BIOINFORMATICS

Vol. 21 Suppl. 2 2005, pages ii245–ii251
doi:10.1093/bioinformatics/bti1141

Text Mining

A probabilistic model for mining implicit 'chemical compound–gene' relations from literature

Shanfeng Zhu[1], Yasushi Okuno[2], Gozoh Tsujimoto[2] and Hiroshi Mamitsuka[1,*]

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji 611-0011, Japan and
[2]Graduate School of Pharmaceutical Sciences, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan

ABSTRACT

Motivation: The importance of chemical compounds has been emphasized more in molecular biology, and 'chemical genomics' has attracted a great deal of attention in recent years. Thus an important issue in current molecular biology is to identify biological-related chemical compounds (more specifically, drugs) and genes. Co-occurrence of biological entities in the literature is a simple, comprehensive and popular technique to find the association of these entities. Our focus is to mine implicit 'chemical compound and gene' relations from the co-occurrence in the literature.

Results: 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 examined the performance of our approach not only by a cross-validation using the data generated from the MEDLINE records but also by a test using an independent human-curated dataset of the relationships between chemical compounds and genes in the ChEBI database. We performed experimentation on three different types of co-occurrence datasets (i.e. compound–gene, gene–gene and compound–compound co-occurrences) in both cases. Experimental results have shown that MAM trained by all datasets outperformed any simple model trained by other combinations of datasets with the difference being statistically significant in all cases. In particular, we found that incorporating compound–compound co-occurrences is the most effective in improving the predictive performance. We finally computed the likelihoods of all unknown compound–gene (more specifically, drug–gene) pairs using our approach and selected the top 20 pairs according to the likelihoods. We validated them from biological, medical and pharmaceutical viewpoints.

Contact: mami@kuicr.kyoto-u.ac.jp

1 INTRODUCTION

Traditional molecular biology tells us that genetic information is transferred from DNA to protein and ultimately shows up as protein functions. The final goal of molecular biology in this 'central dogma' is to identify and understand biological activities regulated by proteins so that they can be managed. The most important protein function is to catalyze biochemical reactions for the synthesis of one chemical compound from another. Thus the first step of the above goal may be compared with detecting one or more chemical compounds for which each protein can catalyze.

Recently the importance of chemical compounds has been emphasized more in molecular biology, and a new research field, called 'chemical genomics', has attracted a great deal of attention. In fact, one of the five items to be taken up by the National Institute of Health (NIH) roadmap initiative is a chemoinformatics project for building small molecular libraries. This chemoinformatics project will develop a new compound database of chemical structures and their biological activities, with the idea of promoting pharmaceutical research, such as discovering new drugs. This database, called PubChem, will house compound information on the screening and probe data newly obtained by the Molecular Libraries Screening Centers Network (MLSCN) as well as those from the current scientific literature. A related fact is that databases of chemical compounds and their biochemical reactions have also been developed in recent years. For example, the KEGG (Kyoto Encyclopedia of Genes and Genomes) database (Kanehisa *et al*., 2004), which is a database of metabolic pathways generated by gathering biochemical reactions, has drastically grown both in the size of stored reactions and the number of citations in the biological and medical sciences. European Bioinformatics Institute (EBI) also developed a freely available database of small molecular entities, ChEBI (Brooksbank *et al*., 2005), which stands for 'Chemical Entities of Biological Interest'.

Thus an important issue in current molecular biology is to identify biological-related chemical compounds and genes, which is a fundamental step of chemical genomics research. Mining biomedical and biological literature databases, such as Medline (Wheeler *et al*., 2005), for identifying these kinds of biological-related entities has been actively tackled in the last few years (Blasckel *et al*., 2002; Yandell *et al*., 2002). Co-occurrence of biological entities in the literature is a simple, comprehensive and popular technique to identify the association of these entities (Stapley *et al*., 2000; Jenssen *et al*., 2001; Chang *et al*., 2004). This technique is based on the following hypothesis: if a biological entity appears with another biological entity in the same document, these two entities should be biologically related with high probability. This hypothesis was already experimentally testified by many researchers (Jenssen *et al*., 2001; Chang *et al*., 2004). We will describe the details of this and related issues in Section 2.

