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

Olivier Gevaert¹,*; Frank De Smet¹,²; Dirk Timmerman³; Yves Moreau¹ and Bart De Moor¹

¹Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium; ²Medical Direction, National Alliance of Christian Mutualities, Haachtsesteenweg 579, 1031 Brussel, Belgium; and ³Department 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 data on 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)).

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

© The Author 2006. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oxfordjournals.org
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 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. A variable in a Bayesian network is conditionally independent of its parents and its children other parents. Here the Markov blanket variables are shown in a grey circle.

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:

$$p(S|D) \propto p(S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \left[ \frac{\Gamma(N_{ij})}{\Gamma(N_{ij} + N_{ijk})} \right]^{n_{ijk} - 1} \frac{1}{\Gamma(N_{ij} + N_{ijk})},$$

with $N_{ij}$ 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: $N_{ij} = \sum_{k=1}^{q_i} N_{ijk}$. $N_{ij}$ and $N_{ijk}$ have similar meanings but refer to prior knowledge for the parameters. When no knowledge is available they are estimated using $N_{ijk} = N(i|j)$ ( 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: $p(S) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} p(l_i = x_i) \prod_{k=1}^{n} p(m_k = x_k) / p_l$, 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 $p(ab)$ 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

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.
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:

\[ p(\theta_j | S) = \text{Dir}(\theta_j | N_{ij1}, \ldots, N_{ijk}, \ldots, N_{ijp_j}) \]  

(3)

with \( \theta_j \) a parameter set where \( i \) refers to the variable and \( j \) to the \( j \)-th instantiation of the parents in the current structure. \( \theta_j \) contains a probability for every value of the variable \( x_j \) given the current instantiation of the parents. \( \text{Dir} \) corresponds to the Dirichlet distribution with \( (N_{ij1}, \ldots, N_{ijp_j}) \) 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,

\[ p(\theta_j | D, S) = \text{Dir}(\theta_j | N'_{ij1} + N_{ij1}, \ldots, N'_{ijk} + N_{ijk}, \ldots, N'_{ijp_j} + N_{ijp_j}) \]  

(4)

with \( N'_{ij} \) 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.ncbi.nlm.nih.gov/geo with 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 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 \( \geq 0.3 \) or \( \leq -0.3 \). 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 put 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 facts 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).

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)).

Next we chose an operating point for BDIM and BPIM by choosing a threshold that corresponds with a maximum for 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 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. Both Tables 2 and 3 show that BPIM and the 70 genes can reliably be used to predict the prognosis in lymph node negative breast cancer.

Figure 3 shows the complete network built with partial integration. The outcome variable and its Markov Blanket is indicated as 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. The number of patients assigned a poor prognosis for the complete test set and for the true poor and good prognosis patients.

<table>
<thead>
<tr>
<th>Method</th>
<th>Total test set (n = 19)</th>
<th>Metastasis within 5 yr (n = 12)</th>
<th>Disease free at 5 yr (n = 7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>St Gallen 1998</td>
<td>13/19 (68%)</td>
<td>10/12 (83%)</td>
<td>3/7 (43%)</td>
</tr>
<tr>
<td>NIH 2000*</td>
<td>15/19 (79%)</td>
<td>10/12 (83%)</td>
<td>5/7 (71%)</td>
</tr>
<tr>
<td>NPI*</td>
<td>11/19 (58%)</td>
<td>9/12 (75%)</td>
<td>2/7 (29%)</td>
</tr>
<tr>
<td>70 genes†</td>
<td>14/19 (74%)</td>
<td>12/12 (100%)</td>
<td>2/7 (29%)</td>
</tr>
<tr>
<td>BPIM†</td>
<td>13/19 (68%)</td>
<td>11/12 (92%)</td>
<td>2/7 (29%)</td>
</tr>
<tr>
<td>BDIM†</td>
<td>11/19 (58%)</td>
<td>9/12 (75%)</td>
<td>2/7 (29%)</td>
</tr>
</tbody>
</table>

†The operating point is determined by maximizing the sum of the sensitivity and specificity on the training set.

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>average AUC</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical data</td>
<td>0.751</td>
<td>0.086</td>
</tr>
<tr>
<td>Microarray data</td>
<td>0.750</td>
<td>0.073</td>
</tr>
<tr>
<td>Decision integration</td>
<td>0.790</td>
<td>0.072</td>
</tr>
<tr>
<td>Partial integration</td>
<td>0.793</td>
<td>0.068</td>
</tr>
<tr>
<td>Full integration</td>
<td>0.747</td>
<td>0.099</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Total test set (n = 19)</th>
<th>Metastasis within 5 yr (n = 12)</th>
<th>Disease free at 5 yr (n = 7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>St Gallen 1998</td>
<td>13/19 (68%)</td>
<td>10/12 (83%)</td>
<td>3/7 (43%)</td>
</tr>
<tr>
<td>NIH 2000*</td>
<td>15/19 (79%)</td>
<td>10/12 (83%)</td>
<td>5/7 (71%)</td>
</tr>
<tr>
<td>NPI*</td>
<td>11/19 (58%)</td>
<td>9/12 (75%)</td>
<td>2/7 (29%)</td>
</tr>
<tr>
<td>70 genes†</td>
<td>14/19 (74%)</td>
<td>12/12 (100%)</td>
<td>2/7 (29%)</td>
</tr>
<tr>
<td>BPIM†</td>
<td>13/19 (68%)</td>
<td>11/12 (92%)</td>
<td>2/7 (29%)</td>
</tr>
<tr>
<td>BDIM†</td>
<td>11/19 (58%)</td>
<td>9/12 (75%)</td>
<td>2/7 (29%)</td>
</tr>
</tbody>
</table>

*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.
‡NIPI is the sum of 0.2 times the tumour diameter in cms, lymph node stage and the tumour grade.
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 \cite{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

---

**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.
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 (Goldhirsh 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, HMASLS 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 et al. (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.

Our ongoing research includes investigating 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.

ACKNOWLEDGEMENTS
This research is supported by: the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWV-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 inbioi), 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

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.

e189


