Genomics & Informatics Vol. 2(2) 99-106, June 2004

PubMiner: Machine Learning-based Text Mining for Biomedical Information Analysis

Jae-Hong Eom and Byoung-Tak Zhang*

Biointelligence Laboratory, School of Computer Science and Engineering, Seoul National University, Seoul 151-744, Korea

Abstract

In this paper we introduce PubMiner, an intelligent machine learning based text mining system for mining biological information from the literature. PubMiner employs natural language processing techniques and machine learning based data mining techniques for mining useful biological information such as protein-protein interaction from the massive literature. The system recognizes biological terms such as gene, protein, and enzymes and extracts their interactions described in the document through natural language processing. The extracted interactions are further analyzed with a set of features of each entity that were collected from the related public databases to infer more interactions from the original interactions. An inferred interaction from the interaction analysis and native interaction are provided to the user with the link of literature sources. The performance of entity and interaction extraction was tested with selected MEDLINE abstracts. The evaluation of inference proceeded using the protein interaction data of *S. cerevisiae* (bakers yeast) from MIPS and SGD.

Keywords: Biomedical Text Mining, Data Mining, Machine Learning, Software Application

Introduction

New scientific discoveries are based on the existing knowledge which has to be accessible and thus usable by the scientific community (Andrade *et al*., 2000). In the 19th century, the spread of scientific information was still done by writing letters with new discoveries to a small number of colleagues. Printed journals took over this job professionally. We are currently on another transition into electronic media. Electronic storage allows the customized extraction of information from the literature and its combination with other data resources such as heterogeneous databases. In fact, it is not only an opportunity, but also a pressing need as the volume of scientific literature is increasing immensely. Furthermore, the scientific community is growing so that even for a rather specialized field it becomes impossible to stay up-to-date just through personal contacts in that particular community. The growing amount of knowledge also increases the chance for new ideas based on the combination of solutions from different fields. And there is a necessity of accessing and integrating all scientific information to be able to judge the own progress and to get inspired by new questions and answers.

Since the human genome sequences have been decoded, especially in biology and bioinformatics, there are more and more people devoted to this research domain and hundreds of on-line databases characterizing biological information such as sequences, structures, molecular interactions, and expression patterns (Chiang *et al*., 2004). Despite the prevalent topic of research, the end result of all biological experiments is a publication in the form of text. However, information in text form, such as MEDLINE (*http://www.pubmed.gov/*), is a greatly underutilized source of biological information to biological researchers. Because it takes lots of time to obtain the important and precise information from huge databases with daily increase. Thus knowledge discovery from a large collection of scientific papers is very importantfor efficient biological and biomedical research. Until now, a number of tools and approaches have been developed to resolve such needs. There are many systems analyzing abstracts in MEDLINE to offer bio-related information services. Suiseki (Blaschke *et al.,* 1999; Blaschke *et al*., 2002) and BioBiblioMetrics (Stapley *et al*., 2000) focus on the protein-protein interaction extraction and visualization. MedMiner (Tanabe *et al*., 1999) utilizes external data sources such as GeneCard (Safran *et al*., 2003) and MEDLINE for offering structured information about specific key-words provided by the user. AbXtract (Andrade *et al*., 1998) labels the protein function in the input text and XplorMed (Perez-Iratxeta *et al*., 2000) presents the user specified information through the interaction with user. GENIES

*Corresponding author: E-mail btzhang@bi.snu.ac.kr, Tel +82-2-880-1833, Fax +82-2-880-1847
Accepted 15 May 2004

(Friedman *et al*., 2001) discovers more complicated information such as pathways from journal abstracts. Recently, MedScan (Daraselia *et al*., 2004) employed full-sentence parsing technique for the extraction of human protein interactions from MEDLINE.

Generally, these conventional systems rely on basic natural language processing (NLP) techniques when analyzing literature data. And the efficacy of such systems heavily depends on the rules for processing raw information. Such rules have to be refined by human experts, entailing the possibility of lack of clarity and coverage. In order to overcome this problem, we used machine learning techniques in combination with natural language processing techniques to analyze the interactions among the biological entities. Our method also incorporated several data mining techniques for the extensive discovery, i.e. detection of the interactions which are not directly described in the text.