Thus we focus on the co-occurrence information in the literature to discover implicit 'chemical compound–gene' relations, being those which are not in existing compound–gene co-occurrences in the literature but could be discovered from the co-occurrence data. All possible combinations of compounds and genes are very large in number, but obviously known co-occurrences of them are very limited, even though the literature is very abundant in size. Thus we attempt to use not only the co-occurrence of a chemical compound

*To whom correspondence should be addressed.

and a gene but also various other types of co-occurrences data, such as gene–gene and compound–compound 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 (Hofmann, 2001), 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. For example, we can have three different components corresponding to three different co-occurrences of compound–gene, gene–gene and compound–compound. These three datasets might be handled by AM by regarding all three as only one type of co-occurrence data. However MAM has roughly two significant advantages, compared with this way of using AM. First, it has a weight for each component, so that the users can control the weight for each co-occurrence dataset. Obviously, AM cannot do this. The second advantage is both time and space efficiency. When we have T datasets all having co-occurrences of N events, MAM considers $T \cdot N^2$ combinations at most, whereas AM must consider a maximum of $T^2 \cdot N^2$. In practice, N (i.e. the number of compounds or genes) reaches at least a few thousands, so that even if T is a relatively small number, this difference would be pronounced.

Our algorithm for estimating the probability parameters of MAM is based on the EM (Expectation–Maximization) algorithm (Dempster *et al*., 1977) 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. MAM can find new biological-related compound–gene pairs that have not yet been found in current biology and medical literatures.

In our experiments, we generated three types of co-occurrence datasets: gene–gene, compound–compound and compound–gene from the Medline records (Wheeler *et al*., 2005). We evaluated our method by not only these datasets, but also an independent (human-curated) dataset of chemical compound and gene relationships in the ChEBI database. 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 the experiment on predicting the biological-related compounds and genes in the ChEBI database, 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. Finally, we computed the likelihood of each of all unknown compound–gene (more precisely, drug–gene) pairs and selected the top 20 of them according to the likelihoods. We thus showed a list of them that have the highest likelihoods given by MAM trained by all given datasets and examined the validity of these pairs from biological, medical and pharmaceutical viewpoints.

2 RELATED WORK

Mining the Medline text 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, effective and popular technique to 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) presented a gene-to-gene co-occurrence network called PubGene using over 10 million Medline records. They randomly selected 500 pairs of genes that co-occurred once and 500 pairs of genes that 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 ∼60% for the first group, and 72% for the second. In further analysis, they found that almost all errors were owing 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 another 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 the failures are because of 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.

Some studies have combined co-occurrence methods with natural language processing techniques, such as shallow parsing, full parsing and constructing templates (Yandell *et al*., 2002; Blasckel *et al*., 2002). Their goal is to extract and clarify the detailed relationships among biological entities, such as protein–protein interaction (Blasckel *et al*., 1999), protein–drug interaction (Rindflesch *et al*., 2000) and gene–mutation pairs (Rebholz-Schuhmann *et al*., 2004).

Some researchers have, however, attempted to find implicit relationships between biological entities of not having direct co-occurrences in the literature (Wiren *et al*., 2004). For example, Perez-Iratxeta *et al*. (2002) used the fuzzy set theory to analyze the relationships between the co-occurrence of MeSH terms in different categories and the co-occurrence of a MeSH term and a GO (Gene Ontology) term in Medline records and scored the implicit associations between symptoms of diseases and GO terms by fuzzy relations.

In contrast to these existing approaches, our focus is placed on implicit ‘compound–gene’ relations in the literature, and our approach is based on statistical learning using a probabilistic model that is an extension of the so-called AM (Hofmann, 2001). This AM has already proved effective in a lot of applications for analyzing co-occurrence data, such as informational retrieval, computational linguistics and collaborative filtering (Hofmann, 2004; Si *et al*., 2003). We emphasize that our statistical learning based approach develops a noise-robust probabilistic model and a systematic and

efficient algorithm for estimating the parameters of our model from different types of multiple co-occurrence data.

