Validating Text Mining Results on Protein-Protein Interactions Using Gene Expression Profiles

Deyu Zhou, Yulan He and Chee Keong Kwoh
School of Computer Engineering, Nanyang Technological University
Nanyang Avenue, Singapore 639798

Abstract

Protein-protein interactions referring to the associations of protein molecules are crucial for many biological functions. Since most knowledge about them still hides in biological publications, there is an increasing focus on mining information from the vast amount of biological literature such as MedLine. Many approaches, such as pattern matching, shallow parsing and deep parsing, have been proposed to automatically extract protein-protein interaction information from text sources, with however limited success. Moreover, to the best of our knowledge, none of the existing approaches have performed automatic validation on the mining results. In this paper, we describe a novel framework in which text mining results are automatically validated using the knowledge mined from gene expression profiles. A probability model is proposed to score the confidence of protein-protein interactions based on both text mining results and gene expression profiles. Experimental results are presented to show the feasibility of this framework.

1 Introduction

How proteins interact with each other gives biologists a deep insight into the mechanism of living cell and provides targets for effective drug designs. Until now, vast knowledge about protein-protein interactions are still locked in the biological publications. As a result, automatically mining protein-protein interactions from literature is crucial to meet the demand of the researchers.

Existing approaches can be broadly categorized into two types, based on simple pattern matching, or employing parsing methods. Approaches using pattern matching [1, 2] rely on a set of predefined or automatically generated patterns to extract protein-protein interactions. Parsing based methods employ either shallow or deep parsing. Unlike word-based pattern matchers, shallow parsers [3, 4] break sentences into non-overlapping phases. They extract local dependencies among phases without reconstructing the structure of an entire sentence. Systems based on deep parsing [5, 6] deal with the structure of an entire sentence and therefore are potentially more accurate. Table 1 shows the best performance reported so far in each category. These are only indicative figures since no benchmarking datasets are available to compare the performance fairly.

Category	Performance		Reference
	Recall(%)	Precision(%)	
Rule-based	86	94	[1]
Shallow parsing	62	89	[7]
Deep parsing	48	80	[8]

Table 1: Performance on mining protein-protein interactions from literature.

More recently, there is a trend that mining results from literature can be integrated with knowledge from experiments and genome analysis to improve the extraction accuracy [9]. One such example is to investigate the relationships between protein-protein interactions and gene expressions since a protein is the product of a gene. From gene expression profiles, co-expressed genes, which are groups of genes that demonstrate coherent patterns on samples, can be found. Grigoriev [10] observed that proteins encoded by co-expressed genes interact with each other more frequently than with random pairs by analyzing physical interactions in yeast. Jansen *et al.* [11] found that apart from a few big known protein complexes that have clearly defined interactions among their subunits, the relationship between the two is weak in yeast. In [12], in order to investigate the global relationship of protein interactions with gene expressions, four diverse species were studied including human, mouse, yeast, and *Escherichia coli*. The results show that in *E. coli* the gene expression profiles of interacting pairs are highly correlated in comparison to random pairs, while in the other three species only slightly stronger relations are revealed than those of random pairs.

Based on the above findings, we may conclude that there exist relations between protein-protein interactions and gene expression profiles. It is therefore natural to investigate the feasibility of validating text mining results based on the gene expression profiles. In this paper, a framework of validating mining results from gene expression data has been proposed. Since the strength of the relationships between protein-protein interactions and gene expression profiles is different across different species, simply combining the knowledge discovered from both

sides may result in poor performance. A probability model is therefore proposed to account for various correlation levels between protein-protein interactions and gene expression profiles.

The rest of the paper is organized as follows. In section 2, the overall system framework is presented. Section 3 then describes the main methods employed in the system in more details. Experimental results are presented and discussed in section 4. Finally, section 5 concludes the paper.

2 System Overview

The process of validating text mining results based on gene expression profiles is conducted in three stages: 1) mining protein-protein interactions from literature, 2) clustering co-expressed genes based on their expression levels, 3) making inference based on the above results. Thus, the system comprises of the three main components which are illustrated in Figure 1. The functions of each component are described as follows.

