BMC Systems Biology

BioMed Central

Poster presentation

Open Access

Predicting implicit associated cancer genes from OMIM and MEDLINE by a new probabilistic model

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

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

Email: Shanfeng Zhu* - zhusf@kuicr.kyoto-u.ac.jp

* Corresponding author

from BioSysBio 2007: Systems Biology, Bioinformatics and Synthetic Biology
Manchester, UK. 11–13 January 2007

Published: 8 May 2007

BMC Systems Biology 2007, **1**(Suppl 1):P16 doi:10.1186/1752-0509-1-S1-P16

This abstract is available from: http://www.biomedcentral.com/1752-0509/1?issue=S1

Background

Discovering cancer associated genes can facilitate the understanding of tumour pathogenesis, the medical diagnoses and the treatment of patients. Here we mined OMIM and MEDLINE to discover implicitly associated cancer genes by applying a new probabilistic model, mixture aspect model (MAM) [1], on cancer gene co-occurrence data in OMIM and MEDLINE. Through cross-validation experiments, the accuracy of predicting associated cancer genes was shown to be improved by incorporating gene-gene co-occurrence pairs from MEDLINE into cancer-gene co-occurrence pairs in OMIM. Furthermore, some implicit associated cancer genes were predicted and analyzed preliminarily. The detailed result was presented on line http://www.bic.kyoto-u.ac.jp/pathway/zhusf/CancerInformatics/Supplemental2006.html for the reference of interested researchers and further validation by biologists.

Materials and methods

We extracted cancer-gene and cancer-cancer co-occurrence pairs from OMIM, a human curated knowledgebase on human genes and inherited diseases. A software tool CGMIM was used to extract the description section of OMIM to obtain cancers and associated genes [2]. This software maps genetic disorders into 21 different types of cancers. To avoid the difficulty of recognizing gene names, we extracted a human curated database, Entrez Gene, to obtain a subset of high quality MEDLINE records, where we obtained gene-gene co-occurrence data. MAM was proposed by us to mine implicit "chemical compound-gene" relations by integrating three types of co-occurrence data (compound-compound, gene-gene and compound-gene) in the literature [1]. The main advantage of MAM is the ability of integrating different type of co-occurrence data from heterogeneous data sources. MAM was first estimated by an EM algorithm to fit the existing co-occurrence data of cancer and gene, and then was used to predict the likelihood of the association of an unobserved pair of a cancer and a gene. See Table 1.

Table 1: The size of co-occurrence datasets

	Gene	Gene-Gene	Cancer	Cancer-Cancer	Cancer-Gene
Size	2,536	16,393	21	207	4,646

Table 2: AUCs and *t*-values obtained in the cross-validation experiment.

Model	The ratio of training to test data		
	3:1	1:1	1:3
3MAM(CG+CC+GG)	75.1	74.0	72.9
2MAM(CG+CC)	75.0(0.09)	73.7(**2.36**)	71.4(**15.4**)
2MAM(CG+GG)	72.4(**26.2**)	70.8(**23.7**)	68.4(**47.0**)
AM(CG)	73.3(**15.1**)	70.0(**23.5**)	64.7(**65.7**)

After training 3MAM with all three types of co-occurrence data, we computed the likelihood of all other cancer-gene pairs that are unknown in the OMIM. For each type of cancer, we present the top specific implicit gene in the Table 3. One interesting result is the top implicit associated gene specific to the prostate cancer, KLK10, which was already verified by Bharaj et al [3].

Table 3: For each type of cancer, we list the top specific implicit associated gene.

Cancer	Gene Name	Cancer	Gene Name	Cancer	Gene Name
BLADDER	IGFBP5	BRAIN	SYNJ2	BREAST	KLK8
CERVIX	PTGER1	COLORECTAL	RPS27	ESOPHAGUS	MAP3K10
KIDNEY	TFEC	LARYNX	MAP2K1	LEUKEMIA	IKZF3
LUNG	CHRNA7	LYMPHOMA	TRD@	MELANOMA	TSPAN7
MYELOMA	PRDM1	ORAL	TIAL1	OVARY	KLK7
PANCREAS	WIPI1	PROSTATE	KLK10	STOMACH	FUT6
TESTIS	PAGE1	THYROID	TBXA2R	BODY_OF_UTERUS	HOXC13

Results

We evaluated the performance of MAM by cross-validation on predicting associated cancer-gene pairs. In addition to training AM on cancer-gene pairs, we trained three other types of MAM by incorporating different type of co-occurrence data. 2MAM (CG+CC) and 2MAM (CG+GG) were built by adding cancer-cancer pairs and gene-gene pairs, respectively. In addition, 3MAM was built by incorporating all three types of co-occurrence data. To explore the effect of the size of the training data set on the performance of the probabilistic model, we set three different ratios of the size of training to test datasets, 3:1, 1:1 and 1:3, in the cross-validation experiment. The negative test examples were randomly generated and it was assured that no negative test example would appear in either training or positive test 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. After estimating the probability parameters of a probabilistic model from training data, we computed the likelihood of each cancer-gene pair in test data and ranked all pairs according to their likelihoods. Then it would be evaluated by AUC (Area under the ROC curve). The t-value was also computed to check the statistical significance of the different performance by two models. Here if the t-value is greater than 3.50 (2.36), the difference is more than 99.9% (98%) statistically significant. As illustrated in Table 2, 3MAM outperforms all other models, and is especially significant in the case of a small size of training data.

Conclusion

In this work, we mined OMIM database and MEDLINE to discover implicitly associated pairs of cancers and genes by applying a new probabilistic model, mixture aspect model (MAM), on the data of co-occurrence of cancers and genes, using OMIM and MEDLINE.

Acknowledgements

This work is partly supported by JSPS (Japan Society for the Promotion of Science) Postdoctoral Fellowship.

References

1. Zhu S, Okuno Y, Tsujimoto G, Mamitsuka H: **A probabilistic model for mining implicit 'chemical compound-gene' relations from literature.** *Bioinformatics* 2005, **21(Suppl 2):**ii245-ii251.
2. Bajdik CD, Kuo B, Rusaw S, Jones S, Brooks-Wilson A: **CGMIM: automated text-mining of Online Mendelian Inheritance in Man (OMIM) to identify genetically-associated cancers and candidate genes.** *BMC Bioinformatics* 2005, **6:**78-84.
3. Bharaj BB, Luo LY, Jung K, Stephen C, Diamandis EP: **Identification of single nucleotide polymorphisms in the human kallikrein 10 (KLK10) gene and their association with prostate, breast, testicular, and ovarian cancers.** *Prostate* 2002, **51(1):**35-41.
4. Zhu S, Okuno Y, Tsujimoto G, Mamitsuka H: **Application of a new probabilistic model for mining implicit associated cancer genes from OMIM and Medline.** *Cancer Informatics* 2006, **2:**361-371.