3 METHODS

3.1 Notations

We define the notations that are 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 $g_1, \ldots, g_S$, each of which corresponds to a gene. Similarly, let C be an observable random variable taking on $c_1, \ldots, c_T$, each of which corresponds to a chemical compound. Let Z be a discrete-valued latent variable taking on values $z_1, \ldots, z_H$, 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 AM for two-mode and co-occurrence data (Hofmann, 2001). With latent clusters z_h $(h = 1, \ldots, H)$, 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-component mixture aspect model (2MAM). The log-likelihood for D with two datasets for this model is given as follows:

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

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 $p(v|z; \theta)$. Therefore, 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. We note that if each of these datasets has a random variable that appears in more than one dataset, this model is different from AM. This is particularly true when estimating its parameters, each of which corresponds to a variable appearing more than once. These parameters must be trained (controlled) by more than one dataset. The detailed algorithm for estimating this type of parameters is described for a particular case of the co-occurrence of a chemical compound and a gene in Section 3.4.

3.3 Mixture aspect model for predicting co-occurrences of compound–gene

When there is only one type of co-occurrence data (i.e. compound–gene pairs), this dataset can be handled by AM. If another dataset like gene–gene pairs is added to this dataset, these two datasets can be handled by 2MAM. For example, if we have two types of co-occurrence data, such as compound–gene and gene–gene pairs, the log-likelihood for all the data D by 2MAM is written as follows:

$$\log p(D; \theta) = \pi_{\mathrm{CG}} \sum_{i,j} \frac{N_{i,j}}{N_{\mathrm{CG}}} \log \sum_h p(c_i|z_h; \theta) p(g_j|z_h; \theta) p(z_h; \theta) + \pi_{\mathrm{GG}} \sum_{j,j'} \frac{M_{j,j'}}{N_{\mathrm{GG}}} \log \sum_h p(g_j|z_h; \theta) p(g_{j'}|z_h; \theta) p(z_h; \theta),$$

where $N_{\mathrm{CG}} = \sum_{i,j} N_{i,j}$ and $N_{i,j}$ is the number of co-occurrences of (c_i, g_j), and $N_{\mathrm{GG}} = \sum_{j,j'} M_{j,j'}$ and $M_{j,j'}$ is the number of co-occurrences of $(g_j, g_{j'})$.

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

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

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

In this paper, even though we used three types of data, it is evident that 3MAM can incorporate another type of co-occurrence data if it can improve the predictive performance of 3MAM.

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:

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

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 (Dempster *et al.*, 1977) 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.

3.4.1 AM We begin to explain the EM algorithm for AM for only one type of co-occurrence data, i.e. compound–gene pairs. The log-likelihood for D is given in Section 3.2, and the E- and M-steps can be given as follows:

E-step:

$$p(z_h|c_i, g_j; \theta) = \frac{p(c_i|z_h; \theta) p(g_j|z_h; \theta) p(z_h; \theta)}{\sum_{h'} p(c_i|z_{h'}; \theta) p(g_j|z_{h'}; \theta) p(z_{h'}; \theta)}.$$

M-step:

$$\theta_{c_i|z_h} \propto \sum_j N_{i,j}\, p(z_h|c_i, g_j; \theta_{\mathrm{old}}),$$

$$\theta_{g_j|z_h} \propto \sum_i N_{i,j}\, p(z_h|c_i, g_j; \theta_{\mathrm{old}}),$$

$$\theta_{z_h} \propto \sum_{i,j} N_{i,j}\, p(z_h|c_i, g_j; \theta_{\mathrm{old}}).$$

3.4.2 2MAM Next, we show the EM algorithm for the case in which we add another type of co-occurrence data, such as gene–gene pairs to compound–gene pairs. In this case we use 2MAM, and so the log-likelihood for these datasets is given in Section 3.3. The E- and M-steps for 2MAM can be given as follows:

