Pacific Symposium on Biocomputing 13:279-290(2008)

INTEGRATION OF MICROARRAY AND TEXTUAL DATA IMPROVES THE PROGNOSIS PREDICTION OF BREAST, LUNG AND OVARIAN CANCER PATIENTS

O. GEVAERT, S. VAN VOOREN, B. DE MOOR
BioI@ESAT-SCD, Dept. Electrical Engineering
Katholieke Universiteit Leuven
Kasteelpark Arenberg 10, Leuven, B-3001, Belgium
E-mail: olivier.gevaert@esat.kuleuven.be

Microarray data are notoriously noisy such that models predicting clinically relevant outcomes often contain many false positive genes. Integration of other data sources can alleviate this problem and enhance gene selection and model building. Probabilistic models provide a natural solution to integrate information by using the prior over model space. We investigated if the use of text information from PUBMED abstracts in the structure prior of a Bayesian network could improve the prediction of the prognosis in cancer. Our results show that prediction of the outcome with the text prior was significantly better compared to not using a prior, both on a well known microarray data set and on three independent microarray data sets.

1. Introduction

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[1,2,3]. This is due to the rise in publicly available databases and also the number of databases has increased significantly[4]. 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 and often suffer from a small sample size. This causes that genes are often differently expressed between clinically relevant outcomes purely by chance. Integration of prior knowledge can improve model building in general and gene selection in particular. In this paper we present an approach to integrate information from litera-

Pacific Symposium on Biocomputing 13:279-290(2008)

ture 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[2]. 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 of a Bayesian network can improve the prediction of prognosis in cancer when modeling expression data. Bayesian networks provide a straightforward way to integrate information in the prior distribution over the possible structures of its network. By mining 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[5,6] where we have shown that structure prior information improves model selection especially when few data is available.

In this study we investigated if a Bayesian network model with a text prior can be used to predict the prognosis in cancer. Bayesian networks and their combination with prior information have already been studied by others[3,7,8,9] however, to the author's knowledge, none have investigated the influence of priors in a classification setting or, more specifically, when predicting the outcome or phenotypic group of cancer patients. First, we will show how the prior performs on a well known breast cancer data set and examine the effect of the prior in more detail. Subsequently, we will validate our approach on three other data sets studying breast, lung and ovarian cancer.

2. Bayesian networks

A Bayesian network is a probabilistic model that consists of two parts: a directed acyclic graph which is called the structure of the model and local probability models[10]. The dependency structure specifies how the variables (i.e. gene expression levels) are related to each other by drawing directed edges between the variables without creating directed cycles. In our case each variable x_i models the expression of a particular gene. Such a variable

Pacific Symposium on Biocomputing 13:279-290(2008)

or gene depends on a possibly empty set of other variables which are called the parents (i.e. their putative regulators):

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i|Pa(x_i)) \tag{1}$$

where $Pa(x_i)$ are the parents of x_i and n is the total number of variables. Usually the number of parents for each variable is small and therefore a Bayesian network is a sparse way of writing down a joint probability distribution. The second part of this model, the local probability models, specifies how the variables or gene expressions depend on their parents. We used discrete-valued Bayesian networks which means that these local probability models can be represented with Conditional Probability Tables (CPTs). Such a table specifies the probability that a variable takes a certain value given the value or state of its parents.

2.1. *Model building*

We already mentioned that a discrete valued Bayesian network consists of two parts: the structure and the local probability models. Consequently, there are two steps to be performed during model building: structure learning and learning the parameters of the CPTs. First the structure is learned using a search strategy. Since the number of possible structures increases super-exponentially with the number of variables, we used the well-known greedy search algorithm K2[11] in combination with the Bayesian Dirichlet (BD) scoring metric[11,12,13]:

$$p(S|D) \propto p(S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \left[\frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})} \right], \tag{2}$$

