社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－BIO－2 (2)
2005／10／7

確率モデルに基づくテキスト共起データからのマイニング

Shanfeng Zhu[a], 奥野 恭史[b], 辻本 豪三[b], 馬見塚 拓[a]
[a] 京都大学化学研究所 バイオインフォマティクスセンター
[b] 京都大学大学院薬学研究科

生物情報データからのテキストマイニングでは、文献上の関連する事象を見つけ出すことが非常に重要である。本研究では、生体内低分子化合物と遺伝子との関係に着目する。すなわち、両者に関連する様々な事象から、関係のある低分子化合物と遺伝子を自動的に抽出することを試みる。この問題に対して、本研究では、確率モデルに基づくアプローチを提案する。より具体的には、アスペクトモデルと呼ばれる既存の共起モデルを拡張した混合アスペクトモデルを提案し、同時に EM(Expectation-Maximization) アルゴリズムに基づくパラメータ推定手法を提案する。本研究手法が「低分子化合物-遺伝子」間の新しい関係を抽出可能であることを、MEDLINE から得られたデータを用いた実験から示す。

Mining literature co-occurrence data using a probabilistic model

Shanfeng Zhu[a], Yasushi Okuno[b], Gozoh Tsujimoto[b], Hiroshi Mamitsuka[a]
[a] Bioinformatics Center, Institute for Chemical Research, Kyoto University
[b] Graduate School of Pharmaceutical Sciences, Kyoto University

A fundamental issue in biomedical text mining is to mine biological related entities from literature. Here we focus on the co-occurrence of chemical compounds and genes in literature. That is, we try to find the co-occurrence of compounds and genes from any other types of co-occurrences which are related to genes as well as compounds. We propose a probabilistic model, called the mixture aspect model (MAM), and an algorithm for estimating its parameters to efficiently handle different types of co-occurrence datasets at once. We evaluated our method through experimentation on three different types of co-occurrence datasets (i.e. compound-gene, gene-gene and compound-compound co-occurrences) generated from the MEDLINE. Experimental results have shown that when predicting co-occurred compound-gene and compound-compound pairs, MAM trained by all datasets outperformed any simple models trained by other combinations of datasets with the difference being statistically significant in all cases. For example, we achieved the AUC of 95% using all types of co-occurrences while the AUC obtained using gene-compound co-occurrences only is just 85%.

1 Introduction

Mining literature for biomedical knowledge discovery has become a very active field in bioinformatics recently. One of the important applications is to discover the relationship among genes, proteins, disease phenotype and chemical compounds. Co-occurrence in MEDLINE is a simple and popular technique for discovering possible biological relationships among different entities. This technique is based on the following hypothesis: if biological entity A co-occurs with biological entity B in the same MEDLINE record (title and abstract), A and B should be biologically related with high probability. This hypothesis was experimentally testified by many researchers[1, 2]. In this study, we also employ co-occurrence technique to identify biologically related genes and chemical compounds. We focus on the co-occurrence information in the literature to discover implicit related entities, e.g. "chemical compound - gene", being those which are not in existing co-occurrences in the literature but could be discovered from the co-occurrence data.

We propose a probabilistic model, which we call a mixture aspect model (MAM), coupled with an efficient algorithm for estimating its parameters. MAM is an extension of a probabilistic model, called the aspect model (AM) developed in natural language processing[3], with one significant difference. MAM can incorporate different types of co-occurrence data efficiently. More formally, the probabilistic

structure of MAM is a weighted mixture of (normalized) AMs, and each component (i.e. AM) handles one type of co-occurrence data. Our algorithm for estimating the probability parameters of MAM is based on the EM (Expectation-Maximization) algorithm[4] that locally maximizes the likelihoods of given data. Once the probability parameters of MAM are estimated, MAM can predict the likelihood for any pair of events, such as a pair of a chemical compound and a gene.

We evaluated our approach by performing experiments using real datasets. We generated three types of co-occurrence data: gene-gene, compound-compound and compound-gene from the MEDLINE records[5]. In our experiments, we first checked the performance of MAM to predict the co-occurrences of compounds and genes by using cross-validation, starting with compound-gene pairs and then adding compound-compound pairs, followed by gene-gene pairs. Experimental results have shown that adding gene-gene (or compound-compound) pairs improved the performance of using compound-gene pairs only, with the difference being statistically significant. In particular, we found that adding compound-compound pairs is the most effective in improving the performance of predicting compound-gene pairs. We then performed a similar experiment for compound-compound pairs and found that the performance improvement was obtained in almost the same way. These results indicate that combining all these datasets is effective in our problem setting, and that MAM and its learning algorithm are extremely useful for obtaining the results.

