BIOINFORMATICS

Vol. 00 no. 00 2006
Pages 1–7

Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks.

Olivier Gevaert [a*], Frank De Smet [a,b], Dirk Timmerman [c], Yves Moreau [a], Bart De Moor [a]

[a]Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium [b]Medical Direction, National Alliance of Christian Mutualities, Haachtsesteenweg 579, 1031 Brussel, Belgium, [c] Department of Obstetrics and Gynecology, University Hospital Gasthuisberg, Katholieke Universiteit Leuven, Herestraat 49, 3000 Leuven, Belgium

ABSTRACT

Motivation: Clinical data, such as patient history, laboratory analysis, ultrasound parameters - which are the basis of day-to-day clinical decision support - are often underused to guide the clinical management of cancer in the presence of microarray data. We propose a strategy based on Bayesian networks to treat clinical and microarray data on an equal footing. The main advantage of this probabilistic model is that it allows to integrate these data sources in several ways and that it allows to investigate and understand the model structure and parameters. Furthermore using the concept of a Markov Blanket we can identify all the variables that shield off the class variable from the influence of the remaining network. Therefore Bayesian networks automatically perform feature selection by identifying the (in)dependency relationships with the class variable.

Results: We evaluated three methods for integrating clinical and microarray data: decision integration, partial integration and full integration and used them to classify publicly available breast cancer patients into a poor and a good prognosis group. The partial integration method is most promising and has an independent test set area under the ROC curve of 0.845. After choosing an operating point the classification performance is better than frequently used indices.

Contact: olivier.gevaert@esat.kuleuven.be

1 INTRODUCTION

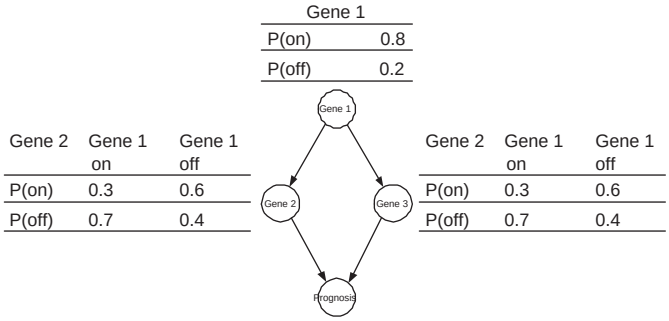
In the past decade microarrays have had a great impact on cancer research. This technology allows to measure the expression of thousands of genes at once; possibly representing the whole genome. Usually a microarray consists of a selection of probes which are applied onto a solid surface and represent a number of genes (Lockhart *et al.* (1996); Brown and Botstein (1999)). Reverse transcribed mRNA extracted from a tumor sample can be hybridized with the probes on this surface. This results in expression levels of thousands of genes for every tumor sample that is hybridized. The resulting data has been used for many applications such as class discovery and the prediction of diagnosis, prognosis or treatment response. Several studies have been conducted using microarray technology studying several types of cancer (Golub *et al.* (1999); Bhattacharjee *et al.* (2001); Singh *et al.* (2002); van 't Veer *et al.* (2002); van de Vijver *et al.* (2002); Spentzos *et al.* (2004, 2005)).

*to whom correspondence should be addressed

However, microarray data is high dimensional, characterized by many variables and few observations. Moreover this technique suffers from a low signal-to-noise ratio. In our opinion, integration of other sources of information could be important to counter randomly generated differences in expression levels. For example Shedden *et al.* (2003) used a pathological framework and showed that this information significantly lowered the number of genes required in their model. Nevertheless, the focus in most studies is on the microarray analysis while the clinical data is not used in the same manner. Clinical data includes for example: patient history, laboratory analysis or ultrasound parameters. This data was the basis of research and fully guided the clinical management of cancer in the pre-microarray era and is, in our opinion, often underused when microarray data is available. Here we propose methods based on Bayesian networks that integrate clinical data and microarray data. These methods treat both the clinical and the microarray variables (i.e. the gene expression levels) in the same manner. For example, Shedden *et al.* (2003) also did not add clinical data to the gene expression levels when classifying tumour samples.

Bayesian networks are popular decision support models (Husmeier *et al.* (2005)) because they inherently model the uncertainty in the data. They are a successful marriage between probability theory and graph theory. They allow to model a multidimensional probability distribution in a sparse way by searching independency relations in the data. Furthermore this model allows different strategies to integrate two data sources. First, it is possible to combine data sources directly or, secondly, by combining them at the decision level. Furthermore, because Bayesian networks are learned from data in two independent steps, we can define a third method to integrate both data sources. These three methods will be presented and evaluated using Receiver Operator Characteristic (ROC) curves on the training set. The method with the highest average ROC performance will be evaluated on an independent test set. To the author's knowledge, the first two methods have not been previously applied in this context and the third method has not been previously defined.