with N_{ijk} the number of cases in the data set D having variable x_i in state k associated with the j-th instantiation of its parents in current structure S. Γ corresponds to the gamma distribution. Next, N_{ij} is calculated by summing over all states of a variable: $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. In our case the state of a variable refers to the expression of the corresponding gene where each variable can have one of three states: over-expressed, under-expressed or no expression. Next, N'_{ijk} and N'_{ij} have similar meanings as N_{ijk} and N_{ij} but refer to prior knowledge for the parameters. When no knowledge is available they are estimated using[13]: $N'_{ijk} = \frac{N}{r_i q_i}$ with N the equivalent sample size, r_i the number of states of variable x_i and q_i the number of instantiations of the parents of variable x_i. K2 uses a prior ordering of

Pacific Symposium on Biocomputing 13:279-290(2008)

the variables to restrict the number of structures that can be built. The order of the variables reflects the causal relationship between the variables, this means that regulators should come before their targets in the ordering. Because the prior ordering of the variables is not known in advance we repeat the model building process for a set of randomly drawn variable orderings and choose the model with the highest posterior BD score. The next step consists of estimating the parameters of the local probability models of each variable in the structure with the highest BD score. This amounts to filling in a CPT for every variable and every possible value of its parents using the data. This Bayesian network can then be used to predict future data.

2.2. *Structure prior*

Previously two approaches have been used to define informative prior distributions over Bayesian network structures[5]. First there are penalization methods that start from a prior structure and score structures based on the difference with the prior structure[13]. Secondly there are pairwise methods which define the prior probability of a Bayesian network structure by combining individual edge scores between variables. This method assumes that being a parent of some node is independent of any other parental relation. We have chosen the second approach to model the structure prior where the prior probability of a structure is decomposed as:

$$p(S) = \prod_{i=1}^{n} p(Pa(x_i) \rightarrow (x_i)) \tag{3}$$

The probability of a local structure (i.e. $p(Pa(x_i) \rightarrow x_i)$) is then calculated by multiplying the probability that there is an edge between the parents of x_i and, the probability there is no edge between the other variables and x_i:

$$p(Pa(x_i) \rightarrow x_i) = \prod_{p \in Pa(x_i)} p(p \rightarrow x_i) \prod_{y \notin Pa(x_i)} p(y \nrightarrow x_i) \tag{4}$$

where $\nrightarrow$ means no edge between y and x_i. These individual edge probabilities can be represented in a matrix. In the Text Prior Section, we will be able to derive a matrix S from the literature where the elements represent the connectedness or similarity between the genes. Rather than using these values immediately as edge probabilities, we will introduce an extra parameter, ν called the mean density, which controls the density of the networks that will be generated from the distribution. We will transform all the matrix elements in the prior with an exponent ζ such that the

Pacific Symposium on Biocomputing 13:279-290(2008)

average of the mean number of parents per local substructure is according to the given mean density[5]. Finding the exponent, ζ that gives rise to the correct mean number of parents can be done with any single variable optimization algorithm. With this mean number of parents, we can control the complexity of the networks that will be learned.

2.3. *Inference*

After learning the model, we can use it to predict unseen data. This means that we can use a Bayesian network model to predict the value of a variable given the values of the other variables. We used the probability propagation in tree of cliques algorithm[13] to predict the state of the class variable (i.e. the prognosis in cancer). This inference algorithm was then used to evaluate the effect of using a text prior in combination with the expression data described below. To accomplish this, we used a randomization approach where we randomly distributed the data in 70% used to build a model and 30% to estimate the Area Under the ROC curve (AUC). This process was repeated 100 times to have a robust estimate of the generalization performance of the two approaches: with text prior and without text prior. Then these 100 AUCs were averaged and reported. Next a model was built using the complete data set for both methods and we investigated the possible differences between the Markov blanket variables (i.e. the set of genes which are sufficient to predict the outcome). The average AUC with and without prior are compared by calculating the p-value with a two-sided Wilcoxon rank sum test. P-values are considered statistically significant if smaller than 0.05.

