Bayesian integration of microarray and proteomics data to predict therapy response in patients with rectal cancer

Olivier Gevaert[1], Anneleen Daemen[1], Annelies Debucquoy[2], Jean-Pascal Machiels[3], Bart De Moor[1], and Karin Haustermans[2]

[1] Katholieke Universiteit Leuven, ESAT/Sista/BioI, Dept. Electrical Engineering, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
{olivier.gevaert,anneleen.daemen,bart.demoor}@esat.kuleuven.be
http://www.kuleuven.be/bioinformatics/

[2] Katholieke Universiteit Leuven, University Hospital Gasthuisberg, Dept. Radiation Oncology and Experimental Radiation, Herestraat 49, 3000 Leuven, Belgium
{annelies.debucquoy,karin.haustermans}@med.kuleuven.be

[3] Université Catholique de Louvain, St. Luc University Hospital, Dept. Medical Oncology, Ave. Hippocrate 10, 1200 Brussel, Belgium
{jean-pascal.machiels@onco.ucl.ac.be}

Abstract. *We developed a Bayesian integration framework to integrate high-dimensional and heterogenous data sources to predict the Rectal Cancer Regression Grade (RCRG) in rectal cancer patients. Fourty tumour and plasma samples were gathered at different time points during treatment to investigate the response of a combination therapy in the preoperative treatment of rectal cancer patients. These tumour and plasma samples were subjected at all time points to Affymetrix microarray and Luminex proteomics analysis, respectively. At surgery the RCRG was registered. When our framework was applied to this data set, we achieved an Area Under the ROC curve (AUC) of 0.73 for predicting RCRG when combining all data sources. Other data source combinations had worse AUC performance. When choosing a threshold maximizing accuracy, the sensitivity, specificity and accuracy where 92%, 30% and 75% respectively. Additionally, the posterior network distribution shows that only 5 genes and 4 proteins are necessary to predict RCRG. Moreover, among these genes and proteins all but one have strong links with tumorigenesis in the literature.*

Key words: rectal cancer, microarray, proteomics, Bayesian networks

1 Introduction

In the past decade microarray technology has had a big impact on cancer research. This technology allows to investigate the complete transcriptomic makeup of a tumour in one experiment. The data resulting from these experiments makes it possible for clinicians to make decisions regarding diagnosis or prognosis

based on the actual biological reasons why subsets of tumours behave differently, instead of relying on general clinical data or morphological characteristics of a tumour (i.e. grade, stage or immunohistochemistry of specific gene products). Although microarray technology hasn't reached the clinic yet, it has led to the discovery of new subclasses [1], gene expression patterns that can distinguish subclasses [2, 3] and in The Netherlands a large clinical trial is underway where a 70 gene expression profile [4, 5] for breast cancer is being collected.

The transcriptome however doesn't tell the full story. Post-translational modifications (e.g. phosphorylation, acetylation,etc.) and alternative splicing affect the final protein product. The previously adopted paradigm 'one gene - one protein' is not true, one gene can produce a variety of proteins. Gene expression, although it contains information on the tumour behaviour, can not completely reflect the final amount of protein that will be active in the cell. By studying the proteome one can obtain complementary information about the behavior of malignant processes [6] and uncover a previously unprobed part of the cancer biology puzzle. Additionally, incorporating the proteomic patterns in the clinical management of cancer could improve prediction of tumour behaviour and subsequently in distinguishing patients with different diagnosis, prognosis or response to therapy.

In this paper we present a Bayesian framework to integrate microarray and proteomics data to predict the behaviour of rectal tumours. This framework forms an extension of our method for integrating clinical and microarray data using Bayesian networks [7]. In this previous paper we showed that the partial integration method whereby each data source is first modeled separately before integrating them, is superior then other integration methods. Here, this idea is extended in two aspects. Firstly, instead of learning only one model for each data source, we adopt a more Bayesian approach and learn a distribution of models for each data source separately. Secondly, each distribution is combined in a prior for the Bayesian network modeling all data sources instead of hard coding the structure as in [7]. Figure 1 shows the main steps of our integration framework which will be explained in detail in section 2.5.

