PubMiner: Machine Learning-Based Text Mining System for Biomedical Information Mining

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. PubMiner, an intelligent machine learning based text mining system for mining biological information from the literature is introduced. PubMiner utilize natural language processing and machine learning based data mining techniques for mining useful biological information such as protein-protein interaction from the massive literature data. The system recognizes biological terms such as gene, protein, and enzymes and extracts their interactions described in the document through natural language analysis. The extracted interactions are further analyzed with a set of features of each entity which were constructed 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 evaluation of system performance proceeded with the protein interaction data of *S.cerevisiae* (bakers yeast) from MIPS and SGD.

Keywords: Natural Language Processing, Data Mining, Machine Learning, Bioinformatics, and Software Application

1 Introduction

New scientific discoveries are founded on the existing knowledge which has to be easy to get to and thus usable by the scientific community. Electronic storage allows the customized extraction of information from the literature and its combination with other data resources such as heterogeneous databases. 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 combining 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].

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 [2]. Even though

C. Bussler and D. Fensel (Eds.): AIMSA 2004, LNAI 3192, pp. 216–225, 2004.

the widespread topics of research, the end result of all biological experiments is a publication in the form of text. Information in text form such as MEDLINE[1], however, is a greatly underutilized source of biological information to the biological researchers. 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 important for 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. For example, Suiseki [3] and BioBiblioMetrics [4] focus on the protein-protein interaction extraction and visualization. MedMiner [5] utilizes external data sources such as GeneCard [6] and MEDLINE for offering structured information about specific keywords provided by the user. AbXtract [7] labels the protein function in the input text and XplorMed [8] presents the user specified information through the interaction with user. GENIES [9] discovers more complicated information such as pathways from journal abstracts. Recently, MedScan [10] employed full-sentence parsing technique for the extraction of human protein interactions from MEDLINE. And there a number of approaches related to text mining for biology and other fields [11].

Generally, these conventional systems rely on basic natural language processing (NLP) techniques when analyzing literature data. And the efficacy of such systems greatly depends on the rules for processing unrefined 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 conventional NLP techniques. Our method also incorporated several data mining techniques for the extensive discovery, i.e., detection of the interactions which are not explicitly described in the text.

We have developed PubMiner (Publication Text Mining system) which performs efficient mining of gene and protein interactions. For the evaluation, the budding yeast (*S. cerevisiae*) was used as a 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), related gene or protein, and relation of gene or proteins. Especially we focus on interactions between entities.

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

2 System Description

The system, PubMiner, consists of three key components: natural language processing, machine learning-based inference, and visualization module.

[1] http://www.pubmed.gov

2.1 Interaction 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 [12] and the biological entity and its interaction was trained with GENIA [13] corpus. The underlying NLP approaches for named entity recognition are based on the system of Hwang *et al*. [14] and Lee *et al*. [15] with collaborations. More detailed descriptions of language processing are elucidated in [16]. Figure 1 shows the schematic architecture of the interaction extraction module.

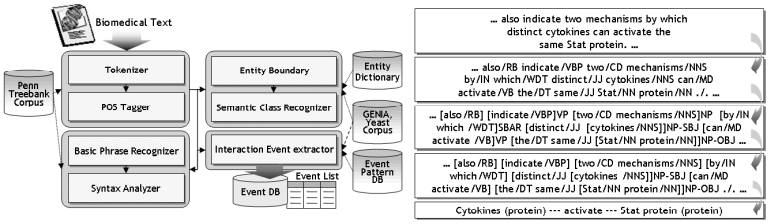

Fig. 1. The schematic architecture of the interaction extraction module (left) and the example of sentence parsing (right). The resulting event DB contains interactions between entities. Event pattern database was constructed from the GENIA corpus and tuned to yeast domain with manually tagged documents.

2.2 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) [17] and database of Munich Information Center for Protein Sequences (MIPS) [18] and represented as a binary feature vector. An association rule discovery algorithm, Apriori [19] was used to extract the appropriate common feature set of interacting biological entities. In addition, a distribution-based clustering algorithm [20] 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 user 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 Section 3. The inferred relations as well as the original relations 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.