1. **Mining protein-protein interactions from literature.**
   This component can be further divided into three sub-components as illustrated in Figure 1.
   - *Preprocessing – identification of protein names, other biological terms and interaction keywords.*
     In our system, protein names are identified based on a dictionary of manually constructed biological terms. In addition, a category/keyword dictionary for identifying terms describing interactions has also been built based on [13]. All identified biological terms and interaction keywords are then replaced with their respective category labels.
   - *Semantic Parsing – parsing sentences using the Hidden Vector State model.*
     A sentence which contains at least two protein names identified by the preprocessing step is then parsed with the Hidden Vector State (HVS) model which were trained using a lightly annotated training corpus. Details about the HVS model will be discussed in section 3.1.
   - *Extraction of protein-protein interactions.*
     Given the HVS parsing results, the protein-protein interactions can be easily extracted using some predefined rules.
2. **Clustering genes based on their expression profiles.**
   This component can be further divided into two sub-components.
   - *Preprocessing – identification of species having strong relations between gene expressions and protein-protein interactions.*
     To reduce the quantity of the data to be processed, the correlation values between gene expressions and protein-protein interactions are first calculated using some statistics measures based on [12] for each species, such as human, mouse, yeast etc. Only the species exhibiting strong correlations will be considered.
   - *Clustering – using the ant-based clustering algorithm to group genes in the same species.*
     The ant-based clustering algorithm has been applied successfully in document clustering [14]. We intend to further investigate this algorithm to handle the gene expression data.
3. **Making inference based on the above results.**
   Considering the text mining results as assertions, the confidence level of each assertion is inferred based on the gene clustering results. The assertions with their confidence levels below a predefined threshold will then be rejected.

3 Methodology

The main methods employed by the framework presented in Section 2 are discussed in detail in this section. The Hidden Vector State (HVS) model which is used for mining protein-protein interactions from literature [6] is described followed by the ant-based clustering algorithm for gene expression data clustering. At last, the probability model for making reference based on mining results and gene expression information is described.

3.1 Hidden Vector State Model for Text Mining

Instead of manually defining semantic rules or patterns to extract protein-protein interactions from literature, we are more interested in investigating statistical approaches which can perform automatic extraction without hand-crafted rules. Here, we propose a Hidden Vector State (HVS) model which is a discrete Hidden Markov Model (HMM) in which each HMM state represents the state of a push-down automaton with a finite stack size. The state transitions may be factored into a stack shift by n positions followed by a push of one or more new preterminal semantic concepts relating to the next input word. Such stack operations are constrained in order to reduce the state space to a manageable size. Natural constraints to introduce are limiting the maximum stack depth and only allowing one new preterminal semantic concept to be pushed onto the stack for each new input word. Such constraints effectively limit the class of supported languages to be right branching.

Given a series of stack shift operations N, concept vector sequence $\mathbf{C}$, and word sequence W, the joint probability $P(N, \mathbf{C}, W)$ can be decomposed as follows

$$P(N, \mathbf{C}, W) = \prod_{t=1}^{T} P(n_t | W_1^{t-1}, \mathbf{C}_1^{t-1}) \cdot P(c_t[1] | W_1^{t-1}, \mathbf{C}_1^{t-1}, n_t) \cdot P(w_t | W_1^{t-1}, \mathbf{C}_1^{t}) \quad (1)$$

where:

