are likely to be complete at any given moment, resulting in some synonymy relationships that may captured exactly.

BACKGROUND STUDY

The entity type ontology includes the well-known Genia ontology of physical types was studied (Ohta *et al*., 2002). For corpus users, the ontologies exactly describe which types of entities and relationships are processed and how they are linked with each other. By binding the corpus text to typed entities and relationships, the annotation also offers a mapping from the open field of language statements up to certain limit, controlled vocabulary of types in the ontologies, identifying the words that are employed to state entities and relationships of each type.

In earlier days, subset of the corpus syntactic annotation is employed to compute the performance of the Link Grammar and Connexor Machinese Syntax dependency parsers in the biomedical area was studied (Pyysalo *et al*., 2004). The annotation permitted a detailed error analysis which find a number of areas for future province adaptation of Link Grammar were studied (Aubin *et al*., 2005).

In past decades, development of a gene module is based on the co-expression property of genes (Sharan *et al*., 2003). By incorporating gene expression with protein-protein interaction data expanded the modules to a great extent and in rare cases improves the functional association of modules (Gu *et al*., 2010). The gene expression and protein interaction information can be employed in designing gene modules, even though these approaches could not contain wide knowledge about the presented gene annotation/function. Thus, these approaches would not be estimated to be maximally efficient in building modules with strong association to biological functions (Wang and Zhang, 2007).

In recent times, the idea of semantic similarity, which was developed for analyzing gene ontology terms and used to foresee and verify protein functions and interactions (Cho *et al*., 2008). Wang *et al*. (2007) presented a Gene Ontology (GO) (Ashburner *et al*., 2000) structure based measure to enumerate semantic similarity between individual terms with genes and showed the advantage of using semantic similarity in organizing complex biological terms.

PROPOSED METHODOLOGY

Biotext corpus was originally annotated for disease and treatment mentions (Rosario and Hearst, 2004) and is part of Biotext Project at UC Berkley. The corpus was obtained from MEDLINE 2001 and contains 3655 annotated sentences. In the proposed work, in order to perform the entity relationship process, biotext corpus is taken as input which is stored in local database. Initially the user query is sent to database and related information of the query is extracted from biotext using the dictionary-based text mining, once the text are extracted then preprocessing is done to remove the non-functional characters like stop words, comma, etc, in the extracted information. Once the preprocessing step is completed then set of the entities is also identified in the preprocessed text from biotext corpus and measure