We will focus as an example on the prediction of the prognosis in lymph node negative breast cancer (without apparent tumor cells in local lymph nodes at diagnosis). We define the outcome as a variable that can have two values: poor prognosis or good prognosis. Poor prognosis corresponds to recurrence within 5 years after diagnosis and good prognosis corresponds to a disease free interval of



Gene 1	
P(on)	0.8
P(off)	0.2

Gene 2	Gene 1 on	Gene 1 off
P(on)	0.3	0.6
P(off)	0.7	0.4

Gene 2	Gene 1 on	Gene 1 off
P(on)	0.3	0.6
P(off)	0.7	0.4

Prognosis	Gene 2 on Gene 3 on	Gene 2 on Gene 3 off	Gene 2 off Gene 2 on	Gene 2 off Gene 3 off
P(good)	0.6	0.1	0.9	0.5
P(poor)	0.4	0.9	0.1	0.5

Fig. 1. A simple example of a Bayesian network with four binary variables. The conditional probability tables are shown next to each node where each column in such a table refers to a specific instantiation of the parents. Gene 1 has no parents therefore the node's table specifies a priori probabilities.

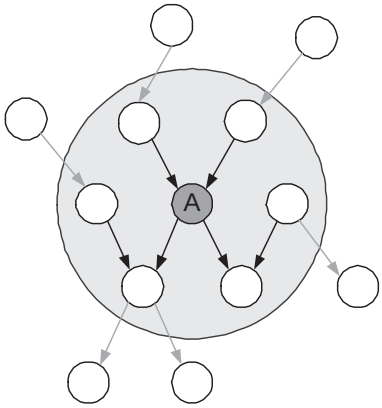


Fig. 2. The Markov blanket of variable A is composed of the variable's parents, its children and its children other parents. Here the Markov blanket variables are shown in a grey circle.

at least 5 years (van 't Veer *et al.* (2002)). If we can distinguish between these two groups, patients could be treated more optimally thus eliminating over- or under-treatment.

2 METHODS

2.1 Bayesian networks

2.1.1 Definition A Bayesian network is a probabilistic model that consists of two parts: a dependency structure and local probability models (Pearl (1988); Neapolitan (2004)). The dependency structure specifies how the variables are related to each other by drawing directed edges between the variables without creating directed cycles. Each variable depends on a possibly empty set of other variables which are called the parents:

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

where $Pa(x_i)$ are the parents of x_i. Usually the number of parents for each variable is small 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 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 of its parents. Figure 1 shows an example of a Bayesian network with four binary variables. The prognosis variable in this example has two parents: gene 2 and gene 3. The CPTs for each variable are shown alongside each node.

2.1.2 Markov Blanket An important concept of Bayesian networks is the Markov blanket of a variable. The Markov blanket of a variable is the set of variables that completely shields off this variable from the other variables. This set consists off the variable's parents, children and its children's other parents. A variable in a Bayesian network is conditionally independent of the other variables given its Markov Blanket. Conditional independency means that when the Markov blanket of a certain variable x is known, adding knowledge of other variables leaves the probability of x unchanged (Korb and Nicholson (2004)). This is an important concept because the Markov blanket is the only knowledge that is needed to predict the behaviour of that variable. For classification purposes we will focus on the Markov Blanket of the outcome variable. The concept of a Markov blanket is shown in Figure 2.

2.2 Bayesian network learning

Previously we mentioned that a discrete valued Bayesian network consists of two parts. Consequently, there are two steps to be performed during model building: structure learning and learning the parameters of the CPTs.

2.2.1 Structure learning 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 (Cooper and Herskovits (1992)) in combination with the Bayesian Dirichlet (BD) scoring metric:

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

with N_{ijk} the number of cases in the data set D having variable i in state k associated with the j-th instantiation of its parents in current structure S. n is the total number of variables. Next, N_{ij} is calculated by summing over all states of a variable: $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. N'_{ijk} and N'_{ij} have similar meanings but refer to prior knowledge for the parameters. When no knowledge is available they are estimated using $N'_{ijk} = N/(r_i q_i)$ (Heckerman *et al.* (1995)) with N the equivalent sample size, r_i the number of states of variable i and q_i the number of instantiations of the parents of variable i. $\Gamma(.)$ corresponds to the gamma distribution. Finally $p(S)$ is the prior probability of the structure. $p(S)$ is calculated by: $p(S) = \prod_{i=1}^{n} \prod_{l_i=1}^{p_i} p(l_i \rightarrow x_i) \prod_{m_i=1}^{o_i} p(m_i \nrightarrow x_i)$ with p_i the number of parents of variable x_i and o_i all the variables that are not a parent of x_i. Next, $p(a \rightarrow b)$ is the probability that there is an edge from a to b while $p(a \nrightarrow b)$ is the inverse, i.e. the probability that there is no edge from a to b. Since we are interested in the prediction of the prognosis, edges with the outcome variable are given a higher prior probability than other edges.