We have developed PubMiner (Publication-based Text Mining system) which performs efficient interaction mining of biological entities such as gene, protein, and enzymes. For the performance evaluation, the budding yeast (*S. cerevisiae*) was used as the model organism. The goal of our text mining system is to design and develop an information system that can efficiently retrieve the biological entity-related information from the MEDLINE, where the biological entity-related information includes biological function of entities (e.g., gene, protein, and enzymes etc.), related gene or protein, and relation of gene or proteins. Especially we focus on interactions between entities.

System Description

PubMiner, a machine learning based text mining platform, consist of three key components: natural language processing, machine learning based inference, and visualization module.

Information Extraction

The interaction extraction module is based on the NLP techniques adapted to take into account the properties of biomedical literature. It includes a part-of-speech (POS) tagger, a named-entity tagger, a syntactic analyzer, and an event extractor. The POS tagger based on hidden Markov models (HMMs) was adopted for tagging biological words as well as general ones. The named-entity tagger, based on support vector machines (SVMs), recognizes the region of an entity and assigns a proper class to it. The syntactic analyzer recognizes base phrases and detects the dependency represented in them. Finally, the event extractor finds the binary relation using the syntactic information of a given sentence, co-occurrence statistics between two named entities, and pattern information of an event verb. General medical term was trained with UMLS meta-thesaurus (Humphreys *et al*., 1998) and the biological entity and its interaction was trained with GENIA corpus (Kim *et al*. 2003). And the underlying NLP approach for named entity recognition is based on the system of Hwang (Hwang *et al*., 2003) and Lee (Lee *et al*., 2003). Figure 1 shows the schematic architecture of information extraction module.

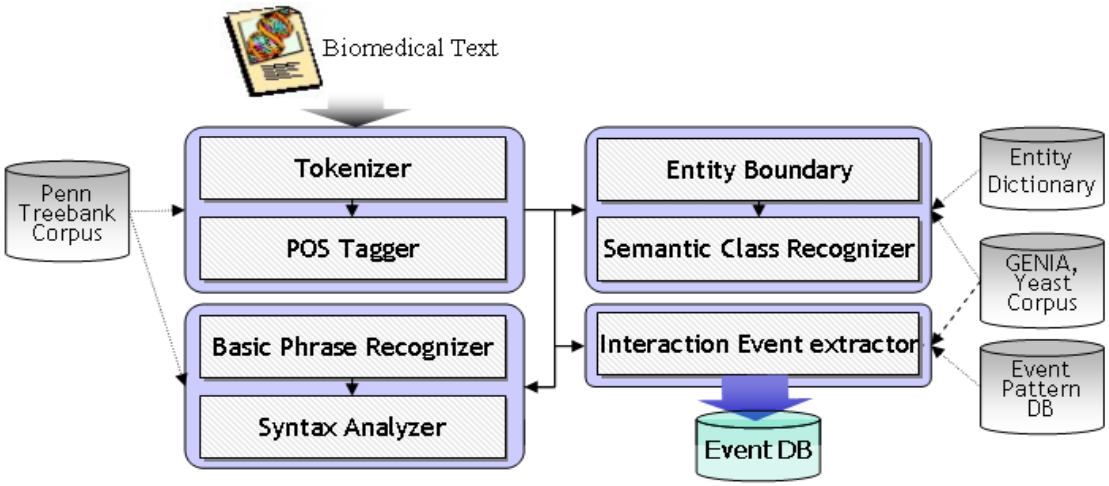

Fig. 1. The schematic architecture of interaction extraction module. The resulting event DB contains interactions between entities. Event pattern database was constructed from the GENIA corpus.

Interaction Inference

The relation inference module, which finds common features and group relations, is based on data mining and machine learning techniques. A set of features of each component of the interaction are collected from public databases such as *Saccharomyces* Genome Database (SGD) (Christie *et al*., 2004) and database of Munich Information Center for Protein Sequences (MIPS) (Mewes *et al*., 2004) and represented as a binary feature vector. An association rule discovery algorithm, Apriori (Agrawal *et al*., 1993) was used to extract the appropriate common feature set of interacting biological entities. In addition, a distribution-based clustering algorithm (Slonim *et al*., 2000) was adopted to analyze group relations. This clustering method collects group relation from the collection of document which contains various biological entities. And the clustering procedure discovers common characteristics among members of the same cluster. It also finds the features describing inter-cluster (between clusters) relations. PubMiner also provides graphical interface to select various options for the clustering and mining. Finally, the hypothetical interactions are generated for the construction of interaction network. The hypotheses correspond to the inferred generalized association rules and the procedure of association discovery is described in the Section of 'Methods.' A set of inferred relations as well as the relations from text analysis are stored in the local database in a systematic way for efficient management of information. Figure 2 describes the schematic architecture of relation inference module.