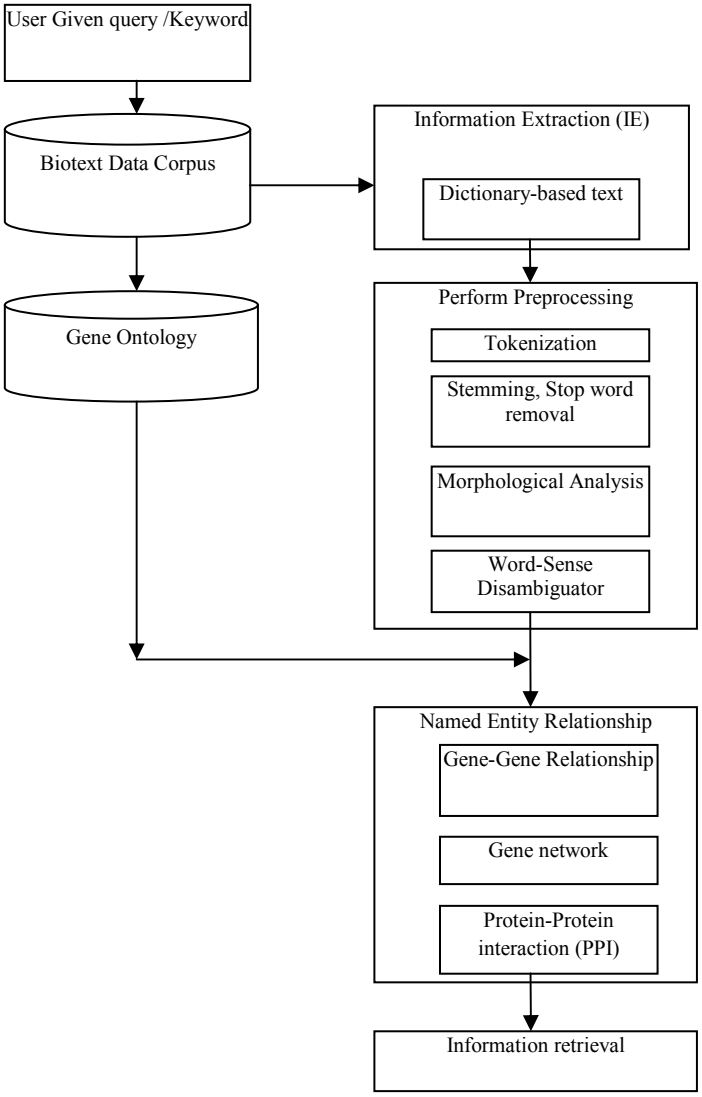

Fig. 1: Proposed architecture representation

the semantic similarity between the terms or entities mentioned from the above step. After the semantic similarity steps the results becomes the entity relationship. The proposed architecture is shown in the Fig. 1.

Information extraction: Bhattacharya *et al*. (2010) mainly focused on dictionary-based text mining and its role in enabling practitioners in recognizing and examining huge text datasets. They described dictionary $D = \mathrm{Dict}(C, X)$ as a set of words, which depicts a semantic protein and protein interaction perception C in a document collection X and as well to build a protein-protein interaction related concept dictionaries for interpreting a set of documents from a particular domain. Here, an online interactive framework is implemented, where the user starts off with a small set of protein or gene words, examines the consequences, chooses and declines protein or gene words from the returned ranking where the process done iteratively until meet the criteria. The user grants positive and negative seed gene or protein at each stage of iteration to the method with interactive regulation.

This procedure often refines the seed sets and the ranking is provided based on the closeness of the user's preference as the iteration maintains. For general purpose gene or protein with English meaning, this framework of building dictionary requires for granting a set of seed gene or protein names for stipulating a concept C and utilizes the WordNet to describe the semantics of seed set definitely. On the other hand, uncertainties may occur while choosing the protein or genes in seed set or some subset of them, the conceptual structure is deficient like WordNet when all proteins and genes are not equal to these perceptions. After that, the dictionaries require to be built for every new Biotext dataset and the existing concept nodes can be utilized for seeding to symbolize the biotext

effectively. Subsequently the ranking revisited by the system is checked to generate the adapted dictionary for the new collection of document. As a result reclaiming dictionaries can considerably make the task of denoting the semantic concept to be simple without semantic structure for dictionary creation for a concept in text dataset.

Algorithm 1: Build dictionary (Gene terms g^s or p^s, Int K, Double t):

- Initialize candidate set C_S to empty set
- For all the gene or protein terms ($G(g^s$ or $p^s)$)
- If the $G(p^s, p') > thershold$ t'
- Add p' to candidate set C_S
- For each gene or protein terms p' in C_S
- Compute similarity between two protein (p^s, p') with p^s
- Reject p' if the similarity below the threshold t
- Sort remaining proteins or genes by the similarity and return the top most information that related to concept it is represented as K

In order to perform the similarity between the protein using the following steps:

Random forest (Breiman, 2001) uses a group of independent decision trees in place of one tree. Θ is represented as the set of possible attributes (or variables on which nodes can be split) and by $h(x, \Theta)$ a tree grown by Θ to classify a vector protein p^s. By these notations a random forest f is defined as:

$$f = \{h(p^s, \Theta_k)\}\ k = 1, .. K \quad (1)$$

where, $\Theta_k \subseteq \Theta$. That is to say, a random forest is a group of trees, where each tree is grown by a subset of all possible attributes. For the k^{th} tree Θ_k is randomly chosen and is independent of the past random vectors $\Theta_1, \ldots, \Theta_k$. To categorize, using each of the trees 'votes' for one of the classes and the most well-liked class is allocated to input text documents related to protein data p^s. For a given forest f, calculate the similarity between two pairs of proteins pairs p^s and p' in the same way. For each of the two pairs initially propagate their values down all trees within f. After that, the terminal node position for each pair in each of the trees is recorded. Let $Z_1 = (Z_{11}, \ldots, Z_{1K})$ be these tree node positions for p^s and similarly define Z_2. Then the similarity between pair p^s and p' is set to: (I is the indicator function.):