Using Equation 2 we can now score structures using the K2 search strategy. K2 consists of a greedy search combined with a prior ordering of the variables. This ordering restricts the search space by only allowing parents if they precede the current variable in the ordering. Then K2 iteratively tries to find the best parents for each variable separately by starting with an empty set of parents and incrementally adding the best parents. When the addition

of a parent does not increase the score, the algorithm stops and moves on to the next variable in the ordering. Since the ordering of the variables is not known in advance, the model building process is iterated a number of times with different permutations of the ordering. Then the network with the highest score is chosen.

2.2.2 Parameter learning The second step of the model building process consists of estimating the parameters of the local probability models corresponding with the dependency structure. In section 2.1.1 we reported that we are using CPTs to model these local probability models. For each variable and instantiation of its parents there exists a CPT that consists of a set of parameters. Each set of parameters was given a uniform Dirichlet prior:

$$p(\theta_{ij}|S) = Dir(\theta_{ij}|N'_{ij1}, ..., N'_{ijk}, ..., N'_{ijr_i}) \quad (3)$$

with θ_{ij} a parameter set where i refers to the variable and j to the j-th instantiation of the parents in the current structure. θ_{ij} contains a probability for every value of the variable x_i given the current instantiation of the parents. Dir corresponds to the Dirichlet distribution with $(N'_{ij1}, ..., N'_{ijr_i})$ as parameters of this Dirichlet distribution. Parameter learning then consists of updating these Dirichlet priors with data. This is straightforward because the multinomial distribution that is used to model the data, and the Dirichlet distribution that models the prior, are conjugate distributions. This results in a Dirichlet posterior over the parameter set:

$$p(\theta_{ij}|D, S) = Dir(\theta_{ij}|N'_{ij1} + N_{ij1}, ..., N'_{ijk} + N_{ijk}, ..., N'_{ijr_i} + N_{ijr_i}) \quad (4)$$

with N_{ijk} defined as before. We summarized this posterior by taking the Maximum A Posteriori (MAP) parameterization of the Dirichlet distribution and used these values to fill in the corresponding CPTs for every variable. Using MCMC could improve our current set-up because this technique allows devising the complete posterior distribution (Neal (1996)).

2.3 Data

We used the data of van 't Veer *et al.* (2002) which is available at http://www.rii.com/publications/default.htm or in the Integrated Tumor Transcriptome Array and Clinical data Analysis database (ITTACA (2006)). This data set consists of two groups of patients. The first group of patients, which we call the training set, consists of 78 patients of which 34 patients belonged to the poor prognosis group and 44 patients belonged to the good prognosis group. The second group of patients, the test set, consists of 19 patients of which 12 patients belonged to the poor prognosis group and 7 patients 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. Each patient also had the following clinical variables recorded: age, diameter, tumor grade, oestrogen and progesterone receptor status, the presence of angioinvasion and lymphocytic infiltration, which together form the clinical data.

2.3.1 Preprocessing The microarray data consists of approximately 25000 expression values per patient, which was already background corrected, normalized and log-transformed. An initial selection was done (similar to van 't Veer *et al.* (2002)) by removing the genes that did not meet the following criteria using only the training data: at least a twofold increase or decrease and a P-value of less than 0.01 in more than 3 tumors. This resulted in a subset of approximately 5000 genes. Then we calculated the correlation between the expression values of these genes with the binary outcome and selected the genes with a correlation of ≥ 0.3 or ≤ -0.3. This resulted in 232 genes that where correlated with the outcome. Missing values were estimated using a 15-weighted nearest neighbours algorithm (Troyanskaya *et al.* (2001)). Then these genes were discretized into three categories: baseline, over-expression or under-expression according to two thresholds. These thresholds depended on the variance of the gene such that a gene with high variance receives a higher threshold than a gene with low variance. The data set that results from these steps was used as input for the Bayesian network software.

2.3.2 Model building We evaluated the performance of the different methods for integrating both data sources (see section 2.4) using the training data. This was done by randomizing the training data set 100 times, in a stratified way, into a set of 70% of the patients used to build the model (model building data set) and a set of 30% to estimate the Area Under the ROC curve (AUC). Then these 100 AUCs were averaged and reported. In this manner we can evaluate the generalizing performance of a specific method and compare with other methods.
Next, the method that performed best in the previous step was used to train 100 models with different orderings using the complete training set. The model with the highest AUC among these 100 models was chosen to predict the outcome on the test set.

2.4 Integration of data sources

2.4.1 Full integration Bayesian networks allow to combine the two data sources, the clinical and microarray data, in different ways. The first method, full integration, is equal to putting both data sources together and treating them as if it is one dataset. This means that both the clinical variables (e.g. age, diameter, grade, etc.) and the microarrays variables (mRNA expressions for each gene) are offered as one data set to the Bayesian network learning algorithm. In this manner the developed model can contain any type of relationship between the clinical variables and the microarray variables.