Visualization

The visualization module shows interactions among the biological entities as a network format. It also shows the documents from which the relations were extracted and inferred. In addition, diverse additional information, such as the weight of association between biological entities could be represented. By this, the user can easily examine the reliability of relations inferred by the system. Moreover, the visualization module shows interaction networks with minimized complexity for comprehensibility and can be utilized as an independent interaction network viewer with predefined input format. Figure 3 shows the overall architecture of visualization module and its interface.

Methods

Feature Selection

In our application, each interaction event is represented by their feature association. Thus it is very important to select optimal feature subset is important to achieve the efficiency of system and to eliminate non-informative association information. Therefore, PubMiner uses feature dimension reduction filter (FDRF) to achieve these objectives.

Each feature of data is considered a random variable and the entropy is used as a measure of the uncertainty of the random variable. The entropy of a variable X is defined as:

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i)) \tag{0.1}$$

And the entropy of X after observing values of another variable Y is defined as:

$$H(X|Y) = -\sum_j P(y_j)\sum_i P(x_i|y_j)\log_2(P(x_i|y_j)) \tag{0.2}$$

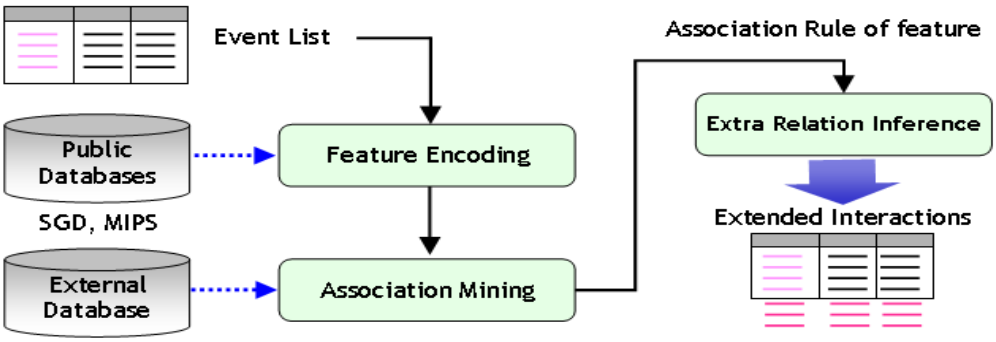

Fig. 2. The schematic architecture of relation inference module.For feature encoding, feature definition of public database such as SGD and MIPS are used. The event list represents the set of interactions which was constructed from previous interaction extraction module. The extended interactions include inferred interaction through the feature association mining.