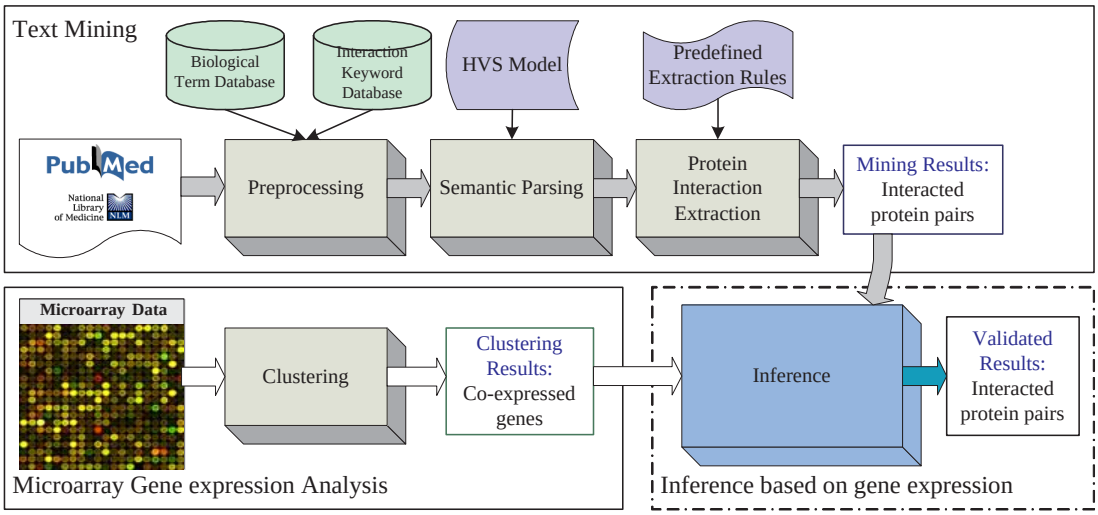

Figure 1: System architecture.

- $\mathbf{C}_1^t$ denotes a sequence of vector states $\mathbf{c}_1..\mathbf{c}_t$. $\mathbf{c}_t$ at word position t is a vector of D_t semantic concept labels (tags), i.e. $\mathbf{c}_t = [c_t[1], c_t[2], .., c_t[D_t]]$ where $c_t[1]$ is the preterminal concept and $c_t[D_t]$ is the root concept ;

- $W_1^{t-1}\mathbf{C}_1^{t-1}$ denotes the previous word-parse up to position $t-1$;

- n_t is the vector stack shift operation and takes values in the range of $0, .., D_{t-1}$ where D_{t-1} is the stack size at word position $t-1$;

- $c_t[1] = c_{w_t}$ is the new preterminal semantic tag assigned to word w_t at word position t.

The details of how this is done are given in [15]. The result is a model which is complex enough to capture hierarchical structure.

Unlike other fully-recursive statistical parsers which need fully-annotated treebank data for training, the HVS model explores the embedded sentence structures using only lightly annotated corpus. To train the HVS model, an abstract annotation needs to be provided for each sentence. For example, for the sentence,
CUL-1 was found to interact with SKR-1, SKR-2, SKR-3, SKR-7, SKR-8 and SKR-10 in yeast two-hybrid system.
The Annotation is:
PROTEIN_NAME(ACTIVATE(PROTEIN_NAME)).

The trained HVS model can then be used to parse the sentences from the medical literature and protein-protein interactions can be extracted based on some simple predefined rules [16].

3.2 Ant-Based Clustering Method

Cluster analysis is concerned with multivariate techniques that can be used to create groups amongst the observations, where there is no *a priori* information regarding the underlying group structure. Clustering of the genes on the basis of the tissues can be used to search for groups of gene that might be regulated together. Available methods of cluster analysis can be categorized broadly as being hierarchical such as Agglomerative Hierarchical Clustering (AHC) or non-hierarchical such as k-means clustering. A major limitation of hierarchical methods is their inability to determine the number of cluster. The limitation of k-means method is its high computational complexity.

The Ant Colony Optimization (ACO) algorithm belongs to the natural class of problem solving techniques which is initially inspired by the efficiency of real ants as they find their fastest path back to their nest when sourcing for food. An ant is able to find this path back due to the presence of pheromone deposited along the trail by either itself or other ants. An open loop feedback exists in this process as the chances of an ant taking a path increases with the amount of pheromone built up by other ants.

Early approaches in applying ACO to clustering are to first partition the search area into grids. A population of ant-like agents then move around this 2D grid and carry or drop objects based on certain probabilities so as to categorize the objects. However, this may result in too many clusters as there might be objects left alone in the 2D grid and objects still carried by the ants when the algorithm stops. Therefore, Some other algorithms such as k-means are normally combined with ACO to minimize categorization errors. More recently, variants of ant-based clustering have been proposed, such as using inhomogeneous population of ants which allow to skip several grid cells in one step, representing ants as data objects and allowing them to enter either the active state or the sleeping state on a 2D grid. Existing approaches are all based on the same scenario that ants

move around in a 2D grid and carry or drop objects to perform categorization.

We have proposed an ant-based clustering algorithm for document clustering based on the travelling salesperson (TSP) scenario [14]. The advantages of our ant-based clustering approach are : 1) It does no rely on a 2D grid structure. 2) It can generate optimal number of clusters without incorporating any other algorithms such as k-means or AHC. 3) When compared with both the classical document clustering algorithms such as K-means and AHC and the Artificial Immune Network (aiNet) based method, it shows improved performance when tested on the subsets of 20 Newsgroup data [17]. We intend to investigate the ant-based clustering algorithm for gene expression data analysis.