2 Background

Co-occurrence in MEDLINE is a simple, effective and popular technique for identify biological relationships among different entities. This technique is based on the hypothesis that entities appearing in the same MEDLINE record are more likely to be biologically related. This hypothesis has been verified by many researchers. Jenssen *et al.* (2001) which presented a gene-to-gene co-occurrence network called PubGene using over 10 million MEDLINE records. They randomly selected 500 pairs of genes co-occurred once and 500 pairs of genes co-occurred more than five times in the MEDLINE, then manually analyzed the biological relationship of these pairs by expertise. They found that the accuracy of biological relationship identification is around 60% for the first group, and 72% for the second. In further analysis, they found that almost all errors were due to the failures in gene name recognition. Chang *et al.* (2004) also identified related genes and drugs based on their co-occurrence in the titles and abstracts of publications in MEDLINE. They manually examined the biological relationship of 100 gene-drug pairs. They found that out of the 100 pairs (50 of them with largest number of co-occurrence, and 50 of them randomly selected), 70 shared some biological relationships. From these studies, we can see that co-occurrence methods can successfully find biological relationships, and most of failures are due to the difficulty of biological entity name identification in extracting MEDLINE texts. We emphasize that in our experiment we generated our co-occurrence data not directly from MEDLINE texts, but from human curated datasets (for further details, see Section 4), consequently avoiding errors that may occur in gene name or chemical compound name identification.

3 Method

3.1 Notation

We define notation that is used throughout this paper. We denote a variable by a capitalized letter, e.g. U, and its value as the same letter in lower case, e.g. u. To explain a particular model for the co-occurrence of a gene and a compound, we define the following symbols in particular. Let G be an observable random variable taking on values $g_1, ..., g_S$, each of which corresponds to a gene. Similarly let C be an observable random variable taking on $c_1, ..., c_T$, each of which corresponds to a chemical compound. Let Z be a discrete-valued latent variable taking on values $z_1, ..., z_H$, each of which corresponds to a latent cluster, where H is the number of clusters. Let θ be a set of parameters for the model to be optimized in the learning process, and let π be a mixture parameter (i.e. weight) of a component of our model that the users can specify. Let D be a set of all examples.

3.2 Mixture Aspect Model (MAM)

We begin by describing the *aspect model* (AM) for two-mode and co-occurrence data[3]. With latent clusters z_h $(h = 1, \cdots, H)$, AM gives the log-likelihood for a co-occurrence of (u, v) in the following form:

$$\log p(u, v; \theta) = \log \sum_h p(u|z_h; \theta) p(v|z_h; \theta) p(z_h; \theta).$$

So the log-likelihood for D by this model is given as follows:

$$\log p(D; \theta) = \sum_{i,j} N_{i,j} \log p(u_i, v_j; \theta),$$

where $N_{i,j}$ is the number of co-occurrences of (u_i, v_j).

The purpose of this paper is to handle multiple different types of co-occurrence data with overlapping variable. More concretely, we can assume that we have two datasets, in which one has two random variables U and V, and the other has V and W. For these two datasets, we now define a new probabilistic model that is a mixture of two AMs, which we call *two-components mixture aspect model* (2MAM). The log-likelihood for D with two datasets for this model is given as follows:

$$\begin{aligned} & \log p(D; \theta) \\ = \; & \pi_{UV} \sum_{i,j} \frac{N_{i,j}}{N_{UV}} \log \sum_h p(u_i|z_h; \theta) p(v_j|z_h; \theta) p(z_h; \theta) \\ + \; & \pi_{VW} \sum_{j,k} \frac{M_{j,k}}{N_{VW}} \log \sum_h p(v_j|z_h; \theta) p(w_k|z_h; \theta) p(z_h; \theta), \end{aligned}$$

where $\pi_{UV} + \pi_{VW} = 1$ for U and V, $N_{i,j}$ and $M_{j,k}$ are the number of co-occurrences of (u_i, v_j) and (v_j, w_k), respectively, $N_{UV} = \sum_{i,j} N_{i,j}$ for U and V, and $N_{VW} = \sum_{j,k} M_{j,k}$ for V and W.