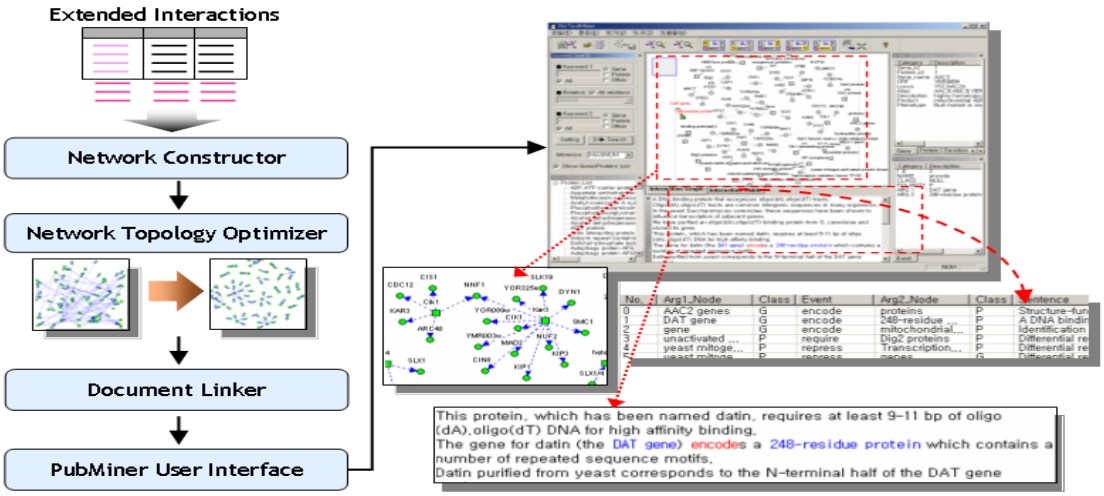

Fig. 3. The workflow diagram of visualization module. The dashed line in the resulting interaction graph stands for the inferred interaction.

where $P(y_i)$ is the prior probability of y_i, and $P(x_i|y_j)$ is the posterior probability of $X=x_i$ given the values of $Y=y_j$. The amount by which the entropy of X decreases reflects additional information about X provided by Y and is called *information gain* (Quinlan *et al.*, 1993), given by:

$$IG(X|Y) = H(X) - H(X|Y) \quad (0.3)$$

According to this measure, a feature Y is considered to be more correlated to feature X than feature Z, if $IG(X|Y) > IG(Z|Y)$. Symmetry is a desired property for a measure of correlations between features and information gain. However, information gain is biased in favor of features with more values and the values have to be normalized to ensure they are comparable and have the same affect. Therefore, here we use symmetrical uncertainty as a measure of feature correlation (Press *et al.,* 1998), defined as follows:

$$SU(X, Y) = 2\left[\frac{IG(X|Y)}{H(X)+H(Y)}\right], 0 \le SU(X, Y) \le 1 \quad (0.4)$$

With symmetrical uncertainty (SU) as feature association measure, we define feature selection procedure which is similar to the definition of Yu (Yu *et al.*, 2003) to reduce the computational complexity. To decide whether a feature is relevant to the protein interaction (interaction class) or not, we use c-correlation and f-correlation which use the threshold SU value δ decided by user. In the proposed method, the class C is divided into two class, conditional protein class (C_C) and result protein class (C_R) of interaction. Figure 4 shows the overall procedure of informative feature selection and the procedure is conducted for each interaction class.

The two feature scoring measures used in the procedure of Figure 4, c-correlation ($SU_{i,c}$) and f-fcorrelation ($SU_{j,i}$), are defined as follows:

Definition 1 (c-correlation $SU_{i,c}$ and f-fcorrelation $SU_{j,i}$). Assume that dataset S contains N ($f_1, \cdots, f_N$) features and a class C (C_C or C_R). Let SU_{ic} denote the SU value that measures the correlation between a feature fi and the class C (call as c-correlation), then the subset S of relevant feature can be decided by a threshold SU value δ, such that $\forall f_i \in S', 1 \le i \le N, SU_{ic} \ge \delta$. And the pair-wise correlation between all features (call as f-correlation) can be defined in same manner of c-orrelation with threshold value δ. f-correlation is used to decide whether relevant feature is redundant or not when considering it with other relevant features.

Mining Feature Association

To predict implicit interaction between entities with feature association, we use conventional data mining method. For this, we adopt association rule discovery algorithm (so-called Apriori algorithm) proposed by Agrawal (Agrawal *et al.*, 1993). Generally, association rule $R(A \Rightarrow B)$ has two values, *support* and *confidence*, representing the characteristics of the association rule. Support (SP) represents the frequency of co-occurrence of all the items appearing in the rule. And confidence (CF) represents the accuracy of the rule computed by dividing the support value by frequency of co-occurrence conditional part items of the rule. These are defined as

Given training dataset $S=(f_1, K, f_N, C)$, where $C=C_C \cup C_R$ and
User-decided threshold δ, do following procedure for each class C_C and C_R.