2.4.2 Decision integration The decision integration method amounts to learning a separate model for the clinical and the microarray data. Then the predictions for the outcome are fused. This comes down to combining the probability of the outcome for the clinical model with the probability of the outcome for the microarray model using weights. The weight parameter is trained using only the model building data set (see section 2.3.2) within each randomization which, in the context of decision integration, is called an outer randomization. This is done by performing again 100 inner randomizations of the model building data set within each outer randomization by again splitting this data set in 70% of the data for training and 30% of the data for testing. For each inner randomization the weight is increased from 0.0 to 1.0 in steps of 0.1. Then the weight value with the highest average AUC on the 30% left out data of the 100 inner randomizations is chosen as weight for the outer randomization.

2.4.3 Partial integration Bayesian networks also allow a third method, which we will call partial integration. This is due to the fact that learning Bayesian networks is a two step process. Therefore we can perform the first step, structure learning, separate for both data sources. This results in a structure for the clinical data and a structure for the microarray data. Both structures have only one variable in common, the outcome, since this variable is present in both data sources. The outcome variable allows joining the separate structures into one structure. Then the second step of learning Bayesian networks (i.e. parameter learning) starts with the combined clinical and microarray data. Partial integration is similar to imposing a restriction during structure learning where no links are allowed between clinical variables and gene expression variables.

3 RESULTS

Model building was done as described in section 2.3.2 for the three integration methods (full, partial and decision integration) and for both data sources (clinical and microarray) separately for comparison. In case of decision integration, we used randomizations to determine the weights to fuse the decisions as described in

2.4.2. This resulted in a weight of 0.6 for predicted probabilities of the clinical model and a weight of 0.4 for predicted probabilities of the microarray model, slightly favouring the clinical model. After choosing these optimal weights, we can compare the methods for integrating the data sources. Table 1 shows the AUCs for the developed models. Partial integration and decision integration are significantly different from the other methods but not significantly different from each other (Wilcoxon rank sum tests).

Table 1. Average AUC performance and standard deviation of the three methods for integrating clinical and microarray data and each data source separately with 100 randomizations. The first two methods, clinical and microarray, are for comparison. The next three methods (decision, partial and full) refer to the methods for integrating the clinical and microarray data.

Method	average AUC	Std
Clinical data	0.751	0.086
Microarray data	0.750	0.073
Decision integration	0.790	0.072
Partial integration	0.793	0.068
Full integration	0.747	0.099

Next, both decision integration and partial integration were chosen as the best methods of integrating the two data sources and 100 models were built using the training set. Then the best performing model for each method was chosen and used to predict the outcome on the test data set. The best partial integration model is referred to as BPIM (Best Partial Integration Model) and the best decision integration model as BDIM (Best Decision Integration Model). Table 2 shows the AUC of these two models on the test set. We compared our models with the 70 genes prognosis profile by applying the methods described in van 't Veer *et al.* (2002) and using the resulting classifier on the test set. The AUC is also shown in table 2, the standard deviations were estimated according to Hanley and McNeil (1983). Both BPIM and the 70 genes model perform in the same manner on the data set while BDIM is worse. However, there are no significant differences between the ROC curves of BDIM, BPIM and the 70 genes model (Hanley and McNeil (1983)).

Table 2. The AUC of the Bayesian network models (BPIM and BDIM) and of the reconstructed model based on van 't Veer *et al.* (2002) based on 70 genes.

	AUC	std
70 genes	0.851	0.132
BPIM	0.845	0.132
BDIM	0.810	0.118

Next we chose an operating point for BDIM and BPIM by choosing a threshold that corresponds with a maximum for the sum of the sensitivity and specificity (Smet *et al.* (2004)). Then we compared the classifications of our models with the 70 genes model and with the following indices: the St. Gallen consensus (Goldhirsch *et al.* (1998)), the National Institute of Health (NIH) index (Eifel *et al.* (2001)) and following (Edén *et al.* (2004)) also with the widely used Nottingham Prognostic Index (NPI) (Blamey *et al.* (1979)). For the NPI we used the standard threshold of 3.4 to determine a good or a poor prognosis. Below this threshold the prognosis is considered good, above this threshold the prognosis is considered moderate or poor (Todd *et al.* (1987)). Table 3 shows the number of patients that is assigned to the poor prognosis group for the complete test set, the set of true poor prognosis patients (i.e. sensitivity) and the set of true good prognosis patients (i.e. 1-specificity). We have applied the St Gallen consensus and the NIH index in the same manner as van 't Veer *et al.* (2002). The results show that both the St Gallen consensus and the NIH consensus criteria have a tendency to produce more false positives than the other models which has been observed before (Boyages *et al.* (2002)). In the test set both indices also have some false negatives which can be due to sample selection and small sample size. Both BPIM and the 70 genes have similar performance and are better than the other models since they produce few false positives and false negatives. Both tables 2 and 3 show that BPIM and the 70 genes have similar performance and are better than BDIM and the frequently used indices. BPIM and 70 genes can reliably be used to predict the prognosis in lymph node negative breast cancer.