3.3 Probability Model for Validation

Given two proteins P_1 and P_2, two genes G_1 and G_2 encode the two proteins respectively. Assuming that $A = \text{Interact}(P_1, P_2)$ refers to the event that protein P_1 interacts with protein P_2, $B = \text{Coexpress}(G_1, G_2)$ refers to the event that G_1 and G_2 belong to one cluster according to some clustering algorithm, the probability of the event A can be decomposed as follows based on the Bayesian theorem:

$$p_g = \Pr(A) = \Pr(A|B)\Pr(B) + \Pr(A|\overline{B})\Pr(\overline{B}) \quad (2)$$

Since $\Pr(B)$ and $\Pr(\overline{B})$ are easy to calculate, it is crucial to compute the conditional probability $\Pr(A|B)$ and $\Pr(A|\overline{B})$.

Assuming that relationship between gene expressions and protein interactions is strong, we build the probability model based on the logistic regression model having this form:

$$\log \frac{\Pr(A|B)}{\Pr(\overline{A}|B)} = \beta_0 + \beta_1 \text{Dist}(G_1, G_2) \quad (3)$$

$$\Pr(A|B) = \frac{\exp\big(\beta_0 + \beta_1 \text{Dist}(G_1, G_2)\big)}{1 + \exp\big(\beta_0 + \beta_1 \text{Dist}(G_1, G_2)\big)} \quad (4)$$

where $\text{Dist}(G_1, G_2)$ denotes the distance between the profiles of two genes, and β_0, β_1 are the coefficients of the logistic regression model. Given D_{Euclid} which denotes the Euclidean distance between two gene expression profiles when considering the profile as a vector and Radius_c which denotes the radius of a cluster based on the Euclidean distance, $\text{Dist}(G_1, G_2)$ is defined as follow:

$$\text{Dist}(G_1, G_2) = \begin{cases} 0, & \text{if } G_1, G_2 \text{ are not in the same cluster,} \\ \frac{\text{D}_{Euclid}(G_1,G_2)}{\text{Radius}_c}, & \text{if } G_1, G_2 \text{ are in the same cluster } C \end{cases}$$

Given a species, we can easily estimate the parameter β_0 and β_1 based on the training data consisting of the gene expression profiles and the corresponding proteins which are known to interact. Note that we don't consider the scenario that G_1 and G_2 are not of the same species.

As investigated in [12], the correlation of gene expressions of interacting pairs is different in different species. Directly using the above method will result in poor performance. Here we use the Pearson correlation (PC) coefficient as the measure of relationships between gene expressions and protein interactions for individual species. For each species, we compute the PC coefficient between the expression profiles of the genes whose corresponding proteins are known to interact. The PC coefficient measures the relative shape of the relationship rather than absolute levels and it captures both positive and negative correlations. The detailed process of making inference is described in Figure 2.

Input: protein P_1 interacts with protein P_2 (text mining result)
Ouput: C, reject the assertion when $C = true$; otherwise keep the assertion.
Algorithm:
1. Find the two genes G_1 and G_2 which encodes the proteins P_1 and P_2 respectively.
2. If G_1 and G_2 are not in the same species, $C = false$, return C.
3. Find the species of the two genes and check its Pearson correlation (PC) coefficient. If PC is below a predefined threshold, $C = false$, return C.
4. Calculate the probability of P_1 interactions with P_2, p_g, based on Equation 2.
5. If p_g is less than a predefined threshold, $C = true$, else $C = false$, return C.

Figure 2: Procedure of making inference using gene expression profiles.

From the procedure shown in Figure 2, it can be seen that the assertion that P_1 interacts with P_2 will only be rejected at the condition that G_1 and G_2 are in the same species, the relation between gene expression and protein interaction is strong for this species, and the probability of assertion p_g is small. The idea behind this is that we assume the text mining results have strong confidence and we would need much more stronger belief to invalidate them. The advantage of this method is that the recall value of the protein-protein interaction extraction results will not decrease greatly while the precision value of the extraction results will increase.

4 Experimental Results

At the time of writing this paper, only the text mining component has been implemented and its experimental results are discussed here. The gene expression data clustering component and the inference component are still under development. We however present a case study on validating the mining results using gene expression profiles to illustrate the feasibility of our framework.