1. **Repeat** Step 1.1 to 1.2, for all i, i = 1 to N.
   1.1 **Calculate** $SU_{i,C}$ for f_i.
   1.2. **Append** f_i to S'_{list} when $SU_{i,c} \geq \delta$.
2. **Sort** S'_{list} in descending order with $SU_{i,c}$ value.
3. **Set** f_p with the first element of S'_{list}
4. **Repeat** Step 4.1 to 4.3, for all $f_p \neq NLL$.
   4.1 **Set** f_q with the next element of f_p in S'_{list}
   4.2 **Repeat** Step 4.2.1 to 4.2.3, for all $f_q \neq NLL$.
      4.2.1 **Set** $f'_q = f_q$.
      4.2.2 if $SU_{p,q} \geq SU_{q,c}$,
         Remove f_q from S'_{list} and Set f_q with the next element of f'_q in f_q.
      else **Set** f_q with the next element of f_q in S'_{list}.
   4.3 **Set** f_q with next the element of f_q in S'_{list}
5. **Set** $S_{best} = S'_{list}$

Output the most informative optimal feature subset: S_{best}

Fig. 4. The procedures of feature dimension reduction filter (FDRF).

follows:

$$SP(A \Rightarrow B) = P(A \cup B),\ CF(A \Rightarrow B) = P(B|A) \qquad (0.5)$$

where $A \Rightarrow B$ represents association rule, A and B represent items (set of features) in that order. Association rule can be discovered by detecting all the possible rules whose supports and confidences are larger than the user-defined threshold values called minimal support ($SP_{\min}$) and minimal confidence ($CF_{\min}$) respectively. Rules that satisfy both minimum support and minimum confidence threshold are called strong. Here we consider this strong association rules as interesting ones.

An interaction is represented as a pair of two entities that directly binds to each other. To analyze interaction of entities with feature association, we consider each interacting entity pair as transaction of mining data. These transactions with binary vector representation are described in Figure 5. Then we extract associative features generally representing the interaction with association rule mining.

Results

Performance of Entity Extraction

In order to test our entity recognition and interaction extraction module, we built a corpus from 1,000 randomly selected scientific abstracts from PubMed identified to contain biological entity names and interactions via manual searches. The corpus was manually analyzed for biological entities such as protein, gene, and small molecule names in addition to any interaction relationships present in each abstract within the corpus by biologist in our laboratory. Analysis of the corpus revealed 5,928 distinct references to biological

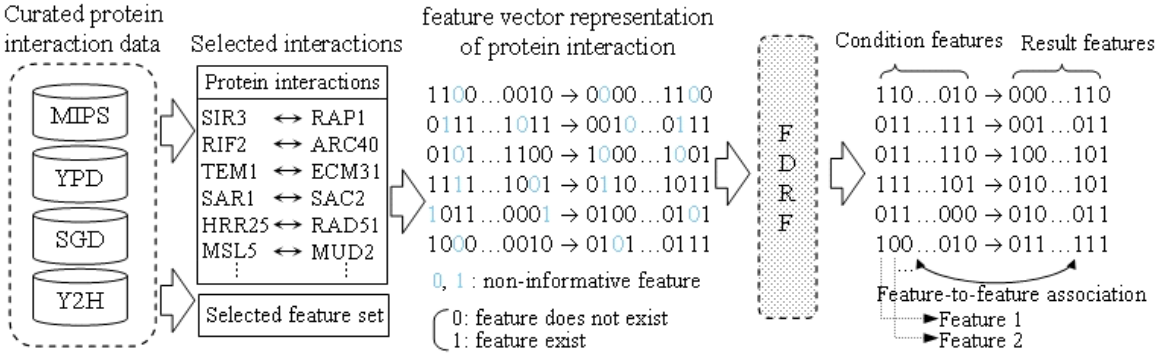

Fig. 5. Binary representation of interactions with feature set. Interactions are represented with feature vector and their associations. After the feature filtering, association mining is applied to the interactions which have filtered features. In this figure, each entities of interaction are presented as 'proteins.'

entities and a total of 3,182 distinct references to interaction relationships. Performance evaluation was done over the same set of 1,000 articles, by capturing the set of entities and interactions recognized by the system and comparing this output against the manually analyzed results previously described. Table 1 shows the statistics of abstract document collection for extraction performance evaluation.

Table 1. The statistics of the document collection.

# of abstracts in collection	# of biological entities	# of interactions
1,000	5,928	3,182

We measured the recall and the precision for both the ability to recognize entity names in text in addition to the ability of the system to extract interactions based on the following calculations:

$$Recall = TP/(TP + FN) \quad (0.6)$$

$$\Pr ecision = TP/(TP + FP) \quad (0.7)$$

where, TP (true positive) is the number of biological entities or interactions that were correctly identified by the system and were found in the corpus. FN (false negative) is the number of biological entities or interactions that the system failed to recognize in the corpus and FP (false positive) is the number of biological entities or interactions that were recognized by the system but were not found in the corpus. Performance test results of the extraction module in the PubMiner are described in Table 2.