$$S(p^s, p') = \frac{1}{N}\sum_{i=1}^{N} I(Z_{1i} == Z_{2i}) \quad (2)$$

Preprocessing for extracted information: Preprocessing step is carried out for extracted information or text.

Tokenization: Generally, in the field of English text, individual English words are obviously employed as tokens. Tokenization can be processed using white spaces as delimiters by using all non-alphanumerical characters as delimiters. Evidently, a simple to kenizer for well-known English text won't work healthy in biomedical text. If every non-alphanumerical character within a named entity is used as delimiters to split the name into numerous tokens, the closeness of these tokens is misplaced in the bag-of-word representation, which may cause in a loss of the semantic meaning of the tokens and cause mismatches. To solve this problem a set of tokenization heuristics is presented that are indiscriminate from previous work on biomedical information retrieval and conduct a systematic evaluation of these heuristics. Specifically, define three sets of break points, three break point normalization methods and a Greek alphabet normalization method.

Heuristic rules to remove non-functional characters:

- Replace the following characters with spaces: ! " # $ % & * < = > ? @ \ / ~
- Remove the following characters if they are followed by a space: . : ; ,
- Remove the following pairs of brackets if the open bracket is preceded by a space and the close bracket is followed by a space: ( ) [ ]
- Remove the single quotation mark if it is preceded by a space or if it is followed by a space: 0
- Remove 's And 'T If they are followed by a space
- Remove slash/if it is followed by a space

Stemmer: Removes the nuance of words for indexing. Semantically associated words should map to the same stem, base or root form and this should reimburse for data sparseness.

Stop-word removal: Stop-words are the words that emerge very often and are of little significant in the discrimination of documents in general. They can be specific to a dataset. In this step, standard stop words such as a, the, for, etc. and words with high frequency (which are not in the standard list) are removed.

Morphological analysis: The preserved words are then examined in order to combine different words having the same root to a single one. For example, the words laugh, laughed, laughter and laughing should be combined to form a single word, laugh. This is logical because all such words convey the same (loosely) meaning.

Word-sense disambiguator: resolves the sense of emotional words (i.e., nouns, adjectives and verbs) based on their context (Sebastiani, 2002). It employs a semantic similarity measure to score the senses of an

affective word with the context words using the Word Net ontology (Seco *et al*., 2004). In addition, the module retrieves the set of synonyms for the resulting sense in order to expand the feature space (Manning *et al*., 2008).

Named entity relationship: The second subtask of information extraction in biomedical domain is relevant to relation extraction, which aims to detect binary relationships among named entities. Examples include gene disease relationships, protein-protein interactions and medical problem treatment relationships. General entities of attention consist of gene and protein names, medical problems and treatments, drug names and their dosages and other semantically definite data separable within the biomedical domain. The integration of semantic similarity based on GO annotation with gene expression and protein-protein interaction data can seriously improve the functional significance of inferred gene modules. Gene Ontology Annotation Similarity (Gene Ontology Distance) relating to all other proteins in the data. Each protein pair was computed by calculating the Gene Ontology Annotation Similarity, taking the gene or protein weights into consideration, can employ any one of the standard similarity measures to estimate $\mathrm{Sim}(p_i, p_j)$. At this point, only present the cosine similarity measure as it is most frequently used in information retrieval:

$$GO(S) = S(p_i, p_j) = \frac{\sum_{i=1}^{k} ce_i(p_i) * ce_i(p_j)}{\sqrt{\sum_{i=1}^{l} c\omega_i^2(p_i)}\sqrt{\sum_{i=1}^{l} c\omega_i^2(p_j)}} \quad (3)$$

where, $ce_i(p_i)$and $ce_i(p_j)$ are the weights of the i^{th} common entity in the expanded query from user given for two different proteins. Gene Ontology Annotation Similarity $S(p_i, p_j)$ between two proteins was less than 1.0, they were well thought-out to be interacting, therefore forming an interaction network. The GO annotations were identified for each protein from Uni Prot. Protein-protein interaction algorithm in which outputs the interaction scores that are annotated on the network as the interaction strength (Palakal *et al*., 2003).