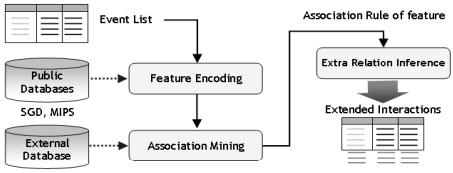

Fig. 2. The schematic architecture of relation inference module. For feature encoding, feature definition of public database such as SGD and MIPS are used. The event list represents the set of interactions which was constructed from previous interaction extraction module. The extended interactions include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

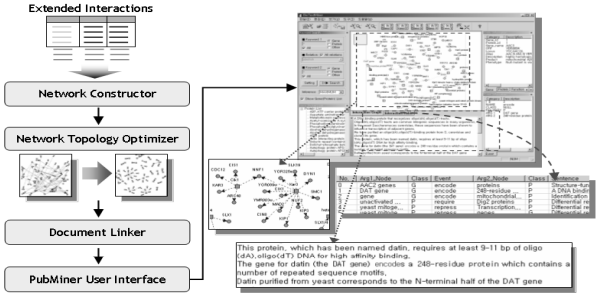

Fig. 3. The workflow diagram of the visualization module (left) and its interface (right). The dashed lines in the resulting interaction graph represent the inferred interactions.

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, this 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 methodology of the relation inference module of PubMiner. Basically, the relation 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 Selection

In our application, each interaction event is represented by their feature association. Thus the selection of 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 which was earlier introduced by Yu *et al*. [21], named fast correlation-based filter (FCBF), to achieve these objectives. Here, we call this FCBC procedure as feature dimension reduction filter (FDRF) for our application.

Each feature of data can be considered as 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:

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

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* [22], given by

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

According to this measure, a feature Y is considered to be more correlated to feature X than feature Z, if $IG(X \mid Y) > IG(Z \mid Y)$. Symmetry is a desired property for a measure of correlation 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 the *symmetrical uncertainty* as a measure of feature correlation [23], defined as:

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

With symmetrical uncertainty (SU) as feature association measure, we use the feature selection procedure of Yu *et al*. [21] to reduce the computational complexity of association mining. To decide whether a feature is relevant to the protein interaction (interaction class) or not, we use two measures, *c*–correlation and *f*–correlation which use the threshold SU value δ decided by user. The class C in Figure 4 is divided into two classes, conditional protein class (C_C) and result protein class (C_R) of interaction.

Definition 1. (*c*–correlation $SU_{i,c}$, *f*–correlation $SU_{j,i}$). Assume that dataset S contains N ($f_1,\ldots, f_N$) features and a class C (C_C or C_R). Let $SU_{i,c}$ denote the SU value that measures the correlation between a feature f_i and the class C (called *c*–correlation), then a subset S' of relevant features can be decided by a threshold SU value δ, such that $\forall f_i \in S'$, $1 \le i \le N$, $SU_{i,c} \ge \delta$. And the pair-wise correlation between all features

(called *f*–correlation) can be defined in same manner of *c*–correlation with a threshold value δ. The value of *f*–correlation is used to decide whether relevant feature is redundant or not when considering it with other relevant features.

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 next the 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).

3.2 Mining Feature Association

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*. [19]. 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*. [24] for performance comparison with respect to the execution time.

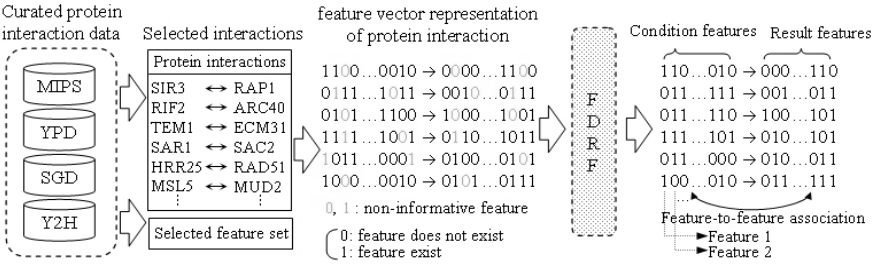

Fig. 5. Representation of protein interaction by feature vectors. Each interaction is represented with binary feature vector (whether the feature exists or not) and their associations. The FDRF sets those features as "don't care" (D/K) which have *SU* value less than given *SU* threshold δ. This is intended 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.

Entity Interaction with Feature Association

An interaction is represented as a pair of two proteins that directly bind to each other. To analyze protein–protein interactions with feature association, we consider each interacting protein pair as a transaction of data mining. 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.