We note that both the first and second terms in this equation use the same probability parameter $p(v|z; \theta)$. Because of this, the parameter must be controlled by both datasets. We can easily see that this mixture model for two datasets can be extended to a mixture model for an arbitrary number of datasets. In this paper we focus on making use of all three types of co-occurrence data to predict co-occurrence of compound and gene. The detail of this model can be referred in [6].

3.3 Mixture Aspect Model for Predicting Co-occurrences of Compound-Gene

We consider three types of co-occurrence data: compound-gene, gene-gene and compound-compound pairs. We present a probabilistic model for this data, which we call *three-components mixture aspect model* (3MAM). The log-likelihood for all data D can be given by 3MAM as follows:

$$\begin{aligned} & \log p(D; \theta) \\ = \; & \pi_{CG} \sum_{i,j} \frac{N_{i,j}}{N_{CG}} \log \sum_h p(c_i|z_h; \theta) p(g_j|z_h; \theta) p(z_h; \theta) \\ + \; & \pi_{GG} \sum_{j,j'} \frac{M_{j,j'}}{N_{GG}} \log \sum_h p(g_j|z_h; \theta) p(g_{j'}|z_h; \theta) p(z_h; \theta) \\ + \; & \pi_{CC} \sum_{i,i'} \frac{L_{i,i'}}{N_{CC}} \log \sum_h p(c_i|z_h; \theta) p(c_{i'}|z_h; \theta) p(z_h; \theta). \end{aligned}$$

In the above equation, $\pi_{CG} + \pi_{GG} + \pi_{CC} = 1$, $N_{CC} = \sum_{i,i'} L_{i,i'}$ and $L_{i,i'}$ is the number of $(c_i, c_{i'})$ pairs.

3.4 Estimating Probability Parameters

Given training data D and the number of clusters H, a popular criterion for estimating the probabilities of a probabilistic model is the maximum likelihood (ML). Parameters are estimated to maximize the log-likelihood of data D:

$$\theta^{ML} = arg\max_{\theta} \log p(D;\theta).$$

The most popular approach for obtaining an ML estimator of a probabilistic model is a time-efficient general scheme called the EM (Expectation-Maximization) algorithm[4] that provides a local maximum. In general, the EM algorithm starts with a random set of initial parameter values and iterates both the expectation step (E-step) and the maximization step (M-step) alternately until a certain convergence criterion is satisfied.

We show the case in which we use all the three types of co-occurrence data such as compound-gene, gene-gene and compound-compound pairs. We use 3MAM, and so the log-likelihood for these datasets is also given in Section 3.3. The E- and M-steps for 3MAM can be given as follows:

E-step:

$$\begin{aligned}
p(z_h|c_i,g_j;\theta) &= \frac{p(c_i|z_h;\theta)p(g_j|z_h;\theta)p(z_h;\theta)}{\sum_{h'} p(c_i|z_{h'};\theta)p(g_j|z_{h'};\theta)p(z_{h'};\theta)} \\
p(z_h|g_j,g_{j'};\theta) &= \frac{p(g_j|z_h;\theta)p(g_{j'}|z_h;\theta)p(z_h;\theta)}{\sum_{h'} p(g_j|z_{h'};\theta)p(g_{j'}|z_{h'};\theta)p(z_{h'};\theta)} \\
p(z_h|c_i,c_{i'};\theta) &= \frac{p(c_i|z_h;\theta)p(c_{i'}|z_h;\theta)p(z_h;\theta)}{\sum_{h'} p(c_i|z_{h'};\theta)p(c_{i'}|z_{h'};\theta)p(z_{h'};\theta)}
\end{aligned}$$

M-step:

$$\begin{aligned}
\theta_{c_i|z_h} &\propto \pi_{CG}\sum_j \frac{N_{i,j}}{N_{CG}} p(z_h|c_i,g_j;\theta_{old}) + \pi_{CC}\sum_{i'} \frac{L_{i,i'}}{N_{CC}} p(z_h|c_i,c_{i'};\theta_{old}) \\
\theta_{g_j|z_h} &\propto \pi_{CG}\sum_i \frac{N_{i,j}}{N_{CG}} p(z_h|c_i,g_j;\theta_{old}) + \pi_{GG}\sum_{j'} \frac{M_{j,j'}}{N_{GG}} p(z_h|g_j,g_{j'};\theta_{old}) \\
\theta_{z_c} &\propto \pi_{CG}\sum_{i,j} \frac{N_{i,j}}{N_{CG}} p(z_h|c_i,g_j;\theta_{old}) + \pi_{GG}\sum_{j',j''} \frac{M_{j',j''}}{N_{GG}} p(z_h|g_{j'},g_{j''};\theta_{old}) \\
&+ \pi_{CC}\sum_{i',i''} \frac{L_{i',i''}}{N_{CC}} p(z_h|c_{i'},c_{i''};\theta_{old})
\end{aligned}$$

4 Experimental Results

4.1 Data

MAM can incorporate any type of co-occurrence data, but as mentioned in Section 1, in this paper we focus on existing literature only. We derived our datasets from all records that have been stored in MEDLINE and were published from 1960 to 2004.