This algorithm consists of following steps:

- Identify the protein pair P(i, j) and its associated structures given in the protein data bank (PDB)
- Find the feasible interacting residues of each PDB structure in the given pair by means of the physico-chemical properties of its residues, hydrophobicity, accessibility and residue propensity
- Calculate the distance between the C-alpha coordinates of the likely interacting residues of the given pair
- Identify the protein pair as interacting or non-interacting based on the given distance threshold
- Evaluate the interaction of the gene protein pair (i, j) satisfies the distance threshold, then the pair is considered interacting:

$$\text{Protein interaction score}_{i,j} = \frac{\text{\#of interacting residues}}{\text{probable number of interacting residues}} \quad (4)$$

$$\text{interaction between proteins score}_{i,j} = \frac{\text{\#of interacting PDB strucutures}}{\text{Total number of PDB strucutures}} \quad (5)$$

Network topology is an important parameter that defines the biological function and performance of the network (Barabasi and Bonabeau, 2003). Network properties such as degree, centrality and clustering coefficients play a significant role in finding the network's underlying biological significance (Kuchaiev *et al*., 2010). For the topological analysis, degree, clustering coefficient and betweenness (centrality) are considered. Degree is the number of edges connected to node i. The clustering coefficient of node i is defined as $C_i = 2n/k_i(k_i - 1)$, where n is the number of connected pairs between all the neighbors of node i and k_i is the number of neighbors of n. Betweenness for node i is the number of times the node is a member of the set of shortest paths that connects all pairs of nodes in the network and it is given as $C_B(n_i) = \sum_{j<k} g_{jk}(n_i)/g_{jk}$, where g_{jk} is the number of links connecting nodes j and k and $g_{jk}(n_i)$ is number of links passing through i. These network properties were calculated by I graph package of statistical tool R. Gene Ontology Annotation Similarity Score by considering the average edge weight of each protein and its interacting neighbors:

$$\text{Gene ontology annotation similarity score}_I = \frac{\sum_{i=1}^{N}\sum_{j=1}^{K}(GO_S)_{ij}}{K} \quad (6)$$

where,

N = The total number of nodes in the network
i = The node in consideration
K = The number of immediate neighbors of node
i, j = The interacting neighbors

The calculation of the Gene Ontology Annotation Similarity Score is illustrated in Additional file 1. The Protein Interaction Propensity Score for a given node was calculated based on the statement that proteins frequently interact with the domains of their own family and therefore it is calculated as:

$$\text{Protein interaction prospensity score}_i = \frac{\sum_{i=1}^{N}\sum_{j=1}^{K}(\text{protein interaction score}_{ij})/k}{\sum_{i=1}^{N}\sum_{j=1}^{K}(\text{protein interaction score}_{ij})/N} \quad (7)$$

where,

N = The total number of nodes in the network
i = The node in consideration
K = The number of immediate neighbors of node i

Protein-Protein Interaction (PPI): Various gene-gene pairwise similarity values, containing information obtained from gene expression, protein-protein interactions and GO annotations, in the construction of modules using affinity propagation clustering.

EXPERIMENTAL RESULTS AND DISCUSSION

Biotext corpus was initially annotated for disease and treatment (Rosario and Hearst, 2004) and it is the part of Biotext Project at UC Berkley. The corpus was attained from MEDLINE 2001 which contains 3655 annotated sentences. These sentences were then searched for the protein pairs that contain the names of at least two proteins that are known to interact. This selection process outcome in a corpus with a much superior proportion of related sentences that is sentences that state actual relationships. The sentences are maintained as they appear in the article abstracts, including spelling errors, grammatical mistakes. Biologically, it is of particular interest to identify the entity relationship and entity that contribute the most to classify protein pairs. Such an analysis can help uncover relationships between different data sources which are not directly apparent. In addition, it can help identify what information should be generated for determining interaction in other species. To compare the efficiency of constructing gene network and gene relationship, precision vs. recall curve to perform the comparisons completely.