Table 2. The precision and recall performance of entity and interaction extraction.

Rcognition Categories	Recall	Precision
Biological entities	83.5	93.1
Interactions of entities	73.9	80.2

Performance of Feature Selection and Association Mining

To test the performance of inference of PubMiner through feature selection (reductions), we used proteinprotein interaction as a metric of entity recognition and interaction extraction. The major protein pairs of the interactions are obtained from the same data source of Oyama (Oyama *et al*., 2002). It includes MIPS, YPD and Y2H by Ito *et al*. and Uetz *et al*., respectively (Mewes *et al*., 2004). Additionally, we also used SGD (Christie *et al*., 2004) to collect more plentiful feature set. Table 3 shows the statistics of interaction data for each data sources and the filtering result with FDRF.

Table 3. The statistics for the proteinprotein interaction dataset.

Data Source	# of interactions	# of initial feature	# of filtered feature
MIPS	10,641	6,232 (total)	1,293 (total)
YPD	2,952		
SGD	1,482		
Y2H (Ito *et al.*)	957		
Y2H (Uetz *et al.*)	5,086		

We performed feature filtering procedure of Figure 4 as a first step of our inference method (δ=0.73) after the

Table 4. Accuracy of the proposed method and the effect (in elapsed time) of filtering optimal informative features with FDRF. Total interactions for prediction are selected from Table 3.

Prediction method	# of interactions			Accuracy (|P|/|T|	Elapsed Tim
	Training set	Test set (T)	Correctly predicted (P)		
Without FDRF	4,628	463	423	91.4 %	212.34 sec
With FDRF	4,628	463	439	94.8 %	143.27 sec
Improvement	—	—	—	3.4 %	32.5 %

feature encoding with the way of Figure 5. Next, we performed association rule mining under the condition of minimal support 10% and minimal confidence 75% on the protein interaction data which have reduced features. And with the mined feature association, we predicted new proteinprotein interaction which have not been used in association training setp. The accuracy of prediction is measured whether the predicted interaction exists in the collected dataset or not. The results are measured with 10 cross-validation.

Discussion

Here, we presented a biomedical text mining system, PubMiner, which screens the interaction data from literature abstracts through natural language analysis, performs inferences based on machine learning and data mining techniques, and visualizes interaction networks with appropriate links to the evidence article. To reveal more comprehensive interaction information, we employed both the data mining approach with optimal feature selection method in addition to the conventional natural language processing techniques. The proposed method achieved the improvement of both accuracy and processing time.

Table 4 gives the advantage of obtained by filtering non-informative (redundant) features and the inference performance of PubMiner. The accuracy of interaction prediction increased about 3.4% with FDRF. And the elapsed time of FDRF based association mining, 143.27 sec, include the FDRF processing time which was 19.89 sec. The elapsed time decrease obtained by using FDRF is about 32.5%. Thus, it is of great importance to reduce number of feature of interaction data for the improvement of both accuracy and execution performance. Thus, we can guess that the information theory based feature filtering reduced a set of misleding or redundnt features of interaction data and this feature reduction eleminated wrong associations and boosted the pocessing time. And the feature association shows the promising results for inferencing implicit interaction of biological entities.

From the result of Table 4, it is also suggested that with smaller granularity of interaction (i.e., not protein, but a set of features of proteins) we could achieve further detailed investigation of the proteinprotein interaction. Thus we cansay that the proposed method is a somewhat suitable approach for an efficient analysis of interactive entity pair which has many features as a back-end module of the generalliterature mining and for the experimentally produced interaction data with moderate false positive ratios.

However, current public interaction data produced by such as high-throughput methods (e.g. Y2H) have many false positives. And several interactions of these false positives are corrected by recent researches through reinvestigation with new experimental approaches. Thus, study on the new method for resolving these problems related to false positive screening further remain as future works.

Acknowledgements

This work was supported by a grant of the Korea Ministry of Science and Technology under the NRL and the Systems Biology programs.

References

Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data* 207-216.

Andrade, M.A. and Valencia, A. (1998). Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics* 14(7), 600-607.

