Integration of expression and textual data enhances the prediction of prognosis in breast cancer

Olivier Gevaert*, Steven Van Vooren, Bart De Moor

BioI@ESAT-SCD
Department of Electrical Engineering
Katholieke Universiteit Leuven
Kasteelpark Arenberg 10, Leuven, B-3001, Belgium

Integration of data sources has become very important in bioinformatics. This is evident from the numerous publications involving multiple data sources to discover new biological knowledge. This is due to the rise in publicly available databases and also the number of databases has increased significantly. Still many knowledge is contained in publications in unstructured from as opposed to being deposited in public databases where they can be amenable to use in algorithms. Therefore we attempted to mine this vast resource and transform it to the gene domain such that it can be used in combination with gene expression data. Microarray data are notorious for there low signal-to-noise ratio which causes that genes are often differently expressed between clinically relevant outcomes purely by chance. Integration of prior knowledge from literature abstracts can improve model building in general and gene selection in particular.

In this contribution we present an approach to integrate information from literature abstracts into probabilistic models of gene expression data. Integration of different data sources into a single framework potentially leads to more reliable models and at the same time it can reduce overfitting. Probabilistic models provide a natural solution to this problem since information can be incorporated in the prior distribution over the model space. This prior is then combined with other data to form a posterior distribution over the model space which is a balance between the information incorporated in the prior and the data.

Specifically, we investigated how the use of text information as a prior can improve the prediction of prognosis in breast cancer when modeling expression data using Bayesian networks. Bayesian networks provide a straightforward way to integrate information in the prior distribution over the possible structures of its network. By mining PUBMED abstracts we can easily represent genes as term vectors and create a gene-by-gene similarity matrix. After appropriate scaling, such a matrix can be used as a structure prior to build Bayesian networks. In this manner text information and gene expression data can be combined in a single framework. Our approach builds further on our methods for integrating prior information with Bayesian networks for other types of data (Antal et al, 2004; Gevaert et al, 2006b), where we have shown that structure prior information improves model selection especially when few data is available.

We investigated two applications of this framework. First, such a model can be used to predict the prognosis in cancer when a class variable describing different outcomes is included. The text prior can be easily extended to cover this class variable by using terms in the vocabulary that describe it. Secondly, this approach provides opportunities to improve the modeling of regulatory networks with Bayesian networks which has received much attention in the past few years. We focused on the first application and show promising results for the second application.

We used publicly available data to assess our framework for integration of textual and gene expression data (van ‘t Veer et al, 2002). This data set consists of breast cancer patients with binary outcomes: poor prognosis or good prognosis. Our goal was to predict the outcome reliably such that we can distinguish between these two groups. First, we assessed the influence of the text prior on the prediction of the outcome. We performed 100 randomizations of the data set with a uniform prior and 100 randomizations with the text prior. This gave a mean Area Under the ROC curve (AUC) of 0.75 for the uniform prior and a mean AUC of 0.80 for the text prior which was significantly different (P-value = 0.000396, two-sided Wilcoxon rank sum test). The text prior thus significantly enhances the prediction of the outcome. The text prior guides model search and favors genes which have a prior record related to prognosis. Apparently, this knowledge significantly improves gene selection and wards off genes which are differentially expressed by chance.

Next, we assessed the predictive performance of the whole network by using it to predict new data using the idea of blanket residuals (Sebastiani et al, 2004). This gave an average number of errors of 2667 (38%) for the text prior and an average number of errors of 2724 (39%) for the uniform prior which was statistically significant (P-value < 2e-10). We recognize that the improvement is small but the fact that the difference is significant means that by using gene-by-gene similarities from PUBMED abstracts we can improve our model of the genetic regulatory network that is related to the outcome of breast cancer.

Next, we evaluated the Markov blanket of the outcome (i.e. the variables which influence the outcome) for a model built with the text prior (TXTmodel) and for a model built with the uniform prior (UNImodel). The TXTmodel has many genes which have been implicated in breast cancer or cancer in general such as TP53, VEGF, MMP9, BIRC5, ADM, CA9 while ACADS, NEO1 and IHPK2 have a weaker link to cancer outcomes. MYLIP has no association. In the UNImodel far less genes are present which have a strong link with cancer outcomes. Only WISP1, FBXO31, IGFBP5 and TP53 have a relation with breast cancer outcome. The other genes have mostly unknown function or are not related. The text prior thus has its expected effect and includes genes which have a prior tendency to be associated with the prognosis of cancer. Moreover the UNImodel has more genes than the TXTmodel which indicates that the text prior selects genes in a more efficient way.

Finally this approach is complimentary to our previously published method to integrate clinical and microarray data with Bayesian networks (Gevaert et al, 2006a). Moreover other sources of information can be combined with the text prior. Possible sources of prior information are known protein-DNA interactions (e.g. Transfac, BIND), known pathways (e.g. KEGG or BIOCARTA) or motif information.

References:

Antal P, Fannes G, Timmerman D, Moreau Y, De Moor B. (2004) Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. *Artif Intell Med* **30**:257-81

Gevaert O, De Smet F, Timmerman D, Moreau Y, De Moor B (2006a) Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics* **15**:e184-90

Gevaert O, De Smet F, Kirk E, Van Calster B, Bourne T, Van Huffel S, Moreau Y, Timmerman D, De Moor B, Condous G (2006b) Predicting the outcome of pregnancies of unknown location: Bayesian networks with expert prior information compared to logistic regression. *Hum Reprod* **21**:1824-31

Sebastiani P, M Abad, MF Ramoni (2004) Bayesian networks for genomic analysis. In Genomic signal processing and statistics Dougherty ER, Shmulevich I, Chen J, Wang ZJ (eds) pp 281-320 New York: Hindawi Publishing Corporation

van ’t Veer,L., Dai,H., van de Vijver,M., He,U., Hart,A., Mao,M., Peterse,H., van der Kooy,K., Marton,M., Witteveen,A., Schreiber,G., Kerkhoven,R., Roberts,C., Linsley, P., Bernards,R. and Friend,S. (2002) Gene expression profiling predicts clinical outcome in breast cancer. *Nature*, **415**, 30–536.