Precision: Among the pairs identified as interacting by the classifier, what is the fraction (or percentage) that is truly interacting:

$$\text{Precision} = \frac{C}{D} \tag{8}$$

Let D are the number pairs identified as interacting by the classifier, C be the number of pairs correctly identified as interacting.

Recall: For the known interaction pairs, what is the percentage that is identified:

$$\text{Recall} = \frac{C}{T} \tag{9}$$

T be the number of pairs labeled as interacting. In other words, precision is the accuracy of our predictor whereas recall is the coverage of the classifier.

Figure 2 shows a comparison between the proposed Dictionary-Based Text and Gene Ontology (DBTGO) method and it is compared with the existing Weighted kNN (k Nearest Neighbor) that uses Euclidean distance (Qi *et al*., 2005), Weighted kNN (k Nearest Neighbor) with Random Forest (RF) (Qi *et al*., 2005) similarity methods for precision and recall curves. If the number of recall values are high the proposed DBTGO based system maintains higher result of the precision value

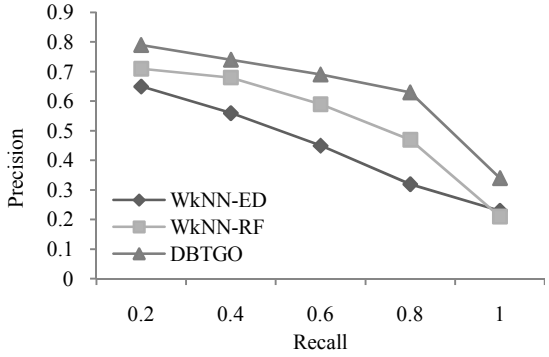

Fig. 2: Precision vs. recall curves

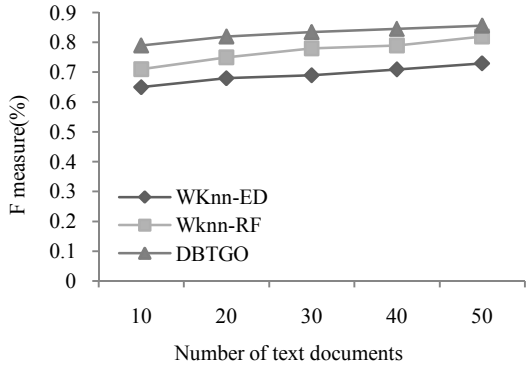

Fig. 3: F-measure vs methods

when compare to the existing methods since it creates gene ontology for each one of the entity in the biotext.

F-measure: For a given impact level, the number of modules enriched with at least one annotation term and the number of annotation terms. The sensitivity and specificity were calculated for all the methods and summarized into a measure of basic enrichment, an F-measure defined as:

$$F = 2 \times \text{precision} \times \text{recall}/(\text{precision} + \text{recall}) \tag{10}$$

Figure 3 shows a comparison between the proposed DBTGO method with existing method in terms of F-measure. It shows that the F measure results of the proposed DBTGO have higher F measure result than the existing WkNN-ED and WkNN-RF algorithm because the proposed work creates a gene ontology network and maintain the gene network to find the named entity relationship for each one of the entities in the network.