3. Prior data

3.1. *Gene prior*

Since microarray data usually references thousands of genes, it is infeasible to manually construct a structure prior as described earlier. Therefore, prior construction involves methods based on an automatic elicitation of relationships between genes. In this paper, we propose the use of priors that consist of gene-by-gene similarity matrices based on biomedical literature mining. To accomplish this, genes are represented in the Vector Space Model . In the VSM model, each position of a gene vector corresponds to a term or phrase in a controlled vocabulary. In our case, we have constructed a cancer specific vocabulary which was extracted from the National Cancer Institute Thesaurus. Using a fixed vocabulary has several advantages.

Pacific Symposium on Biocomputing 13:279-290(2008)

Firstly, simply using all terms that occur in the corpus of literature linked to the genes involved in the microarray experiment at hand, will result in vectors of considerable size, which means genes are represented in a high dimensional space. As this 'curse of dimensionality' is detrimental to the strength of a metric, the use of only a relatively small set of concepts will improve the quality of calculated gene-to-gene distances. Further reduction of the dimensionality is accomplished by performing stemming, which will allow different terms that in essence convey a same meaning (coughing, coughs, coughed) to be treated as a single concept (cough). Secondly, the use of phrases reduces noise in the data set, as genes will only be compared to each other from a highly domain specific view. Thirdly, a structured vocabulary will enable the use of multi-word phrases as opposed to just single terms, without having to resort to co-occurrence statistics on the corpus to detect them. Fourthly, there is no need to filter out articles and stop words, as only highly specific cancer related terms are considered. The gene vectors themselves are constructed as follows. For each gene, manually curated literature references are extracted from Entrez Gene. All PUBMED abstracts linked to these genes are then indexed using the aforementioned vocabulary. As a result, all PUBMED abstracts are represented in a high dimensional vector space using IDF (Inverse Document Frequency) weights for non-zero vector positions. The resulting vectors (which represent abstracts, not genes) are normalized to bring them on the union hyper sphere in the vector space, which facilitates cosine similarity calculation. Gene vectors are then constructed by averaging the vectors of all the abstracts associated to that gene by Entrez Gene. Finally, the cosine measure is used to obtain gene-to-gene distances between 0 and 1. These gene-to-gene distances can then be represented as a symmetric matrix S which forms the structure prior for the Bayesian network modeling.

3.2. *Class variable prior*

We have already defined the way the prior is determined between the genes. Since we are developing models which predict the prognosis in cancer, the need exists for an additional variable in the model, namely the outcome class of the patients. This variable describes to which group each sample belongs, for example, good prognosis and poor prognosis. Hence, we need to define the prior relation between the class variable and the genes. To accomplish this, we used terms in the vocabulary which are related to the prediction of the prognosis of cancer, such as outcome, prognosis and

Pacific Symposium on Biocomputing 13:279-290(2008)

metastasis. Next, we counted the number of associations each gene had with prognosis related terms and increased the gene-to-outcome similarity for every additional term the gene was associated with. Genes which had no association with either term were given a prior probability of 0.5. This information was added to the gene prior creating a structure prior for all the variables studied (i.e. genes and patient outcome). This structure prior is then, after scaling according to the mean density, used in Bayesian network learning.

4. Data

To test our approach we used publicly available microarray data on breast cancer[14] (Veer data). This data set consists of 46 patients that belonged to the poor prognosis group and 51 patients that belonged to the good prognosis group. DNA microarray analysis was used to determine the mRNA expression levels of approximately 25000 genes for each patient. Every tumour sample was hybridized against a reference pool made by pooling equal amounts of RNA from each patient. The ratio of the sample and the reference was used as a measure for the expression of the genes and they constitute the microarray data set. This data set was already background corrected, normalized and log-transformed. Preprocessing was done similarly as in[14]. This resulted in 232 genes that where correlated with the patient outcome which were used in our models.