4.1 Text Mining Results

Experiments have been conducted on the corpus obtained from [2]. The initial corpus consists of 1203 sentences. The protein interaction information for each sentence is also provided. All sentences were examined manually to ensure the correctness of the protein interactions. After manually cleaning up the sentences which do not provide protein interaction information, 800 sentences were kept.

The results reported here are based on the values of TP, FN, and FP. TP is the number of correctly extracted interactions. (TP+FN) is the number of all interactions in the test set and (TP+FP) is the number of all extracted interactions. F-score is computed using the formula below:

$$\text{F-score} = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} \quad (5)$$

where Recall is defined as $TP/(TP + FN)$ and Precision is defined as $TP/(TP + FP)$.

Table 2 lists the results generated by the HVS model.

Experiment	Recall (%)	Precision (%)	F-Score (%)
1	61.7	71.8	66.4
2	52.6	91.0	66.7
3	60.2	72.7	65.8
overall	58.3	76.8	66.3

Table 2: Text Mining Results using the HVS model.

4.2 Validating Extracted Results Using Gene Expression Profiles

A case study on validating the extracted text mining results using gene expression profiles is shown in Figure 3 to illustrate the feasibility of our framework. Firstly, a pair of interacting proteins, *TonB* and *FhuA*, is extracted from literature using the HVS model, as shown in Figure 3. This protein interaction information is in fact false positive (FP) by comparing with the reference results manually.

Secondly, we found that both proteins are of the same species, *Escherichia coli*. The correlation of protein interactions with gene expression profiles in *Escherichia coli* is calculated and is considered strong since the correlation value exceeds the predefined threshold. Here, the threshold is set to 0.8 which ensures a strong relationship between protein interactions and co-expressed genes. The corresponding gene expression profiles can be obtained from the Stanford MicroArray database [18]. Following the process described in Figure 2, the confidence value of the interaction between the two proteins based on the gene expression profiles can then be calculated using Equation 2. Since the computed value is below the predefined threshold 0.2, we may conclude that there is no interaction between these two proteins.

It can be observed from the above example that the validating component can indeed decrease the FP value of the text mining results which in turn improves the overall performance.

5 Conclusion and Future work

In this paper, we have presented a novel framework to validate text mining results based on information from gene expression profiles. It consists of three major stages: text mining using the HVS model, gene expression data clustering using the ant-based clustering algorithm, and finally perform validation on the extracted protein-protein interactions based on a probability model. Preliminary experimental results and a case study have been presented to illustrate its feasibility. In future work we will continue on the development of the gene expression data clustering component and the inference component and conduct a large scale of experiments to evaluate the system performance.

References

[1] Toshihide Ono, Haretsugu Hishigaki, Akira Tanigam, and Toshihisa Takagi. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, 17(2):155–161, 2001.

[2] Minlie Huang, Xiaoyan Zhu, and Yu Hao. Discovering patterns to extract protein-protein interactions from full text. *Bioinformatics*, 20(18):3604–3612, 2004.

[3] Craven Mark and Kumlien Johan. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*, pages 77–86, Heidelberg, Germany, 1999.

[4] J. Pustejovsky, J. Castano, J. Zhang, M. Kotecki, and B. Cochran. Robust Relational Parsing Over Biomedical Literature: Extracting Inhibit Relations. In *Proceedings of the Pacific Symposium on Biocomputing.*, pages 362–373, Hawaii, U.S.A, 2002.

[5] A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii. Event extraction from biomedical papers using a full parser. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 6, pages 408–419, Hawaii, U.S.A, 2001.

[6] Deyu Zhou, Yulan He, and Chee Keong Kwoh. Extracting Protein-Protein Interactions from the Literature using the Hidden Vector State Model. In *International Workshop on Bioinformatics Research and Applications*, Reading, UK, 2006.

[7] Gondy Leroy, Hsinchun Chen, and Jesse D. Martinez. A Shallow Parser Based on Closed-Class Words to Capture Relations in Biomedical Text. *Journal of Biomedical Informatics*, 36(3):145–158, 2003.

