BioPubMiner: Machine Learning Component-Based Biomedical Information Analysis Platform

Jae-Hong Eom and Byoung-Tak Zhang

Biointelligence Lab., School of Computer Science and Engineering,
Seoul National University, Seoul 151-744, South Korea
{jheom,btzhang}@bi.snu.ac.kr

Abstract. In this paper we introduce BioPubMiner, a machine learning component-based platform for biomedical information analysis. BioPubMiner 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 database to infer more interactions from the original interactions. 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.

1 Introduction

Normally, novel scientific discoveries are based on the existing knowledge which has to be accessible and thus usable by the scientific community. In the 19th century, the spread of scientific information was still done by writing letters with new discoveries to a small number of colleagues. This job was taken over professionally by printed journals. Currently, we are on another switch into electronic media. Electronic storage with huge capacity allows the customized extraction of information from the literature and its combination with other data resources such as heterogeneous databases. Indeed, 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 [1].

After 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 [2]. Despite the prevalent topic of research, the end result of all biological experiments is a publication in the form of textbook. However, information in text form, such as MEDLINE (http://www.pubmed.gov), is a greatly underutilized source of biological information

G. Das and V.P. Gulati (Eds.): CIT 2004, LNCS 3356, pp. 11–20, 2004.

to biological researchers. Because it takes lots of time to obtain important and accurate information from this huge databases with daily increase. Thus knowledge discovery from a large collection of scientific papers is become very important for efficient biological and biomedical research. So far, 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 [3, 4] and BioBiblioMetrics [5] focus on the protein-protein interaction extraction and visualization. MedMiner [6] utilizes external data sources such as GeneCard [7] and MEDLINE for offering structured information about specific key-words provided by the user. AbXtract [8] labels the protein function in the input text and XplorMed [9] presents the user specified information through the interaction with user. GENIES [10] discovers more complicated information such as pathways from journal abstracts. Recently, MedScan [11] 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. We 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 BioPubMiner (Biomedical Publication Mining & Analysis System) which performs efficient interaction mining of biological entities. For the evaluation of performance, literature and interaction data of the budding yeast (*S. cerevisiae*) was used as the model organism.

The paper is organized as follows. In Section 2, the major three component of BioPubMiner is described. In Section 3, we describe the methodology of the interaction inference module of BioPubMiner. In Section 4, performance evaluation of each component is given. Finally, concluding remarks and future works are given in Section 5.

2 System Description of BioPubMiner

BioPubMiner, a machine learning based text mining platform, consist of three key components: literature processing, interaction inference, and visualization component.

2.1 Literature Processing

The literature processing module is based on the NLP techniques adapted to take into account the properties of biomedical literature and extract interactions between biological entities. 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 [12] and the biological entity and its interaction was trained with GENIA corpus [13]. And the underlying NLP approach for named entity recognition is based on the system of Hwang *et al*. [14] and Lee *et al*. [15] with collaborations (More detailed descriptions of language processing component are explained in these two papers).

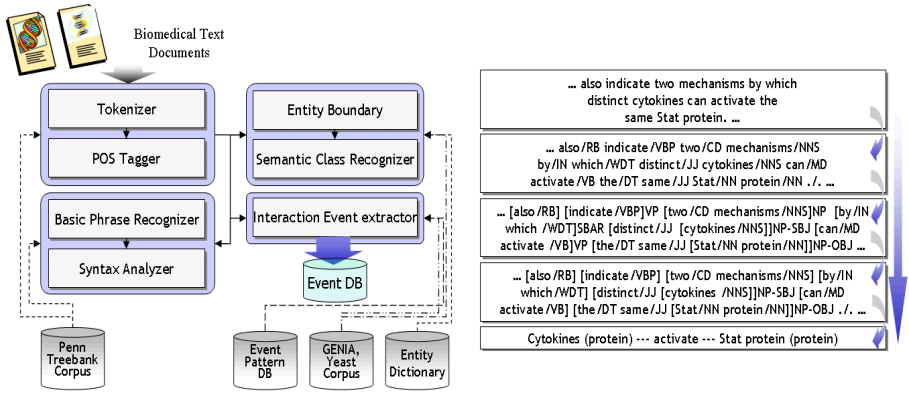

Fig. 1. The block architecture of literature processing module (left) and the example of text processing (right). The resulting event DB contains interactions between entities extracted from domain-documents. Event pattern database was constructed from the GENIA corpus

2.2 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) [16] and database of Munich Information Center for Protein Sequences (MIPS) [17] and represented as a binary feature vector. An association rule discovery algorithm (Apriori [18]) and information theory based feature filter were used to extract the appropriate common feature set of interacting biological entities. In addition, a distribution-based clustering algorithm [19] 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 relations. BioPubMiner 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.' Figure 2 describes the schematic architecture of relation inference module.