E-step:

$$p(z_h|c_i,g_j;\theta)=\frac{p(c_i|z_h;\theta)p(g_j|z_h;\theta)p(z_h;\theta)}{\sum_{h'}p(c_i|z_{h'};\theta)p(g_j|z_{h'};\theta)p(z_{h'};\theta)},$$

$$p(z_h|g_j,g_{j'};\theta)=\frac{p(g_j|z_h;\theta)p(g_{j'}|z_h;\theta)p(z_h;\theta)}{\sum_{h'}p(g_j|z_{h'};\theta)p(g_{j'}|z_{h'};\theta)p(z_{h'};\theta)}.$$

M-step:

$$\theta_{c_i|z_h}\propto\sum_j N_{i,j}\,p(z_h|c_i,g_j;\theta_{\mathrm{old}}),$$

$$\theta_{g_j|z_h}\propto\pi_{\mathrm{CG}}\sum_i\frac{N_{i,j}}{N_{\mathrm{CG}}}p(z_h|c_i,g_j;\theta_{\mathrm{old}})+\pi_{\mathrm{GG}}\sum_{j'}\frac{M_{j,j'}}{N_{\mathrm{GG}}}p(z_h|g_j,g_{j'};\theta_{\mathrm{old}}),$$

$$\theta_{z_h}\propto\pi_{\mathrm{CG}}\sum_{i,j}\frac{N_{i,j}}{N_{\mathrm{CG}}}p(z_h|c_i,g_j;\theta_{\mathrm{old}})+\pi_{\mathrm{GG}}\sum_{j',j''}\frac{M_{j',j''}}{N_{\mathrm{GG}}}p(z_h|g_{j'},g_{j''};\theta_{\mathrm{old}}).$$

3.4.3 3MAM Finally, 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:

$$p(z_h|c_i,g_j;\theta)=\frac{p(c_i|z_h;\theta)p(g_j|z_h;\theta)p(z_h;\theta)}{\sum_{h'}p(c_i|z_{h'};\theta)p(g_j|z_{h'};\theta)p(z_{h'};\theta)},$$

$$p(z_h|g_j,g_{j'};\theta)=\frac{p(g_j|z_h;\theta)p(g_{j'}|z_h;\theta)p(z_h;\theta)}{\sum_{h'}p(g_j|z_{h'};\theta)p(g_{j'}|z_{h'};\theta)p(z_{h'};\theta)},$$

$$p(z_h|c_i,c_{i'};\theta)=\frac{p(c_i|z_h;\theta)p(c_{i'}|z_h;\theta)p(z_h;\theta)}{\sum_{h'}p(c_i|z_{h'};\theta)p(c_{i'}|z_{h'};\theta)p(z_{h'};\theta)}.$$

M-step:

$$\theta_{c_i|z_h}\propto\pi_{\mathrm{CG}}\sum_j\frac{N_{i,j}}{N_{\mathrm{CG}}}p(z_h|c_i,g_j;\theta_{\mathrm{old}})+\pi_{\mathrm{CC}}\sum_{i'}\frac{L_{i,i'}}{N_{\mathrm{CC}}}p(z_h|c_i,c_{i'};\theta_{\mathrm{old}}),$$

$$\theta_{g_j|z_h}\propto\pi_{\mathrm{CG}}\sum_i\frac{N_{i,j}}{N_{\mathrm{CG}}}p(z_h|c_i,g_j;\theta_{\mathrm{old}})+\pi_{\mathrm{GG}}\sum_{j'}\frac{M_{j,j'}}{N_{\mathrm{GG}}}p(z_h|g_j,g_{j'};\theta_{\mathrm{old}}),$$

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

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	3454
Compound–compound	20 443
Compound–gene	47 217

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 the 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' (Pruitt and Maglott, 2001) 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 in December 2004. 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.) 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 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 dataset 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 four subsets

[1]For example, Medline ID 12477932 has more than 9000 human genes.