Table 3. The number of patients assigned a poor prognosis for the complete test set and for the true poor and good prognosis patients.

	Total test set (n=19)	Metastasis within 5 yr (n=12)	Disease free at 5 yr (n=7)
St Gallen 1998♮	13/19 (68%)	10/12 (83%)	3/7 (43%)
NIH 2000○	15/19 (79%)	10/12 (83%)	5/7 (71%)
NPI♢	11/19 (58%)	9/12 (75%)	2/7 (29%)
70 genes†	14/19 (74%)	12/12 (100%)	2/7 (29%)
BPIM†	13/19 (68%)	11/12 (92%)	2/7 (29%)
BDIM†	11/19 (58%)	9/12 (75%)	2/7 (29%)

♮ Either one of the following criteria equals poor prognosis: ER negative, tumour diameter ≥ 2 cm, grade 3 or age < 35
○ Poor prognosis if tumour diameter > 1 cm.
♢ NPI is the sum of 0.2 times the tumour diameter in cms, lymph node stage and the tumour grade.
† The operating point is determined by maximizing the sum of the sensitivity and specificity on the training set.

Figure 3 shows the complete network built with partial integration. The outcome variable and its Markov Blanket is indicated with triangle nodes. Figure 4 shows the Markov Blanket in detail with the gene names where possible. There are three clinical variables: age, grade and angioinvasion and 13 genes, 12 annotated and 1 unannotated.

4 DISCUSSION

We have developed Bayesian networks to integrate clinical and microarray data using the data of van 't Veer *et al.* (2002) and investigated if an improvement was made for the prediction of metastasis in breast cancer. We investigated three methods for integrating the two data sources with Bayesian networks: full integration, partial integration and decision integration.
Table 1 showed that only partial integration and decision integration perform significantly better than each data source separately.

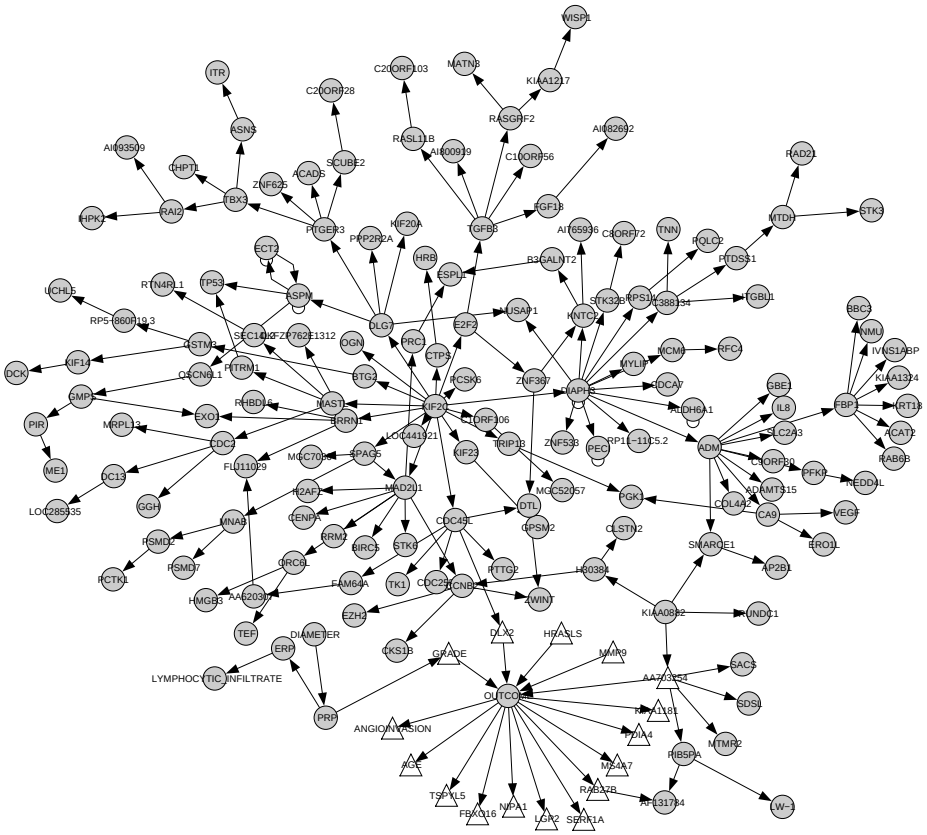


Fig. 3. The complete Bayesian network for the best model using partial integration of clinical and microarray data. The Markov blanket of the outcome variable is indicated with triangle white nodes.