To validate our results we used three publicly available data sets from Bild et al.[15] studying breast, lung and ovarian cancer (Bild data). These data sets contained data on 171 breast cancer patients, 147 ovarian cancer patients and 91 lung cancer patients. The three groups of tumours were analysed on different Affymetrix chips; the breast tumours were hybridized on Hu95Av2 arrays, the ovarian tumours on Hu133A arrays and the lung tumours on Human U133 2.0 plus arrays. The data were already preprocessed using RMA. For all cancer sites survival data was available and patients were split up in two groups according to the following thresholds: 53 months for breast cancer, 62 months for ovarian cancer and 36 months for lung cancer. The thresholds were chosen to make sure both classes contained approximately the same number of samples. Genes were selected similarly as in the Veer data set by selecting the top 100 genes after ranking them by their correlation with patient survival data.

Pacific Symposium on Biocomputing 13:279-290(2008)

4.1. *Discretization*

We have chosen discrete valued Bayesian networks therefore the microarray data has to be discretized. We specifically tried to minimize the loss of relationships between the variables by applying the algorithm of Hartemink[16]. The gene expression values were discretized in three categories or bins: baseline, over-expression and under-expression. This was done using a multivariate discretization method which minimizes the loss of mutual information between the gene expression measurements[16]. First a simple discretisation method with a large number of bins is used as a starting point (e.g. interval discretisation where the complete range of values is divided in a number of equally large bins). Then the multivariate algorithm starts and for each variable it joins the neighboring bins together which have the smallest decrease in mutual information. This is iterated until each variable has three bins. The resulting discretized data set is used as input into the Bayesian network learning algorithms.

5. Implementation

The software implementation is based on a combination of c++, java, matlab and perl. The Bayesian network algorithms were implemented in C++. Java Lucene was used for indexing Pubmed. Matlab scripts were used for discretization and to construct the structure priors. Perl was used to glue the different steps in the workflow together. A typical analysis took between 6 and 25 minutes depending on the data set size. All analysis where run on AMD dual core opteron 2.4 GHz with between 4 and 16 GB RAM memory.

6. Results and discussion

6.1. *Veer data*

First, we assessed the performance of the text prior regarding prediction of outcome on the Veer data set. We performed 100 randomizations of the data set without a prior and 100 randomizations with the text prior (as described in the Model building and testing Section in Materials and methods). We repeated the analysis for different values of the mean density to asses if this parameter had an influence on the results. Table 1 shows the mean AUC for both methods and for increasing mean density. The most important conclusion that can be drawn from Table 1 is that using the text prior significantly enhances the prediction of the outcome (P-value

Pacific Symposium on Biocomputing 13:279-290(2008)

$\prec 0.05$). The text prior guides model search and favors genes which have a prior record related to prognosis. This knowledge improves gene selection and most likely wards off genes which are differentially expressed by chance. Additionally, Table 1 shows that the mean density has no influence on the result in the tested range. The mean density controls the complexity of the network therefore large values should be avoided since the danger of overfitting increases. Note that the results for the mean AUC without prior are essentially the same as our previously obtained result[12].

Table 1. Results of 100 randomizations of the Veer data set with the Text prior and without prior. The mean AUCs are reported together with the p-value.

Mean Density	Text prior mean AUC	Uniform prior mean AUC	P-value
1	0.80	0.75	0.000396
2	0.80	0.75	0.000002
3	0.79	0.75	0.005770
4	0.79	0.74	0.000006