Normalized Discounted Cumulative Gain (NDCG): NDCG measures the performance of a proposed system based on the graded relevance of the proposed units. It may vary from 0.0 to 1.0, with 1.0 representing the ideal ranking of the entities. This metric is generally used in information retrieval and it is computed as:

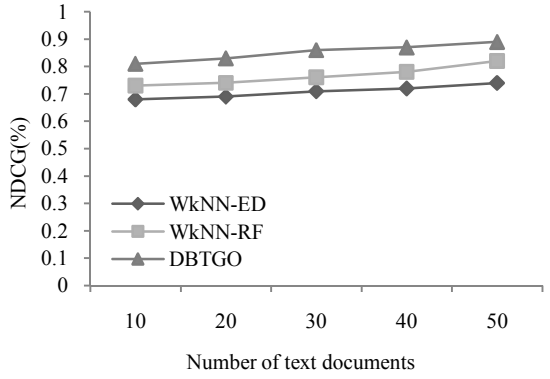

Fig. 4: NDCG vs. methods

$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i - 1}}{\log_2(i+1)} \quad (11)$$

$IDCG_k$ is the maximum possible (ideal) DCG for a given set of queries, documents and relevance's:

$$nDCG_k = \frac{DCG_k}{IDCG_k} \quad (12)$$

Figure 4 shows a comparison between the proposed and the existing method in terms of NDCG parameter. It shows that the NDCG results of the proposed DBTGO have higher NDCG result than the existing WKNN-ED and WKNN-RF algorithm because proposed work return more relevant named entity relationship by the creates a gene ontology network, it retrieves exact Protein-Protein Interaction(PPI) result than the existing WkNN-ED and WkNN-RF method.

CONCLUSION

In this research a new method is proposed for predicting protein-protein interactions by integrating diverse high-throughput biological datasets. The proposed work consists of the following steps: they are, information extraction from biotext corpus by using the dictionary based text methods and then performs preprocessing process using various methods such as stemming, stop word removal, tokenizer, Morphological Analysis, Word-Sense Disambiguator (WSD). Then a similarity measure is computed between protein pairs based on the gene ontology similarity measurement. The proposed DBTGO algorithm outperforms previous methods like WkNN-ED and WkNN-RF and can also derive meaningful biological results for known protein and gene relationship. It exactly identifies the PPI, Gene-Gene relationship for named entity relationship results and results are measured in terms of precision, Recall, F-Measure and NDCG parameters. Interestingly, many of the features determined to be important in this proposed method are through direct measurements (protein-protein interaction and gene-protein interaction). This becomes an open issue in this proposed work, so, in the future work, this study aims to find the important features in the gene to determine interacting pairs.

REFERENCES

Abacha, A.B. and P. Zweigenbaum, 2011. Automatic extraction of semantic relations between medical entities: A rule based approach. J. Biomed. Semant., 2(Suppl. 5): S4.

Aebersold, R. and M. Mann, 2003. Mass spectrometry-based proteomics. Nature, 422(6928): 198-207.

Ananiadou, S., S. Pyysalo, J. Tsujii and D.B. Kell, 2010. Event extraction for systems biology by text mining the literature. Trends Biotechnol., 28: 381-390.

Aronson, A.R. and F.M. Lang, 2010. An overview of MetaMap: Historical perspective and recent advances. J. Am. Med. Inform. Assn., 17: 229-236.

Ashburner, M., C.A. Ball, J.A,. Blake D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight and J.T. Eppig, 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet., 25: 25-29.

Aubin, S., A. Nazarenko and C. Nédellec, 2005. Adapting a general parser to a sublanguage. In: Angelova, G., K. Bontcheva, R. Mitkov, N. Nicolov and N. Nikolov (Eds.), Proceeding of the International Conference on Recent Advances in Natural Language Processing (RANLP, 05). Borovets, Incoma, Bulgaria, pp: 89-93.

Barabasi, A.L. and E. Bonabeau, 2003. Scale-free networks. Sci. Am., 288(5): 60-69.

Bhattacharya, I., S. Godbole, A. Gupta and A. Verma, 2010. Building re-usable dictionary repositories for real-world text mining. Proceeding of the 9th ACM international conference on Information and knowledge management (CIKM'10). Toronto, Ontario, Canada, October 26-30.

Breiman, L., 2001. Random forests. Mach. Learn., 45: 5-32.

Cho, Y.R., L. Shi, M. Ramanathan and A. Zhang, 2008. A probabilistic framework to predict protein function from interaction data integrated with semantic knowledge. BMC Bioinformatics, 9: 382.