4 Experimental 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 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 within our laboratory. Analysis of the corpus revealed 5,928 distinct references to biological entities and a total of 3,182 distinct references to interaction relationships. Performance evaluation was conducted 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 for 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:

$$\text{Recall} = TP / (TP + FN), \quad \text{Precision} = TP / (TP + FP)$$

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 the entities 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 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*. [24]. It includes MIPS, YPD and Y2H by Ito *et al*. and Uetz *et al*., respectively [18]. Additionally, we use SGD [17] to collect more abundant feature set. Table 3 shows the statistics of interaction data for each data source and filtering 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 (δ=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. And with the mined feature association, 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.

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	Elapsed Time
	Total	Excluded	Predicted		
Without FDRF	4,628	463	423	91.4 %	212.34 sec
With FDRF	4,628	463	439	94.8 %	143.27 sec

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 32.5%. Thus, it is of great importance to reduce number of feature of interaction data for the improvement of both accuracy and execution time. 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 over all pocessing time. And the feature association shows the promising results for inferencing implicit interaction of biological entities.

5 Conclusions

In this paper, 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. 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 suitable for an efficient analysis of interactive entity pair which has many features and this approach is also suitable as a back-end module of general literature mining.

But, current public interaction data produced by 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 with respect to literature mining 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.

References

1. Andrade, M.,A., and Borka, P.: 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) 120–121
3. Blaschke, C., *et al*.: Automatic extraction of biological information from scientific text: protein–protein interactions. In *Proc. of ISMB–1999*, Heidelberg, Germany (1999) 60–67
4. BioBiblioMetrics. http://www.bmm.icnet.uk/~stapleyb/biobib/
5. Tanabe, L., *et al*.: MedMiner: an internet text-mining tool for biomedical information, with application to gene expression profiling. *BioTechniques* **27** (1999) 1210–1217
6. 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) 142–146
7. Andrade, M.,A., Valencia, A.,: Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics* **14**(7) (1998) 600–607
8. Perez-Iratxeta, C., *et al*.: XplorMed: a tool for exploring MEDLINE abstracts. *Trends Biochem. Sci.* **26** (2001) 573–575
9. 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
10. Daraselia, N., *et al*.: Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics* **20**(5) (2004) 604–611
11. Nédellec C., *et al*.: Machine learning for information extraction in genomics – state of the art and perspectives. In: Sirmakessis, S. (ed.) : Text Mining and its Applications. *Studies in Fuzzi. and Soft Comp.* **138**. Springer–Verlag, Berlin Heidelberg New York (2004) 99–118
12. Humphreys, B. L., et al.: The Unified Medical Language System: an informatics research collaboration. *J. Am. Med. Inform. Assoc.* **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. *Int. J. Comp. Proc. Orient. Lang.* **16**(1) (2003) 1–20
15. Lee, K.J., *et al*.: Two-phase biomedical NE recognition based on SVMs. In *Proc. of ACL 2003 Workshop on Natural Language Processing in Biomedicine* (2003) 33–40
16. Eom, J.H., *et al*.: PubMiner – a machine learning-based biomedical text mining system. *Technical Report* (*BI–TR0401*), Biointelligence Lab., Seoul National University (2004)
17. 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–D314
18. Mewes, H.W., *et al*.: MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.* **32**(1) (2004) D41–D44
19. Agrawal, R., *et al*.: Mining association rules between sets of items in large databases. In *Proc. of ACM SIGMOD–1993*, Washington D.C., USA (1993) 207–216
20. Slonim, N., and Tishby, N.: Document clustering using word clusters via the information bottleneck method. In *Proc. of SIGIR–2000*, Athens, Greece (2000) 208–215
21. Yu, L. and Liu, H.: Feature selection for high dimensional data: a fast correlation-based filter solution. In *Proc. of ICML–2003*, Washington D.C., USA (2003) 856–863
22. Quinlan, J.: C4.5: Programs for machine learning. Morgan Kaufmann. (1993)
23. Press, W.H., *et al*.: Numerical recipes in C. Cambridge University Press. (1988)
24. Oyama, T., *et al*.: Extraction of knowledge on protein–protein interaction by association rule discovery. *Bioinformatics* **18** (2002) 705–714