Next, we used the complete data set and we built one model with text prior and one model without the text prior, to evaluate the set of genes which are sufficient to predict the outcome (i.e. the genes in the Markov blanket of the outcome). We call the former, the TXTmodel and the latter UNImodel. Table 2 shows the gene names that appear in both models. The average text score (i.e. the probability the gene is related to patient outcome according to literature) of the genes in the TXTmodel is 0.85 compared to only 0.58 for the UNImodel. The text prior thus has its expected effect and includes genes which have a prior tendency to be associated with the prognosis of cancer. There are only 10 genes in the TXTmodel compared to 15 genes in the UNImodel which indicates that TXTmodel needs fewer genes. Moreover, the TXTmodel has many genes which have been implicated in breast cancer or cancer in general such as TP53, VEGF, MMP9[17], BIRC5, ADM[18] and CA9. Next ACADS, NEO1 and IHPK2 have a weaker link to cancer outcomes whereas MYLIP has no association. In the UNImodel, as expected, far less genes are present which have a strong link with cancer outcomes which likely increases the probability of false positives. Only WISP1, FBXO31, IGFBP5 and TP53 have a relation with breast cancer outcome. The other genes have mostly unknown function or are not related. Finally two genes appear in both set: TP53 and IHPK2.

Pacific Symposium on Biocomputing 13:279-290(2008)

TP53 is perhaps the best known gene to be involved in cancer. Therefore it is bound to appear in the TXTmodel and it is no surprise that it is also present in the UNImodel. IHPK2 however has a weak prior relation with prognosis in cancer therefore this gene proves that genes with a low text prior still can be selected in the TXTmodel. Additionally, genes which appear in both models can be considered more reliable.

Table 2. Genes sufficient to predict the outcome variable for the TXTmodel and the UNImodel.

TXTmodel:	MYLIP,TP53,ACADS,VEGF,ADM,NEO1,IHPK2,CA9,MMP9,BIRC5
UNImodel:	PEX12,LOC643007,WISP1,SERF1A,QSER1,ARL17P1,LGP2,IHPK2, TSPYL5,FBCO31,LAGE3,IGFBP5,AYTL2,TP53,PIB5PA

6.2. *Bild data*

Finally we validated our approach on three independent data sets on breast, ovarian and lung cancer[15] to assess if the results on the Veer data set can be confirmed. Based on the results presented in Table 1 we chose a mean density of 1 for these data sets. Again 100 randomizations of the data set with and without the text prior were performed. Table 3 shows the average AUC for the three Bild data sets and confirms that the text prior significantly improves the prediction of the prognosis on independent data sets and for other cancer sites.

Table 3. Results of 100 randomizations of the three Bild data sets with the Text prior and without prior. The mean AUCs are reported together with the p-value.

Mean Density	Text prior mean AUC	Uniform prior mean AUC	P-value
Breast	0.79	0.75	0.00020
Lung	0.69	0.63	0.00002
Ovarian	0.76	0.74	0.02540

7. Conclusions

In this paper we have shown a method to integrate information from literature abstracts with gene expression data using Bayesian network models. This prior information was integrated in the prior distribution over the

Pacific Symposium on Biocomputing 13:279-290(2008)

possible Bayesian network structures after scaling. The results of the randomization analysis in Table 1 and 3 have shown that for both the Veer data set and the three Bild data sets the text prior improves the prediction of the prognosis of cancer patients significantly.

A possible limitation of our approach is the discretization of the data. It is inevitable that some information is lost in the process of discretization. We have chosen discrete valued Bayesian networks because the space of arbitrary continuous distributions is large. A solution could be to restrict ourselves to the use of Gaussian Bayesian networks but this class of models assumes linear interactions between the variables which, in our opinion, would restrict too much the type of relations among genes that are modeled. Moreover, by using the algorithm of Hartemink we are performing a multivariate discretization, keeping the relationships between the variables as much as possible intact.

Secondly by using text information, which is often described as highly biased, one could run the risk of focussing too much on the hot genes disregarding novel important genes. However, in our case the emphasis is not so much on biomarker discovery and more on developing models which can accurately predict the prognosis of disease. There are already many genes known to be involved in different types of cancer based on individual studies or because they are member of a cancer profile. Finding the minimal set of genes which is able to predict the prognosis of disease however, is still an open problem. Our Bayesian network framework attempts to address this issue by tackling the disadvantages of cancer microarray data sets (low signal-to-noise ratio, high dimensional, small sample size, ...) by using information from the literature as a guide.