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**)

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 by AUC (area under the ROC curve).

An ROC (receiver 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 different models (Bradley, 1997), 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 >3.50 then the difference is $>99.9\%$ statistically significant in any ratio.

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 <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. When we set $H = 128$, there were altogether $\sim$3 250 000 [$\fallingdotseq (22\,292 + 3454) \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 $\sim$30 min to execute on a Linux workstation with dual Intel Xeon 3.0 GHz processors and 8 GB of main memory.

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

Table 3. Percentage of the AUCs and the *t*-values (in parentheses) obtained by predicting biological-related compound–gene pairs in ChEBI

Model	3MAM (CG + CC + GG)	2MAM (CG + CC)	2MAM (CG + GG)	AM (CG)
AUC (*t*-value)	**86.3**	86.1 (1.42)	82.3 (**21.6**)	80.1 (**30.9**)

4.2.4 Results 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 $\sim$7–10%. This performance improvement is significant, because the AUC of AM reached 82–89% 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 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 one-eighth of) that of gene–gene pairs.

4.3 Performance evaluation by ChEBI

4.3.1 Test data description We further examined the performance of our probabilistic models by using an independent dataset, i.e. we trained our models using our datasets generated from the Medline records, and the trained models were applied to another dataset.

We used a list of the relationships between chemical compounds and proteins in the ChEBI database to generate a test dataset. We first extracted 26 091 pairs of chemical compounds and human genes from this list by replacing a protein with one or more corresponding genes according to the UniProt database (Bairoch *et al*., 2005). We then removed a compound–gene pair from them if it appears in our training dataset or if one component of this pair does not appear in the

Table 4. Top 20 pairs of drugs and genes

CAS registry number	Drug name	Locus ID	Gene name	Log-likelihood
19545-26-7	Wortmannin	5594	*MAPK1*: Mitogen-activated protein kinase 1	−2.615
16561-29-8	Tetradecanoylphorbol acetate	5590	*PRKCZ*: Protein kinase C, zeta	−2.764
23214-92-8	Doxorubicin	1029	*CDKN2A*: Cyclin-dependent kinase inhibitor 2A	−2.992
73-22-3	Tryptophan	5705	*PSMC5*: Proteasome 26S subunit	−3.000
10102-43-9	Nitric Oxide	959	*TNFSF5*: Tumor necrosis factor, member 5	−3.027
66-81-9	Cycloheximide	5970	*RELA*: V-rel reticuloendotheliosis viral oncogene homolog A	−3.030
33419-42-0	Etoposide	4193	*MDM2*: Transformed 3T3 cell double minute 2	−3.033
50-02-2	Dexamethasone	3458	*IFNG*: Interferon, gamma	−3.037
15663-27-1	Cisplatin	581	*BAX*: BCL2-associated X protein	−3.060
521-18-6	Dihydrotestosterone	2099	*ESR1*: Estrogen receptor 1	−3.061
53-85-0	Dichlororibofuranosylbenzimidazole	2963	*GTF2F2*: General transcription factor IIF, polypeptide 2	−3.103
50-07-7	Mitomycin	7157	*TP53*: Tumor protein p53	−3.104
320-67-2	Azacitidine	6622	*SNCA*: Synuclein, alpha	−3.111
33069-62-4	Paclitaxel	581	*BAX*: BCL2-associated X protein	−3.148
133407-82-6	Leucine aldehyde	7124	*TNF*: Tumor necrosis factor, member 2	−3.203
10540-29-1	Tamoxifen	5241	*PGR*: Progesterone receptor	−3.208
7722-84-1	Hydrogen peroxide	596	*BCL2*: B-cell CLL/lymphoma 2	−3.213
67526-95-8	Thapsigargin	5580	*PRKCD*: Protein kinase C, delta	−3.215
59-14-3	Bromodeoxyuridine	1027	*CDKN1B*: Cyclin-dependent kinase inhibitor 1B	−3.221

compound–gene pairs in the training dataset of AM for the benefit of AM. The number of final compound–gene pairs was 11 743, and we used them as positive examples in the test dataset. We generated the same number of negative examples, which are not in the training dataset and the positive test dataset.