We first used the 'Locus ID'[7] to check if a gene is in an abstract, by using a list of links which is available at ftp://ftp.ncbi.nih.gov/refseq/LocusLink. Each link connects a Locus ID with a PubMed ID. We focused on "human" genes only and selected MEDLINE records containing one or more human genes using this list. We then generated co-occurrence data on genes from the selected MEDLINE records. In order to produce meaningful gene-gene co-occurrence pairs, we skipped the MEDLINE records, each of which has more than 103 genes. This is because some MEDLINE records report all genes in the microarray experiment[1].

We then used the CAS Registry numbers as defined in the records of MEDLINE to find a chemical compound in a document. Using the selected MEDLINE records, we generated co-occurrence data on compound pairs. (For the details of the CAS Registry numbers, see www.cas.org/EO/regsys.html.)

[1]For example, MEDLINE ID 12477932 has more than 9,000 human genes.

Table 1: The size of co-occurrence datasets.

Item	Size
MEDLINE records	63,940
gene type	22,292
gene-gene	174,077
chemical compound type	3,454
compound-compound	20,443
compound-gene	47,217

We finally generated co-occurrence data on compound-gene pairs from the selected MEDLINE records using both the CAS registry numbers and the above link list of genes. We used this list on Locus IDs because it is a curated list and is probably the most reliable data source, to the best of our knowledge. Table 1 shows the sizes of the three co-occurrence datasets. We note that "MEDLINE records" in the table is the number of MEDLINE records that we used to derive our three types of datasets.

4.2 Performance Evaluation by Cross-Validation

4.2.1 Evaluation Procedure

We evaluated the performance of MAM using cross-validation on predicting compound-gene (and additionally, compound-compound) pairs.

We tested four types of models to predict compound-gene pairs. That is, we first tested AM using the co-occurrence data of compound-gene pairs only, and then tested two different 2MAM by adding compound-compound (2MAM (CG+CC)) and gene-gene (2MAM (CG+GG)) pairs. Finally, we made use of all three types of co-occurrence data to train 3MAM.

To examine the effect of the size of the training data set to the performance of the probabilistic model, we set five different ratios of the size of training to test data, 3:1, 2:1, 1:1, 1:2 and 1:3, in the cross-validation experiment. For example, in the 3:1 case, we randomly divide the original compound-gene data into 4 subsets of roughly equal size, and then alternatively select one subset as the test data and the other three subsets as training data. We carried out 50 rounds of this cross-validation to reduce possible biases occurring in only a few rounds and averaged the results obtained. When we add another type of training data, keeping the same training compound-gene pairs for each round of cross-validation, we added one or more other types of co-occurrence data to train 2MAM or 3MAM. Then, the prediction was performed on the same test dataset.

We note that AM cannot make any predictions on a compound-gene pair in the test data if one component of this pair does not appear in the training data. Thus, we removed all such co-occurrence pairs in the test data, and the remaining pairs were used as positive test examples[2]. We then randomly generated the same number of compound-gene pairs which are not found in both training and test as negative test examples. We checked the performance of each of the models tested by the ability to discriminate positive from negative test examples.

4.2.2 Evaluation Measures

Once we estimated the probability parameters of a probabilistic model from training data, we computed the likelihood of each compound-gene pair in test data and ranked all pairs according to their likelihoods. We evaluated these ranked pairs in two ways: AUC (Area Under the ROC curve) and recall.

An ROC (Recciver Operator Characteristic) curve is drawn by plotting "sensitivity" against "false positive rate", using the ranked compound-gene pairs. The sensitivity (or true positive rate) is the proportion of the number of correctly predicted positive examples to the total number of positive examples. The false positive rate is the proportion of the number of false positive examples to the total number of negative examples. The AUC, a popular metric for measuring the performance of

[2]We emphasize that this experimental setting is advantageous to AM and not to MAM.