Finally, the presented framework is complimentary to our previously published method to integrate clinical and microarray data with Bayesian networks[12]. Thus creating a Bayesian network framework which enables modeling of various data sources (i.e. clinical, microarray and text) to improve decision support of outcome (i.e phenotypic group) prediction in cancer or other genetic diseases. Moreover our definition of the structure prior makes no assumptions about the nature of prior information. Therefore other sources of information can be combined with the text prior (e.g. known protein-DNA interactions from Transfac, known pathways from KEGG or motif information). Thus, creating a white box framework that visualizes how decisions are made by a model. This is in contrast to for example a kernel framework where model parameters are not readily interpretable.

Pacific Symposium on Biocomputing 13:279-290(2008)

Acknowledgments

This research is supported by the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen) and GOA AM-BioRICS, CoE EF/05/007 SymBioSys, FWO: G.0499.04 (Statistics), IUAP P6/25 BioMaGNet 2007-2011; FP6-NoE Biopattern; FP6-IP e-Tumours

References

1. Z. Bar-Joseph, G. K. Gerber, T. I. Lee, N. J. Rinaldi, J. Y. Yoo, F. Robert, D. B. Gordon, E. Fraenkel, T. S. Jaakkola, R. A. Young and D. K. Gifford, *Nat Biotechnol* **21**, 1337(November 2003).
2. G. R. G. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan and W. S. Noble, *Bioinformatics* **20**, 2626 (2004).
3. A. Bernard and A. J. Hartemink, *PSB* **10**, 459 (2005).
4. M. Y. Galperin, *Nucl. Acids Res.* **34**, D3 (2006).
5. P. Antal, G. Fannes, D. Timmerman, Y. Moreau and B. De Moor, *Artif Intell Med* **30**, 257 (2004).
6. O. Gevaert, F. De Smet, E. Kirk, B. Van Calster, T. Bourne, S. Van Huffel, Y. Moreau, D. Timmerman, B. De Moor and G. Condous, *Human reproduction* **21** (2006).
7. N. Friedman, M. Linial, I. Nachman and D. Pe'er, *J Comput Biol* **7**, 601 (2000).
8. N. Nariai, S. Kim, S. Imoto and S. Miyano, *PSB* **9**, 336 (2004).
9. T. Ideker, O. Ozier, B. Schwikowski and A. F. Siegel, *Bioinformatics* **18 Suppl 1** (2002).
10. J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann Publishers, San Matteo, California, 1988).
11. G. F. Cooper and E. Herskovits, *Machine Learning* **9** (1992).
12. O. Gevaert, F. De Smet, D. Timmerman, Y. Moreau and B. De Moor, *Proceedings of the 14th Annual ISMB, Bioinformatics special issue* (2006).
13. D. Heckerman, D. Geiger and D. M. Chickering, *Machine Learning* **20**, 197 (1995).
14. L. Van 't Veer, H. Dai, M. J. Van de Vijver, U. D. He, A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards and S. H. Friend, *Nature* **415**, 530 (2002).
15. A. H. Bild, G. Yao, J. T. Chang, Q. Wang, A. Potti, D. Chasse, M.-B. Joshi, D. Harpole, J. M. Lancaster, A. Berchuck, J. A. Olson, J. R. Marks, H. K. Dressman, M. West and J. R. Nevins, *Nature* (November 2005).
16. A. J. Hartemink, Principled computational methods for the validation and discovery of genetic regulatory networks, PhD thesis, MIT2001.
17. J. L. Owen, V. Iragavarapu-Charyulu and D. M. Lopez, *Breast Dis* **20**, 145 (2004).
18. M. K. Oehler, D. C. Fischer, M. Orlowska-Volk, F. Herrle, D. G. Kieback, M. C. Rees and R. Bicknell, *Br J Cancer* **89**, 1927(November 2003).