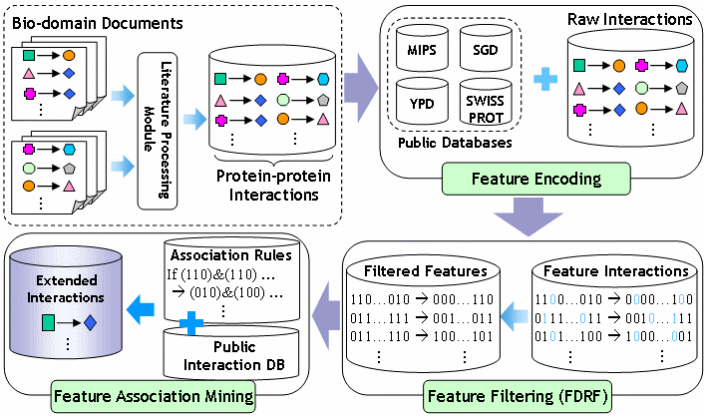

Fig. 2. The block diagram of relation inference module. For feature encoding, feature definition from public database such as SGD and MIPS are used. The raw interaction represents the set of interactions which was constructed from previous literature processing module. The extended interactions include original interactions and inferred interactions through the feature association rule mining

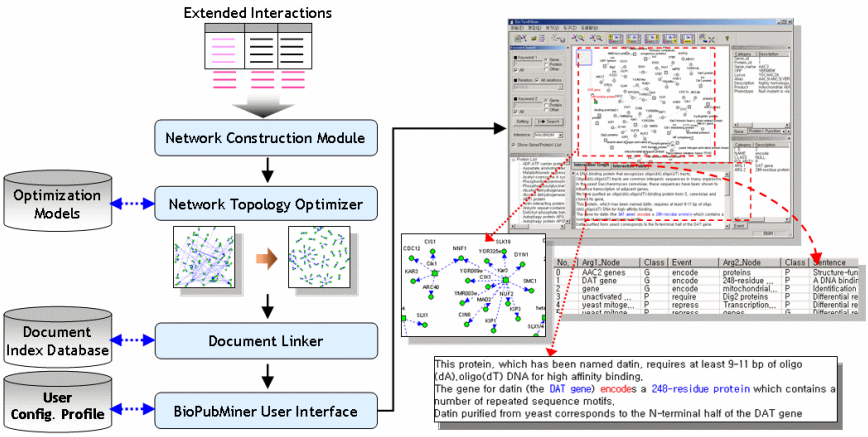

Fig. 3. The workflow diagram of the visualization module. The dashed line in the resulting interaction graph stands for the inferred interactions

2.3 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. Thus 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.

3 Methods

In this section we describe the approaches of the relation inference module of BioPubMiner. Basically, the interaction inference is based on the machine learning theory to find the optimal feature sets. Additionally, association rule discovery method which is widely used in data mining field is used to find general association among the selected optimal features.

3.1 Feature Dimension Reduction by Feature Selection

The correlation between two random variables can be measured by two broad approaches, based on classical linear correlation and based on information theory. Linear correlation approaches can remove features with near zero linear correlation to the class and reduce redundancy among selected features. But, linear correlation measures may not be able to capture correlations that are not linear in nature and the calculation requires all features contain numerical values [20]. In this paper, we use an information theory-based correlation measure to overcome these drawbacks.

Each feature of data can be considered as a random variable and the uncertainty of it can be measured by *entropy*. The entropy of a variable X is defined as

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i)), \tag{1}$$

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

$$H(X \mid Y) = -\sum_j P(y_j)\sum_i P(x_i \mid y_j)\log_2(P(x_i \mid y_j)), \tag{2}$$

where $P(y_j)$ is the prior probability of the value y_j of Y, and $P(x_i \mid y_j)$ is the posterior probability of X being x_i given the values of Y. The amount by which the entropy of X decreases reflects additional information about X provided by Y and is called *information gain* [21], which is given by

$$IG(X \mid Y) = H(X) - H(X \mid Y) \tag{3}$$

It is said that symmetry is a desired property for a measure of correlation between features and information gain [20]. 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, we used the *symmetrical uncertainty* as a measure of feature correlation [22], defined as

$$SU(X,Y) = 2\left[\frac{IG(X \mid Y)}{H(X) + H(Y)}\right],\ 0 \le SU(X,Y) \le 1 \tag{4}$$

Figure 4 shows the overall procedure of the correlation-based feature dimension reduction filter which was earlier introduced by Yu *et al*. [20], named fast correlation-based filter (FCBF). In this paper, we call this FCBC procedure as feature dimension reduction filter (FDRF) for our application. The algorithm finds a set of principal features S_{best} for the class concept. First, the procedure in Figure 4 calculates the symmetrical uncertainty (SU) values for each feature, selects relevant feature into S'_{list} based on the predefined threshold δ, and constructs an ordered list of them in descending order according to their SU values. Next, it further processes the ordered list to remove redundant features and only keeps principal ones among all the selected relevant features.