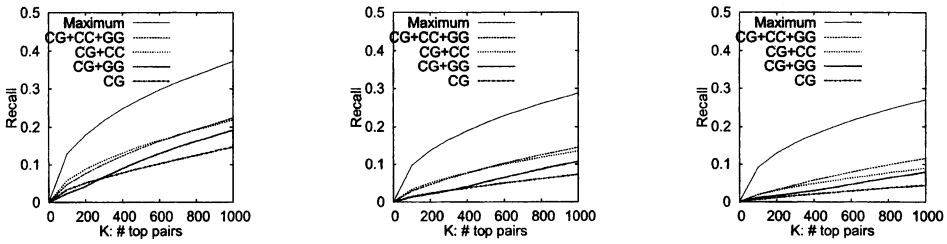

Figure 1: Recalls of the compound-gene experiments at the three ratios of training to test examples: (a) 3:1, (b) 1:1 and (c) 1:3.

different models[8], can be computed as the area under this ROC curve. We can see that the larger the AUC, the better the performance of the model. We further used the paired sample two-tailed t-test to statistically evaluate the performance difference of the two models. Since we run cross-validation 50 times, we have at least 100 values in each of the five different ratios, and so if the t-value is greater than 3.50 then the difference is more than 99.9% statistically significant in any ratio.

The recall used is slightly different from recall as used in the field of information retrieval, because we consider a positive example (i.e. co-occurrence pair) to have one or more co-occurrences. (Note that the number of co-occurrences of a negative example is zero.) When we choose the top K examples from the ranked pairs, our recall is defined as follows:

$$\text{Recall} = \frac{\sum_i^K O_i}{\sum_i^N O_i},$$

where O_i is the number of co-occurrences of the i-th ranked pair, and N is the number of all examples. We note that the maximum recall value occurs when K is given as follows:

$$\text{Maximum recall} = \frac{\sum_i^K O_i^{sort}}{\sum_i^N O_i},$$

where O_i^{sort} is the number of co-occurrences of the i-th ranked pairs in a list of pairs sorted according to the number of co-occurrences. When we choose the top K examples, if all of them are negatives, the recall is zero. So this recall takes a value in the range from zero to the above maximum. A unique property of this recall is that even if all chosen top K examples are positives, the recall can be smaller than the above maximum, unless their likelihoods are the K largest likelihoods.

4.2.3 Parameter Settings

The stopping condition we adopted for our EM estimation was when the improvement of the observed log-likelihood between two successive EM iteration is less than 0.001. We used a uniform distribution for the weights (i.e. π) of both 2MAM and 3MAM in all cases. As mentioned in Section 1, MAM is a space-efficient model, but we note that our datasets require a huge memory space. As we set $H = 128$, there were altogether around 3,250,000 ($\fallingdotseq (22,292 + 3,454) \times 128$ for $p(g|z)$ and $p(c|z)$) parameters to be estimated for 3MAM. It costed around 800 MB of memory and took around 30 minutes to execute on a Linux workstation with dual Intel Xeon 3.0 GHz processors and 8 GB of main memory.

4.2.4 Results on Compound-Gene Pairs

Table 2 shows the AUC for each model at different data settings and the t-values between the AUC of 3MAM and that of another model. This table clearly showed that 3MAM outperformed the other three models and its performance was followed by 2MAM (CG+CC), 2MAM (CG+GG), and AM. We note that the difference between 3MAM and AM reached around 7 to 10%. This performance improvement is significant, because the AUC of AM reached 82 to 89% already, and so it is usually hard to improve these values. Furthermore, the t-values showed that 3MAM outperformed all other models by a statistically significant factor in all cases. These results indicate that incorporating compound-compound and gene-gene pairs improved the predictive performance obtained by compound-gene pairs

Table 2: Percentage of the AUCs and the t-values (in parentheses) obtained by 50 rounds of cross-validation on compound-gene pairs.

Model	Ratio of training to test data				
	3:1	2:1	1:1	1:2	1:3
3MAM (CG+CC+GG)	**96.0**	**95.5**	**94.5**	**92.8**	**91.5**
2MAM (CG+CC)	95.0 (**81.4**)	94.5 (**73.9**)	93.2 (**60.3**)	91.1 (**88.6**)	89.6 (**94.9**)
2MAM (CG+GG)	92.3 (**193.8**)	91.6 (**168.0**)	89.8 (**158.6**)	87.7 (**209.2**)	86.4 (**197.4**)
AM (CG)	89.0 (**232.2**)	88.0 (**202.4**)	86.0 (**190.5**)	83.6 (**285.5**)	82.0 (**357.4**)

Table 3: Percentage of the AUCs and the t-values (in parentheses) obtained by 50 rounds of cross-validation on compound-compound pairs.