4.3.2 Evaluation procedure We trained four different models, AM (CG), 2MAM (CG + GG), 2MAM (CG + CC) and 3MAM (CG + CC + GG) using the datasets generated from the Medline records and then predictions were performed on the test dataset generated in the above manner. We randomly generated the negative test dataset 50 times and the results were averaged over the 50 runs to reduce a possible bias. The evaluation measures and parameter settings were the same as those of the cross-validation experiment.

4.3.3 Results Table 3 shows the average AUC for each model and the *t*-values between the AUC of 3MAM and that of another model. This table clearly showed that 3MAM outperformed three other models, and was followed by 2MAM (CG + CC), 2MAM (CG + GG) and AM. This result indicates that our model is extremely effective to improve the predictive performance for an independent dataset as well. The improvement over 2MAM (CG + GG) and AM were especially significant, but the advantage over 2MAM (CG + CC) was slight (statistically insignificant). This result confirms the empirical finding that incorporating compound–compound pairs was more effective in improving the predictive performance than incorporating gene–gene pairs.

4.4 Mining and analyzing unknown drug-gene relationships

We trained 3MAM by using all datasets, i.e. all gene–gene, compound–compound and compound–gene pairs, and computed the likelihood of each of all possible compound–gene pairs that are not in the training dataset. We repeated this run 50 times and computed the sum of all 50 likelihoods for each compound–gene pair. We sorted all the pairs according to their likelihoods and removed the pairs whose chemical compounds do not have pharmacological activity. Thus each chemical compound of the remaining pairs is a drug. From this list of pairs, we finally selected the top 20 pairs so that a drug did not appear again when we scanned the sorted pairs from the top to the bottom. Table 4 shows the list of these 20 pairs with their log-likelihoods.

These pairs are unknown pairs in the literature, but our method suggested that each of them has a strong relationship between a drug and a gene. In fact, we can see a biological relationship for each pair of a drug and a gene. Below, we briefly describe the biological, medical and pharmaceutical relationship on each pair of the list, for only the top five pairs owing to the space limitations.

Wortmannin, of the first pair, is an inhibitor for the phosphatidylinositide kinase (PI3-kinase) pathway. Substrates of PI kinases are important signaling molecules that affect a wide range of biological processes (Zewail *et al*., 2003). In particular, recent studies have revealed that wortmannin affects proteasome-mediated degradation and chromatin remodeling. MAPK1 is also a popular kinase, and so wortmannin is expected to be an inhibitor of the MAPK pathway too.

The second pair of tetradecanoylphorbol acetate and protein kinase C seems to have more direct relation than that of the first pair. A pharmacological action of tetradecanolyphorbol acetate is a carcinogen. However, a number of processes important in certain diseases, such as solid tumors, are facilitated by the action of protein kinase C, and so inhibitors of protein kinase C have the potential to be anticancer drugs (Shih *et al*., 1999). Tetradecanolyphorbol acetate might become one of them.

Doxorubicin of the third pair is known as an antineoplastic agent, i.e. a drug intended to inhibit or prevent the maturation and proliferation of neoplasms that may become malignant, by targeting

the DNA. However, CDKN2A is an inhibitor of cyclic-dependent kinase (CDK) which is involved in replicative senescence, cell immortalization and tumor generation. Thus, these facts imply that the two molecules of this pair are strongly related.

A biological relation on the fourth pair of tryptophan and proteasome 26s subunit would be easily expected, because 26S proteasomes are mainly involved in the degradation of tryptophan hydroxylase (Kojima *et al*., 2002). The fifth pair would also have a clear biological relation. A member of the tumor necrosis factors can be a ligand for CD40, and CD40 ligation can stimulate nitric oxide production (Bingaman *et al*., 2000).

