

# Mining Implicit Biological Related Entities from Literature Using a Probabilistic Model

Shanfeng Zhu<sup>1</sup>

zhusf@kuicr.kyoto-u.ac.jp

Yasushi Okuno<sup>2</sup>

okuno@pharm.kyoto-u.ac.jp

Gozoh Tsujimoto<sup>2</sup>

gtsuji@pharm.kyoto-u.ac.jp

Hiroshi Mamitsuka<sup>1</sup>

mami@kuicr.kyoto-u.ac.jp

<sup>1</sup> Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan

<sup>2</sup> Graduate School of Pharmaceutical Science, Kyoto University, Japan

**Keywords:** text mining, probabilistic model

## 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. Here we also employ co-occurrence technique to identify biologically related genes and chemical compounds. We focus on discovering 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.

## 2 Materials and Methods

We made use of a probabilistic model, which we call a mixture aspect model (MAM), coupled with an efficient algorithm for estimating its parameters [3]. MAM is an extension of a probabilistic model, called the aspect model (AM) developed in natural language processing [1], with one significant difference of the ability of incorporating different types of co-occurrence data efficiently. Our algorithm for estimating the probability parameters of MAM is based on the EM (Expectation-Maximization) algorithm 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 on three types of co-occurrence data: gene-gene, compound-compound and compound-gene from the MEDLINE records [2]. We extract these data from RefSeq database and corresponding MEDLINE records. In our dataset, we have 22,292 genes and 3,454 chemical compounds. There are altogether 174,077 gene-gene pairs, 20,443 compound-compound pairs and 47,217 compound-gene pairs occurring in 63940 MEDLINE documents.

## 3 Experimental Results

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

Table 1: 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)	<b>96.0</b>	<b>95.5</b>	<b>94.5</b>	<b>92.8</b>	<b>91.5</b>
2MAM (CG+CC)	95.0 ( <b>81.4</b> )	94.5 ( <b>73.9</b> )	93.2 ( <b>60.3</b> )	91.1 ( <b>88.6</b> )	89.6 ( <b>94.9</b> )
2MAM (CG+GG)	92.3 ( <b>193.8</b> )	91.6 ( <b>168.0</b> )	89.8 ( <b>158.6</b> )	87.7 ( <b>209.2</b> )	86.4 ( <b>197.4</b> )
AM (CG)	89.0 ( <b>232.2</b> )	88.0 ( <b>202.4</b> )	86.0 ( <b>190.5</b> )	83.6 ( <b>285.5</b> )	82.0 ( <b>357.4</b> )

Table 2: 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)	<b>96.6</b>	<b>96.2</b>	<b>95.1</b>	<b>93.1</b>	<b>91.7</b>
2MAM (CC+CG)	96.4 ( <b>13.3</b> )	95.9 ( <b>17.1</b> )	94.7 ( <b>17.8</b> )	92.8 ( <b>19.0</b> )	91.5 ( <b>14.4</b> )
AM (CC)	95.3 ( <b>87.1</b> )	94.4 ( <b>96.0</b> )	92.2 ( <b>140.5</b> )	88.8 ( <b>194.7</b> )	86.5 ( <b>219.7</b> )

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. We then randomly generated the same number of compound-gene pairs which are not found in both training and test as negative test examples.

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 AUC (Area Under the 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. Experimental results in Table 1 and 2 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%.

## References

- [1] Hofmann, T., Unsupervised learning by probabilistic latent semantic analysis, *Machine Learning*, 42:177–196, 2001.
- [2] Wheeler, D. et al., Database resources of the National Center for Biotechnology Information, *Nucleic Acids Res.*, 33:D39–D45, 2005.
- [3] Zhu, S., Okuno, Y., Tsujimoto, G., and Mamitsuka, H., A probabilistic model for mining implicit Chemical Compound - Gene Relations from Literature, *Proc. of ECCB2005 (Bioinformatics 21 Supplement 2)*, ii245–ii251, 2005.