Model	Ratio of training to test data				
	3:1	2:1	1:1	1:2	1:3
3MAM (CC+CG+GG)	**96.6**	**96.2**	**95.1**	**93.1**	**91.7**
2MAM (CC+CG)	96.4 (**13.3**)	95.9 (**17.1**)	94.7 (**17.8**)	92.8 (**19.0**)	91.5 (**14.4**)
AM (CC)	95.3 (**87.1**)	94.4 (**96.0**)	92.2 (**140.5**)	88.8 (**194.7**)	86.5 (**219.7**)

only. Another empirical finding from these results is that incorporating compound-compound pairs was more effective in improving the predictive performance than incorporating gene-gene pairs, even though the size of the compound-compound pairs is smaller than (in fact, less than a quarter of) that of gene-gene pairs.

Figure 1 shows the recalls obtained by changing the top K examples selected. (We show only the three ratios due to the space limitations.) This figure also supports all the findings obtained in Table 2. Another finding emphasized in this figure is that the performance improvement from incorporating different types of data was pronounced, especially in the situation of less training data (e.g. the 1:3 case).

4.2.5 Results on Compound-Compound Pairs

We also carried out an experiment on predicting compound-compound pairs using 50 rounds of cross-validation, following the experimental procedure done for predicting compound-gene pairs. That is, AM was trained with compound-compound pairs only, 2MAM was trained with both compound-compound and compound-gene pairs, and gene-gene pairs were further added for training 3MAM. Table 3 summarizes the AUC values obtained by this experiment. From this table, we can see that 3MAM outperformed the other two models again with the difference being statistically significant, but the improvement in AUC was slightly smaller than that of compound-gene pairs.

Figure 2 shows the recalls obtained by changing the top K examples selected (for only the three ratios again due to the space limitations). Interestingly, for compound-compound pairs, the recall values in this figure showed the advantage of 3MAM over the other two models more clearly than the AUC in Table 2. These results confirmed again the effectiveness of incorporating other types of data, especially in the absence of enough training data.

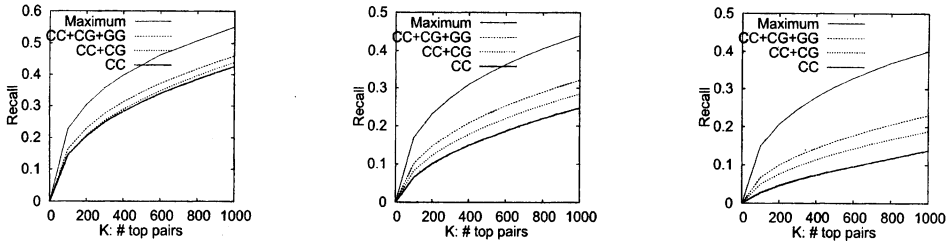

Figure 2: Recalls of the compound-compound experiments at the three ratios of training to test examples: (a) 3:1, (b) 1:1 and (c) 1:3.

5 Concluding Remarks

We have proposed a probabilistic model, composed of a mixture of aspect models, each of which is for one type of co-occurrence data, coupled with its learning algorithm. Our model can combine a number of different types of co-occurrence data efficiently, and in fact our experimental results have shown that incorporating different type of datasets improved the predictive performance drastically.

Acknowledgments

This work is supported in part by Bioinformatics Education Program "Education and Research Organization for Genome Information Science" and Kyoto University 21st Century COE Program "Knowledge Information Infrastructure for Genome Science" with support from MEXT (Ministry of Education, Culture, Sports, Science and Technology), Japan.

References

[1] J. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.*, 28:21–28, 2001.

[2] J.T. Chang and R.B. Altman. Extracting and characterizing gene-drug realtionships from the literature. *Pharmacogenetics*, 14:577–586, 2004.

[3] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196, 2001.

[4] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38, 1977.

[5] D. Wheeler et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, 33:D39–D45, 2005.

[6] S. Zhu, Y. Okuno, G. Tsujimoto, and H. Mamitsuka. A probabilistic model for mining implicit chemical compound - gene relations from literature. In *Proc. of ECCB2005*, 2005.

[7] K. Pruitt and D. Maglott. Refseq and locuslink: Ncbi gene-centered resources. *Nucleic Acids Res.*, 29:137–140, 2001.

[8] A. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159, 1997.