With symmetrical uncertainty as a feature association measure, we reduce the feature dimension through the feature selection. In Figure 4, the class C is divided into two classes, conditional protein class (C_C) and result protein class (C_R) of interaction. The relevance of a feature to the protein interaction (interaction class) is decided by the value of *c*–correlation and *f*–correlation, where an SU value δ is used as a threshold value. These two correlations are defined in the paper of Yu *et al*. [20].

Given training dataset $S = (f_1, \ldots, f_N, C)$, 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 $f_p \neq NULL$.
   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 $f_q \neq NULL$.
      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 S'_{list}.
         else **Set** f_q with the next element of f_q in S'_{list}.
      4.2.3 **Set** f_q with the next element of f_q in S'_{list}
   4.3 **Set** f_p with the next element of f_p 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)

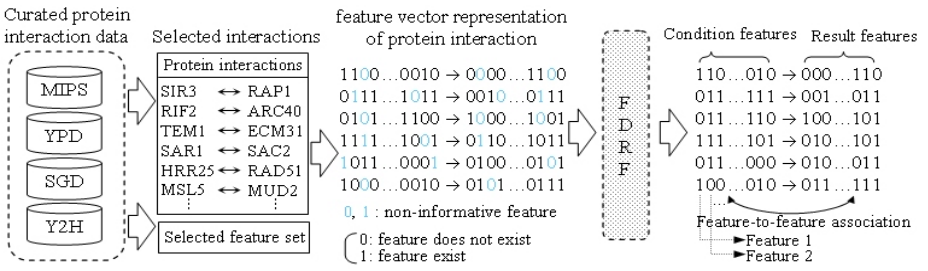

Fig. 5. Protein interaction as feature vector forms. Each interaction is represented with binary feature vector (whether the feature exists or not) and their associations. The FDRF sets those features as "don't care" (D/K) which have SU value less than given SU threshold . This is intended to consider in association mining only those features that have greater SU value than a given threshold. The features marked D/K are regarded as D/K also in association rule mining (i.e., these features are not counted in the calculation of support and confidence). These features are not shown in the vector representation of right side of Figure 5

3.2 Feature Association Mining

Entity Interaction as Feature Association

After the extraction of interaction from literature, each 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. Using association rule mining, then, we extract association of features which generalize the interactions.

Association Mining

To predict protein–protein interaction with feature association, we adopt the association rule discovery data mining algorithm (so-called Apriori algorithm) proposed by Agrawal *et al*. [18]. Generally, an 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*) is the accuracy of the rule, which is calculated by dividing the *SP* value by the frequency of the item in conditional part of the rule.

$$SP(A \Rightarrow B) = P(A \cup B),\ CF(A \Rightarrow B) = P(B \mid A) \tag{5}$$

where $A \Rightarrow B$ represents association rule for two items (set of features) A and B 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 value called minimal support (SP_{min}) and minimal confidence (CF_{min}) respectively. Rules that satisfy both minimum support and minimum confidence threshold are taken as to be *strong*. Here we consider these strong association rules as interesting ones.

In this work, we use the same association rule mining and the scoring approach of Oyama *et al*. [23] for performance comparison with respect to the execution time.

4 Experimental Results

Performance of Literature Processing

To test the performance of entity recognition and interaction extraction of our literature processing module, we built a corpus from 1,500 randomly selected scientific abstracts from PubMed identified to contain biological entity names and interactions through 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 6,988 distinct references to biological entities and a total of 4,212 distinct references to interaction relationships. Performance evaluation was done over the same set of 1,500 articles, by capturing the set of entities and interactions recognized by the system and comparing this output against the manually analyzed results previously mentioned. Table 1 shows the statistics of abstract document collection for extraction performance evaluation.

Table 1. The statistics for the test document collection

# of abstracts in collection	# of biological entities	# of interactions
1,500	6,988	4,212

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:

$$\begin{aligned} \text{Recall} &= TP / (TP + FN) \\ \text{Precision} &= TP / (TP + FP) \end{aligned} \tag{6}$$

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 are described in Table 2.

Table 2. The precision and recall performance of the entities and interaction extraction

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

Performance of Inference Through Feature Selection and Association Mining

To test the performance of inference module of BioPubMiner through feature selection (reductions), we used protein–protein 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 *et al*. [23]. It includes MIPS [17], YPD and Y2H by Ito *et al*. [24] and Uetz *et al*. [25]. Additionally, we used SGD [16] to collect more lavish feature set. Table 3 shows the statistics of interaction data for each data source and the filtering result with FDRF of Figure 4.