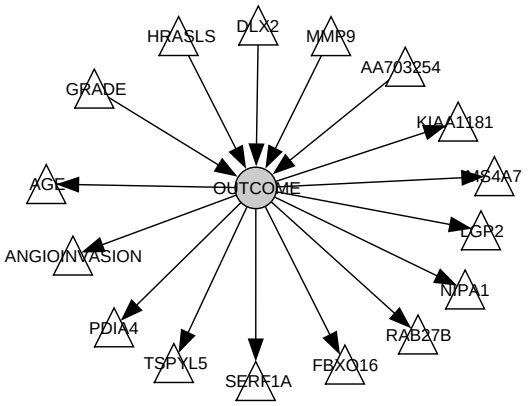


Fig. 4. Markov blanket of the outcome variable for the BPIM model. Gene names have been used where possible.

We believe that this is due to the different nature of the data sources. Clinical data has a low noise level, there are mostly fewer variables than observations and there are both discrete and continuous-valued variables. Microarray data on the other hand has a much higher noise level. There are a lot more variables than observations and all the variables are continuous. Therefore, it could be advisory to treat them separately in some way. Partial integration uses separate structure learning while decision integration builds separate models but fuses the outcome probabilities. Full integration does not make a distinction between these two heterogeneous data sources which causes that the clinical variables are submerged by the microarray variables and mostly have few connections. This leads to a model where the Markov Blanket only consists of microarray variables and explains the similar performance between full integration and using only the microarray data.

Next, table 2 showed that BPIM generalizes best to unseen data compared to BDIM. The difference between these two models is that BPIM is integrated at the parameter level and BDIM at the decision level. The former combines clinical and microarray variables in a more sophisticated way because combined parameter learning results in different parameters for every instantiation of the clinical variables. The latter method combines the outcome probabilities using a weighting scheme and relies on the weights for each model. Furthermore BPIM outperforms the prognostic indices and has comparable performance with the 70 genes prognosis profile (van 't Veer *et al.* (2002)) despite having fewer genes. This suggests that using

clinical data decreases the number of genes required to reliably predict the prognosis. Moreover the low number of genes in BPIM could allow the design of a cheaper test for breast cancer prognosis while still benefiting from data at the molecular level.

Next, we also looked more closely at the BPIM model to investigate the performance of the model when the links of the outcome variable with either the clinical variables or the microarray variables in the Markov blanket are removed. This resulted in worse performance of the model. When the links between the outcome and the clinical variables are removed the AUC performance drops to 0.804 (std 0.130). Similarly when the links between the outcome and the genes are removed the AUC performance drops to 0.798 (std 0.128). This is strong evidence that the combination between the clinical and the microarray variables boosts the performance. Also the formation of a prognostic index from a combination of clinical variables and a small number of genes seems possible.

Furthermore we searched the literature for relations of the variables in the Markov blanket of BPIM (see Figure 4) with breast cancer prognosis and metastasis. The presence of the clinical variables can be explained because they are used as conventional prognostic markers and in prognostic indices. Age in particular because patients with breast cancer at young age have been correlated with poor prognosis (Goldhirsch *et al*. (1998)) while grade is part of the NPI (Blamey *et al.* (1979)). Moreover, recently a large study has shown that lymphovascular invasion, which is related to angioinvasion, is an independent prognostic factor in node-negative breast cancer and improves the NPI (Lee *et al*. (2006)). Furthermore there are 13 genes, 12 annotated and 1 unannotated. Among the annotated genes, MMP9, HRASLS and RAB27B have strong associations with cancer (Owen *et al.* (2004); Kaneda *et al.* (2004)). MMP9 is associated with tumor invasion and angiogenesis since matrix metalloproteases are an important family of proteases that degrade a path through the extra-cellular matrix and the stroma. This process allows tumor cells to invade the surrounding tissue (Pecorino (2005)). HRASLS is associated with the RAS pathway (Malaney and Daly (2001)) and is thought to function as a tumor suppressor. Furthermore RAB27B is a member of the RAS oncogene family.

On the other hand BDIM also showed interesting characteristics. This decision integration model used a weight of 0.6 for the clinical model and a weight of 0.4 for the microarray model. This emphasizes the importance of the clinical data for classification compared to the microarray data. In addition, the clinical data generalizes better to new data since the test set performance is similar to the training set performance (average training set AUC of 100 clinical data models is 0.838) while the microarray data allows bitter fitting but with the danger of overfitting (average training set AUC of 100 microarray data models is 0.981) (also see Table 1). Therefore combining both data sources can lead to models benefiting from the complementary advantages of each data source separately. The results of BDIM and BPIM show that this is possible.