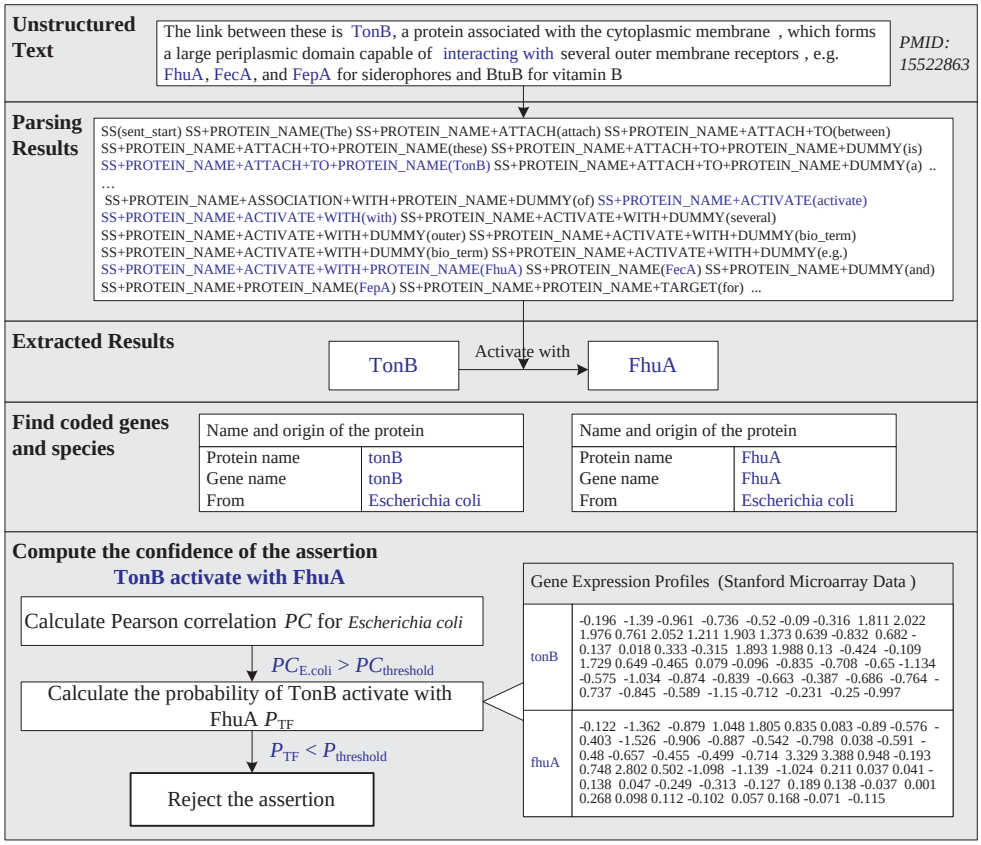

Figure 3: An example of validating extracted results using gene expression profiles.

[8] J. Park, H. Kim, and J. Kim. Bidirectional incremental parsing for automatic pathway identification with combinatory categorical grammar. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 6, pages 396–407, Hawaii, U.S.A, 2001.

[9] M. Scherf, A. Epple, and T. Werner. The next generation of literature analysis: Integration of genomic analysis into text mining. *Briefings in Bioinformatics*, 6(3):287–297, 2005.

[10] Andrei Grigorieva. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. *Nucleic Acids Research*, 29(17).

[11] R. Jansen, D. Greenbaum, and M. Gerstein. Relating whole-genome expression data with protein-protein interactions, 2002.

[12] Nitin Bhardwaj and Hui Lu. Correlation between gene expression profiles and proteincprotein interactions within and across genomes. *Bioinformatics*, 21(11):2730–2738, 2005.

[13] Joshua M. Temkin and Mark R. Gilder. Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics*, 19(16):2046–2053, 2003.

[14] Yulan He, Siu Cheung Hui, and Yongxiang Sim. A Novel Ant-Based Clustering Approach for Document Clustering. In *Asia Information Retrieval Symposium*, Singapore, 2006.

[15] Yulan He and Steve Young. Semantic processing using the hidden vector state model. *Computer Speech and Language*, 19(1):85–106, 2005.

[16] Deyu Zhou, Yulan He, and Chee Keong Kwoh. Extracting Protein-Protein Interactions from the Literature using the Hidden Vector State Model. In *The First International Conference on Computational Systems Biology*, Shanghai, China, 2006.

[17] *20 Newsgroups Data Set*, 2006. http://people.csail.mit.edu/jrennie/20Newsgroups/.

[18] Jeremy Gollub, Catherine A. Ball, Gail Binkley, Janos Demeter, and et al. The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Research*, 31(1), 2003.