Chun, H.W., Y. Tsuruoka, J.D. Kim, R. Shiba, N. Nagata, T. Hishiki and J. Tsujii, 2006. Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. Proceeding of the Pacific Symposium on Biocomputing, pp: 4-15.

Gu, J., Y. Chen, S. Li and Y. Li, 2010. Identification of responsive gene modules by network-based gene clustering and extending: Application to inflammation and angiogenesis. BMC Syst. Biol., 4: 47.

Huang, M., X. Zhu, D.G. Payan, K. Qu and M. Li, 2004. Discovering patterns to extract protein-protein interactions from full biomedical texts. Bioinformatics, 20: 3604-3612.

Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori and Y. Sakaki, 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. P. Natl. Acad. Sci. USA, 98(8): 4569-4574.

Kuchaiev, O., T. Milenkovic, V. Memisevic, W. Hayes and N. Przulj, 2010. Topological network alignment uncovers biological function and phylogeny. J. Roy. Soc. Interface, 7(50): 1341-1354.

Manning, C.D., P. Raghavan and H. Schütze, 2008. Introduction to Information Retrieval. Cambridge University Press, Cambridge, MA.

Ohta, T., Y. Tateisi, H. Mima and J. Tsujii, 2002. GENIA corpus: An annotated research abstract corpus in molecular biology domain. Proceeding of the Human Language Technology Conference (HLT, 2002). San Diego, California, pp: 73-77.

Palakal, M., M. Stephens, S. Mukhopadhyay, R. Raje and S. Rhodes, 2003. Identification of biological relationships from text documents using efficient computational methods. J. Bioinform. Comput. Biol., 1(2): 307-342.

Pyysalo, S., F. Ginter, T. Pahikkala, J. Boberg, J. Järvinen, T. Salakoski and J. Koivula, 2004. Analysis of link grammar on biomedical dependency corpus targeted at protein-protein interactions. Proceeding of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA). Geneva, Switzerland, pp: 15-21.

Qi, Y., J. Klein-Seetharaman and Z. Bar-Joseph, 2005. Random forest similarity for protein: Protein interaction prediction from multiple sources. Proceeding of the Pacific Symposium on Biocomputing, 10: 531-542.

Rosario, B. and M. Hearst, 2004. Classifying semantic relations in bioscience texts. Proceeding of the 42nd Annual Meeting of Association of Computing Linguistics.

Schulze, A. and J. Downward, 2001. Navigating gene expression using microarrays: A technology review. Nat. Cell Biol., 3(8): E190-E195.

Sebastiani, F., 2002. Machine learning in automated text categorization. ACM Comput. Surv., 34: 1-47.

Seco, N., T. Veale and J. Hayes, 2004. An intrinsic information content metric for semantic similarity in WordNet. Proceeding of the European Conference on Artificial Intelligence (ECAI'04), pp: 1089-1090.

Sharan, R., A. Maron-Katz and R. Shamir, 2003. Click and expander: A system for clustering and visualizing gene expression data. Bioinformatics, 19: 1787-1799.

Uetz, P., L. Giot and G. Cagney, 2000. A comprehensive analysis of protein' protein interactions in *Saccharomyces cerevisiae*. Nature, 403: 623-627.

Wang, J.Z., Z. Du, R. Payattakool, P.S. Yu and C.F. Chen, 2007. A new method to measure the semantic similarity of GO terms. Bioinformatics, 23: 1274-1281.

Wang, Z. and J. Zhang, 2007. In search of the biological significance of modular structures in protein networks. PLoS Comput. Biol., 3: e107.

Winnenburg, R., T. Wachter, C. Plake, A. Doms and M. Schroeder, 2008. Facts from text: Casn text mining help to scale-up high-quality manual curation of gene products with ontologies? Brief. Bioinform., 9: 466-478.

Zweigenbaum, P., D. Demner-Fushman, H. Yu and K.B. Cohen, 2007. Frontiers of biomedical text mining: Current progress. Brief. Bioinform., 8: 358-375.