The advantages of the probabilistic approach are that the current models can be extended with prior information. This can be done both at the structure level and the parameter level. This will influence the variables that show up in the Markov blanket and results in a feature selection method based on data and prior biological knowledge with automatic tuning of the balance between data and prior knowledge. Possible sources of prior information are literature abstracts (Glenisson *et al.* (2004)), known pathways (e.g. KEGG or BIOCARTA) or motif information (Thijs *et al.* (2002)). Moreover publicly available microarray data sets studying the same clinical problem can be combined via the prior.

Furthermore, since Bayesian networks are not tuned for classification - they provide a more general framework by modeling a multi-dimensional probability distribution - the reported performance could be improved by using more traditional classifiers. Our ongoing research includes investigating the use of Bayesian networks as feature selector followed by Least Squares Support Vector Machines for classification (Pochet *et al.* (2004)).

In conclusion, the integrated use of clinical and microarray data outperforms the indices based on clinical data (NIH, St. Gallen and NPI) and has comparable performance with the 70 genes prognosis profile. Therefore this approach offers possibilities for the use of Bayesian networks to integrate data sources for other types of cancer and data. Furthermore BPIM has comparable performance as the 70 genes prognosis profile (van 't Veer *et al.* (2002)) but allows interpretation and contains fewer genes. When more public data becomes available the described approach and BPIM in particular can be validated.

ACKNOWLEDGEMENT

This research is supported by: the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, IDO (Genetic networks), several PhD/postdoc and fellow grants Flemish Government: FWO PhD/postdoc grants, projects G.0407.02 (support vector machines), G.0413.03 (inference in bioi), G.0388.03 (microarrays for clinical use), G.0229.03 (ontologies in bioi), G.0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), research communities (ICCoS, ANMMM, MLDM); IWT PhD Grants, GBOU-McKnow (Knowledge management algorithms), GBOU-SQUAD (quorum sensing), GBOU-ANA (biosensors), TAD-BioScope, Silicos. Belgian Federal Science Policy Office: IUAP P5/22 ('Dynamical Systems and Control: Computation, Identification and Modelling, 2002-2006); EU-RTD: FP5-CAGE (Compendium of Arabidopsis Gene Expression); ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain

REFERENCES

Bhattacharjee,A., Richards,W., Staunton,J., Li,C., Monti,S., Vasa,P., Ladd,C., Beheshti,J., Bueno,R., Gillette,M., Loda,M., Weber,G., Mark,E., Lander,E., Wong,W., Johnson,B., Golub,T., Sugarbaker,D. and Meyerson,M. (2001) Classification of human lung carcinomas by mrna expression profiling reveals distinct adenoma subclasses. *PNAS*, **98**, 13790–5.

Blamey,R., Davies,C., Elston,C., Johnson,J., Haybittle,J. and Maynard,P. (1979) Prognostic factors in breast cancer - the formation of a prognostic index. *Clin Oncol*, **5**, 227–236.

Boyages,J., Chua,B., Taylor,R., Bilous,M., Salisbury,E., Wilcken,N. and Ung,O. (2002) Use of the st gallen classification for patietns with node-negative breast cancer may lead to overuse of adjuvant chemotherapy. *British journal of surgery*, **89**, 789–796.

Brown,P. and Botstein,D. (1999) Exploring the new world of the genome with dna microarrays. *Nature*, **21**, 33–7.

Cooper,G. and Herskovits,E. (1992) A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, **9**, 309 –347.

Edén,P., Ritz,C., Rose,C., Fern,M. and Peterson,C. (2004) Good old clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. *European journal of cancer*, **40**, 1837–1841.

Eifel,P., Axelson,J., Costa,J. and et al. (2001) National institutes of health consensus development conference statement: adjuvant therapy for breast cancer. *J Natl Cancer Inst*, **93**, 979–989.

Glenisson,P., Coessens,B., Vooren,S.V., Mathys,J., Moreau,Y. and Moor,B.D. (2004) Txtgate: profiling gene groups with text-based information. *Genome Biology*, **5**.

Goldhirsch,A., Glick,J., Gelber,R. and Senn,H. (1998) Meeting highlights: international consenus panel on the treatment of primary cancer. *J Natl Cancer Inst*, **90**, 1601–1608.

Golub,T., Slonim,D., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J., Coller,H., Loh,M., Downing,J., Caligiuri,M., Bloomfield,C. and Lander,E. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–7.

Hanley,J. and McNeil,B. (1983) A method of comparing the areas under receiver operating characteristics curves derived from the same cases. *Radiology*, **148**, 839–43.

Heckerman,D., Geiger,D. and Chickering,D. (1995) Learning bayesian networks: the combination of knowledge and statistical data. *Machine Learning*, **20**, 197–243.

Husmeier,D., Dybowski,R. and Roberts,S., eds (2005) *Probabilistic modelling in bioinformatics and medical informatics*. Springer-Verlag, London, UK.

ITTACA (2006). http://bioinfo-out.curie.fr/ittaca.