Andrade, M.,A., and Borka, P. (2000). Automated extraction of information in molecular biology. FEBS *Letters* 476, 12-17.

Blaschke, C., Andrade, M.A., Ouzounis, C., and Valencia, A. (1999). Automatic extraction of biological information from scientific text: protein-protein interactions. In *Proc.*

Int. Conf. Intell. Syst. Mol. Biol., 60-67.

Blaschke, C. and Valencia, A. (2002). The frame-based module of the SUISEKI information extraction system. *IEEE Intelligent Systems* 17(2), 14-20.

Chiang, J.,H., Yu, H.,C., and Hsu, H., J. (2004). GIS: a biomedical text-mining system for gene information discovery. *Bioinformatics* 20(1), 120-121.

Christie, K.R., Weng, S., Balakrishnan, R., Costanzo, M.C., Dolinski, K., Dwight, S.S., Engel, S.R., Feierbach, B., Fisk, D.G., Hirschman, J.E., Hong, E.L., Issel-Tarver, L., Nash, R., Sethuraman, A., Starr, B., Theesfeld, C.L., Andrada, R., Binkley, G., Dong, Q., Lane, C., Schroeder, M., Botstein, D., and Cherry, J.M. (2004). Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. *Nucleic Acids Res*. 32(1), D311-D314.

Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A., and Mazo, I. (2004). Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics* 20(5), 604-611.

Friedman, C., Kra, P., Yu, H., Krauthammer, M.,and Rzhetsky, A. (2001). GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* 17(Suppl.1), S74-S82.

Humphreys, B.L., Lindberg, D.A., Schoolman, H.M., and Barnett, G.O. (1998). The Unified Medical Language System: an informatics research collaboration. *J. Am. Med. Inform. Assoc*. 5(1), 1-11.

Hwang, Y.S., Chung, H.J, and Rim, H.C. (2003). Weighted Probabilistic Sum Model based on Decision Tree Decomposition for Text Chunking, *Journal of Computer Processing of Oriental Languages* 16(1), 1-20.

Kim, J.D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GENIA corpus - semantically annotated corpus for bio-textmining. *Bioinformatics* 19(Suppl. 1), i180-182.

Lee, K.J., Hwang, Y.S., and Rim, H.C. (2003). Two-Phase Biomedical NE Recognition based on SVMs. In *Proc. of ACL 2003 Workshop on Natural Language Processing in Biomedicine*, 33-40.

Mewes, H.W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N., Stumpflen, V., Warfsmann, J., and Ruepp, A. (2004). MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res*. 32(1), D41-D44.

Oyama, T., Kitano, K., Satou, K., and Ito, T. (2002). Extraction of knowledge on proteinprotein interaction by association rule discovery. *Bioinformatics* 18, 705-714.

Perez-Iratxeta, C., Bork, P., and Andrade, M.A. (2000). XplorMed: a tool for exploring MEDLINE abstracts. *Trends Biochem Sci*. 26, 573-575.

Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. (1998). Numerical recipes in C (Cambridge: Cambridge University Press).

Quinlan, J.R. (1993). C4.5: Programs for machine learning (San Francisco: Morgan Kaufmann Publishers Inc.).

Safran, M., Chalifa-Caspi, V., Shmueli, O., Olender, T., Lapidot, M., Rosen, N., Shmoish, M., Peter, Y., Glusman, G., Feldmesser, E., Adato, A., Peter, I., Khen, M., Atarot, T., Groner, Y., and Lancet, D. (2003). Human gene-centric databases at the Weizmann institute of science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res*. 31(1), 142-146.

Slonim, N. and Tishby, N. (2000). Document clustering using word clusters via the information bottleneck method. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 208-215.

Stapley, B.J. and Benoit, G. (2000). Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. In *Proc. of Pac. Symp. Biocomput*., 529-40.

Tanabe, L., Scherf, U., Smith, L.H., Lee. J.K., Hunter. L., and Weinstein, J.N. (1999). MedMiner: an internet text-mining tool for biomedical information, with application to gene expression profiling. *BioTechniques* 27, 1210-1217.

Yu, L. and Liu, H. (2003). Feature selection for high dimensional data: a fast correlation-based filter solution. In Proceeding of the *20th International Conference on Machine Learning*, 856-863.