These facts indicate that unknown pairs that are predicted by our method to be related with each other have, in fact, strong biological relations. This analysis would be possible for most of the pairs in this list. Above all, the analysis on these pairs show that our probabilistic model and its learning algorithm can capture significant relationships between chemical compounds (more specifically, drugs) and genes.

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 types of datasets improved the predictive performance drastically.

In our experiments, we used a uniform distribution for the component weights (π) of our mixture model to allow users additional control. Interesting future work would adjust the weights to achieve the maximum predictive performance. It would also be interesting to investigate the possibility of incorporating totally different types of co-occurrence data (e.g. microarray co-expressions and protein–protein interactions) to the current literature data. Another appealing extension is to extract large scale co-occurrence data directly from Medline by incorporating latest natural language processing (NLP) techniques.

ACKNOWLEDGEMENTS

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.

Conflict of Interest: none declared.

REFERENCES

Bairoch,A. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.

Bingaman,A. *et al*. (2000) The role of CD40L in T cell-dependent nitric oxide production by murine macrophages. *Transpl. Immunol.*, **8**, 195–202.

Blaschke,C. *et al*. (1999) Automatic extraction of biological information from scientific text: protein–protein interactions. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **7**, 60–67.

Blaschke,C. *et al*. (2002) Information extraction in molecular biology. *Brief Bioinform.*, **3**, 154–165.

Bradley,A. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, **30**, 1145–1159.

Brooksbank,C. *et al*. (2005) The European Bioinformatics Institute's data resources: towards systems biology. *Nucleic Acids Res.*, **33**, D46–D53.

Chang,J.T. and Altman,R.B. (2004) Extracting and characterizing gene-drug relationships from the literature. *Pharmacogenetics*, **14**, 577–586.

Dempster,A. *et al*. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B*, **39**, 1–38.

Hofmann,T. (2001) Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, **42**, 177–196.

Hofmann,T. (2004) Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*, **22**, 89–115.

Jenssen,T. *et al*. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.*, **28**, 21–28.

Kanehisa,M. *et al*. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.

Kojima,M. *et al*. (2000) Rapid turnover of tryptophan hydroxylase is driven by proteasomes in RBL2H3 cells, a serotonin producing mast cell line. *J. Biochem. (Tokyo)*, **127**, 121–127.

Perez-Iratxeta,C. *et al*. (2002) Association of genes to genetically inherited diseases using data mining. *Nat. Genet.*, **31**, 316–319.

Pruitt,K. and Maglott,D. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.

Rebholz-Schuhmann,D. *et al*. (2004) Automatic extraction of mutations from Medline and cross-validation with OMIM. *Nucleic Acids Res.*, **32**, 135–142.

Rindflesch,T. *et al*. (2000) EDGAR: extraction of drugs, genes and relations from biomedical literature. *Pac. Symp. Biocomput.*, **5**, 517–528.

Shih,S.C. *et al*. (1999) Role of protein kinase C isoforms in phorbol ester-induced vascular endothelial growth factor expression in human glioblastoma cells. *J. Biol. Chem.*, **274**, 15407–15414.

Si,L. and Jin,R. (2003) Flexible mixture model for collaborative filtering. In *Proceedings of the 12th International Conference on Machine Learning*, August 21–24, 2003, Washington, DC, USA. AAAI Press, pp. 704–711.

Stapley,B. and Benoit,G. (2000) Biobibliometrics: information retrieval and visualization from co-occurrence of gene names in MEDLINE abstracts. *Pac. Symp. Biocomput.*, **5**, 529–540.

Wheeler,D. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33**, D39–D45.

Wren,J.D. *et al*. (2004) Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics*, **20**, 389–398.

Yandell,M.D. and Majoros,W.H. (2002) Genomics and natural language processing. *Nat. Rev. Genet.*, **3**, 601–610.

Zewail,A. *et al*. (2003) Novel functions of the phosphatidylinositol metabolic pathway discovered by a chemical genomics screen with wortmannin. *Proc. Natl Acad. Sci. USA*, **100**, 3345–3350.