Table 3. The statistics for the protein–protein interaction dataset

Data Source	# of interactions	# of initial features	# of filtered features
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 ($\delta = 0.73$) after the feature encoding with the way of Figure 5. Next, we performed association rule mining under the condition of minimal support 9 and minimal confidence 75% on the protein interaction data which have reduced features. Next, we predicted new protein–protein interaction which have not 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 for more realistic evaluation.

Table 4 gives the advantage of obtained by filtering non-informative (redundant) features and the inference performance of BioPubMiner. 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.

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

Prediction method (Association ming)	# of interactions			Accuracy (|P|/|T|)	Elapsed Time
	Total	Excluded (T)	Predicted (P)		
Without FDRF	4,628	463	423	91.4 %	212.34 sec
With FDRF	4,628	463	439	94.8 %	143.27 sec

5 Conclusions

In this paper, we presented a component-based biomedical text analysis platform, BioPubMiner, 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 main two component of the proposed system (literature processing and interaction inference) achieved some improvement. 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 protein–protein interaction. Thus we can say 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 general literature 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 research was supported by the Korean Ministry of Science and Technology under the NRL Program and the Systems Biology Program. The RIACT at Seoul National University provided research facilities for this study.

References

1. Andrade, M.,A., *et al*.: Automated extraction of information in molecular biology. *FEBS Letters* **476** (2000) 12-17.
2. Chiang, J.,H., *et al*.: GIS: a biomedical text-mining system for gene information discovery. *Bioinformatics* **20**(1) (2004) 120-21.
3. Suiseki. http://www.pdg.cnb.uam.es/suiseki/index.html.
4. Blaschke, C., *et al*.: Automatic extraction of biological information from scientific text: protein-protein interactions. In *Proceedings of ISMB'99* (1999) 60-67.
5. BioBiblioMetrics. http://www.bmm.icnet.uk/~stapleyb/biobib/.
6. Tanabe, L., *et al*.: MedMiner: an internet text-mining tool for biomedical information, with application to gene expression profiling. *BioTechniques* **27** (1999) 1210-17.
7. Safran, M., *et al*.: Human gene-centric databases at the Weizmann institute of science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res.* **31**(1) (2003) 142-46.
8. Andrade, M.A., *et al*.: Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics* **14**(7) (1998) 600-07.
9. Perez-Iratxeta, C., *et al*.: XplorMed: a tool for exploring MEDLINE abstracts. *Trends Biochem. Sci.* **26** (2001) 573-75.
10. Friedman, C., *et al*.: GENIS: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* **17**(Suppl.1) (2001) S74-S82.
11. Daraselia, N., *et al*.: Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics* **20**(5) (2004) 604-11.
12. Humphreys, B. L., *et al*.: The Unified Medical Language System: an informatics research collaboration. *J. American Medical Informatics Association* **5** (1998) 1-11.
13. Kim J.D., *et al*.: GENIA corpus - semantically annotated corpus for bio-textmining. *Bioinformatics* **19**(Suppl 1) (2003) i180-182.
14. Hwang, Y.S., *et al*.: Weighted Probabilistic Sum Model based on Decision Tree Decomposition for Text Chunking, *J. Computer Processing of Oriental Languages* **16**(1) (2003) 1-20.
15. Lee, K.J., *et al*.: Two-Phase Biomedical NE Recognition based on SVMs. In *Proceedings of ACL'03 Workshop on Natural Language Processing in Biomedicine* (2003) 33-40.
16. Christie, K.R., *et al*.: 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) (2004) D311-14.
17. Mewes, H.W., *et al*.: MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.* **32**(1) (2004) D41-44.
18. Agrawal, R., *et al*.: Mining association rules between sets of items in large databases. In *Proceedings of ACM SIGMOD'93* (1993) 207-16.
19. Slonim, N., *et al*.: Document clustering using word clusters via the information bottleneck method. In *Proceedings of SIGIR'2000* (2000) 208-15.
20. Yu, L., *et al*.: Feature selection for high dimensional data: a fast correlation-based filter solution. In *Proceeding of ICML'03* (2003) 856-63.
21. Quinlan, J.: C4.5: Programs for machine learning. *Morgan Kaufmann* (1993).
22. Press, W.H., et al.: Numerical recipes in C. *Cambridge University Press* (1988).
23. Oyama, T., *et al*.: Extraction of knowledge on protein–protein interaction by association rule discovery. *Bioinformatics* **18** (2002) 705-14.
24. Ito, T., *et al*.: A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA* **98** (2001) 4569-74.
25. Uetz, P., *et al*.: A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403** (2000) 623-27.