Kaneda,A., Wakazono,K., Tsukamoto,T., Watanabe,N., Yagi,Y., Tatematsu,M., Kaminishi,M., Sugimura,T. and Ushijima,T. (2004) Lysyl oxidase is a tumor suppressor gene inactivated by methylation and loss of heterozygosity in human gastric cancers. *Cancer Res.*, **64**, 6410–6415.

Korb,K. and Nicholson,A. (2004) *Bayesian artificial intelligence*. Chapman and Hall, Boca Raton, Florida.

Lee,A., Pinder,S., Macmillan,R., Mitchell,M., Ellis,I., Elston,C. and Blamey,R. (2006) Prognostic value of lymphovascular invasion in women with lymph node negative invasive breast carcinoma. *European journal of cancer*, **42**, 357–362.

Lockhart,D., Dong,H., Byrne,M., Follettie,M., Gallo,M., Chee,M., Mittman,M., Wang,C., Kobayashi,M., Horton,H. and Brown,E. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, **14**, 1675–80.

Malaney,S. and Daly,R. (2001) The ras signaling pathway in mammary tumorigenesis and metastasis. *J Mammary Gland Biol Neoplasia*, **6**, 101–113.

Neal,R. (1996) *Bayesian learning for neural networks*. Springer-Verlag, New York.

Neapolitan,R. (2004) *Learning Bayesian networks*. Prentice Hall, Upper Saddle River, NJ.

Owen,J., Iragavarapu-Charyulu,V. and Lopez,D. (2004) T cell-derived matrix metalloproteinase-9 in breast cancer: friend or foe? *Breast Dis*, **20**, 145–153.

Pearl,J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Matteo, California.

Pecorino,L. (2005) *Molecular biology of cancer*. Oxford university press, New York.

Pochet,N., Smet,F.D., Suykens,J. and Moor,B.D. (2004) Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. *Bioinformatics*, **20**, 3185–95.

Shedden,K., Taylor,J., Giordano,T., Kuick,R., Misek,D., Rennert,G., Schwartz,D., Gruber,S., Logsdon,C., Simeone,D., Kardia,S., Greenson,J., Cho,K., Beer,D., Fearon,E. and Hanash,S. (2003) Accurate molecular classification of human cancers based on gene expression using a simple classifier with a pathological tree-based framework. *American journal of pathology*, **163**, 1985–1995.

Singh,D., Febbo,P., Ross,K., Jackson,D., Manola,J., Ladd,C., Tamayo,P., Renshaw,A., D'Amico,A., Richie,J., Lander,E., Loda,M., Kantoff,P., Golub,T. and Sellers,W. (2002) Gene expression correlates of clinical prostate cancer behaviour. *Cancer Cell*, **1**, 203–9.

Smet,F.D., Moreau,Y., Engelen,K., Timmerman,D., Vergote,I. and Moor,B.D. (2004) Balancing false positives and false negatives for the detection of differential expression in malignancies. *British Journal of Cancer*, **91**, 1160–1165.

Spentzos,D., Levine,D., Kolia,S., H,O., Boyd,J., Libermann,T. and Cannistra,S. (2005) Unique gene expression profile based on pathologic response in epithelial ovarian cancer. *J Clin Oncol.*, **23**, 7911–8.

Spentzos,D., Levine,D., Ramoni,M., Joseph,M., Gu,X., Boyd,J., Libermann,T. and Cannistra,S. (2004) Gene expression signature with independent prognostic significance in epithelial ovarian cancer. *J Clin Oncol.*, **22**, 4700–10.

Thijs,G., Moreau,Y., Smet,F.D., Mathys,J., Lescot,M., Rombauts,S., Rouze,P., B,B.D.M. and Marchal,K. (2002) Inclusive: integrated clustering, upstream sequence retrieval and motif sampling. *Bioinformatics*, **18**, 331–2.

Todd,J., Dowle,C., Williams,M., Elston,C., Ellis,I., Hinton,C., Blamey,R. and Haybittle,J. (1987) Confirmation of a prognostic index in primary breast cancer. *Br J Cancer*, **56**, 489–492.

Troyanskaya,O., Cantor,M., Sherlock,G., Brown,P., Hastie,T., Tibshirani,R., Botstein,D. and Altman,R. (2001) Missing value estimation methods for dna microarrays. *Bioinformatics*, **17**, 520–525.

van de Vijver,M., He,Y., van t Veer,L., Dai,H., Hart,A., Voskuil,D., Schreiber,G., Peterse,J., Roberts,C., Marton,M., Parrish,M., Atsma,D., Witteveen,A., Glas,A., Delahaye,L., van der Velde,T., Bartelink,H., Rodenhuis,S., Rutgers,E., Friend,S. and Bernards,R. (2002) A gene expression signature as a predictor of survival in breast cancer. *The new England journal of medicine*, **347**, 1999–2009.

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**, 530–536.