Our framework is applied to a set of patients with rectal cancer subjected to a pre-operative combination therapy of cetuximab, capecitabine and radiotherapy. Fourty patients were recruited in this study and microarray and proteomics data were gathered at three timepoints (T0, T1 and T2) during therapy. At surgery, the Rectal Cancer Regression Grade (RCRG) was registred which includes a measurement of tumour response after preoperative therapy. In this paper, patients were divided into two groups according to this RCRG: the positive group (RCRG pos) corresponding to good responsiveness and the negative group (RCRG neg) corresponding to moderate responsiveness and poor responsiveness. Besides having both microarray and proteomics data for the same patients, both technologies were used at three timepoints. Since our goal is to predict the RCRG before or early in therapy, we will not use the microarray and proteomic data in the last timepoint (i.e. at surgery). This leaves us with four data sources that can be integrated and used for the prediction of RCRG:

microarray at T0, microarray data at T1, proteomics data at T0 and proteomics data at T1 (See Figure 1 and section 2.4).

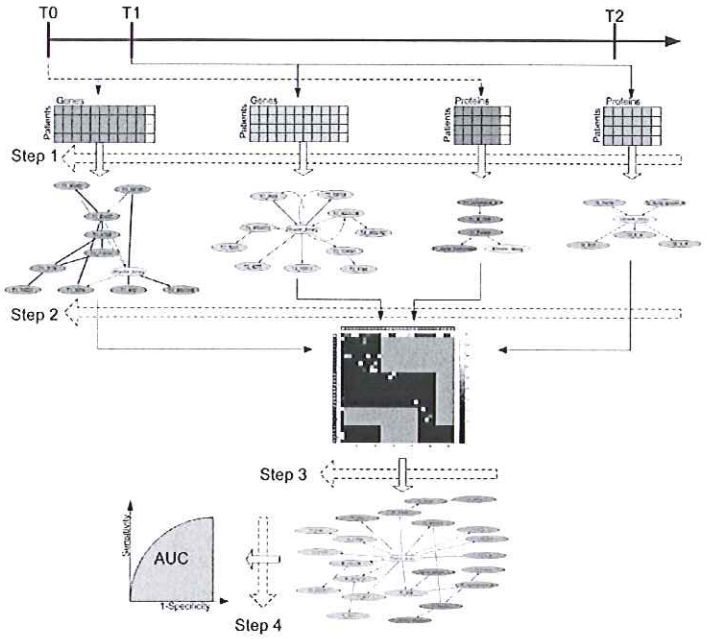


Fig. 1. Bayesian network integration framework: this figure explains the four steps involving data integration using Bayesian networks. The data sets are represented with matrices where rows correspond to patients and columns to genes or proteins. The yellow variable refers to the RCRG. Step 1 consist of learning a Bayesian network distribution for each data source separately by repeating the model building process 100 times. Step 2 involves integrating each separate distribution into the structure prior of a Bayesian network modeling all data sources. Each entry represents the probability a certain edge occurs. Step 3 involves learning a Bayesian network over all data sources using the previously constructed structure prior. Finally, step 4 involves predicting the RCRG of a new patient.

2 Methods

2.1 Bayesian networks

A Bayesian network is a probabilistic model that consists of two parts: a dependency structure and local probability models [8, 9]. 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:

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

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.

2.2 Bayesian network learning

We already 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.
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 [10] in combination with the Bayesian Dirichlet (BD) scoring metric:

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

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 corresponds to 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 $N'_{ijk} = N/(r_i q_i)$ [11] 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 and will be discussed in section 2.3. 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.
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 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:

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

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:

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

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.

2.3 Structure prior

Previously two approaches have been used to define informative prior distributions over Bayesian network structures [12]. First there are penalization methods that start from a prior structure and score structures based on the difference with the prior structure [11]. 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:

$$p(S) = \prod_{i=1}^{n} p(Pa(x_i) \rightarrow (x_i)) \tag{5}$$

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:

$$p(Pa(x_i) \rightarrow x_i) = \prod_{p \in Pa(x_i)} p(p \rightarrow x_i) \prod_{y \notin Pa(x_i)} p(y \nrightarrow x_i) \tag{6}$$

where $\nrightarrow$ means no edge between y and x_i. These individual edge probabilities can be represented in a matrix such as in Step 2 of Figure 1 and can be used to guide model search through its influence on Equation 2.

2.4 Data

Fourty patients with rectal cancer (T3-T4 and/or N+) from seven Belgian centers were enrolled in a phase I/II study investigating the combination of cetuximab, capecitabine and radiotherapy in the preoperative treatment of patients with rectal cancer [13]. Tissue and plasma samples were gathered before treatment (T0), after one dose of cetuximab but before radiotherapy with capecitabine (T1) and at moment of surgery (T2). At all these three time points, the frozen tissues were used for Affymetrix microarray analysis while the plasma samples were used for Luminex proteomics analysis. Some patients had to be excluded which ultimately resulted in a data set containing 36 patients.
The tumour samples were hybridized to Affymetrix human U133 2.0 plus gene chip arrays. The resulting data was first preprocessed for each time point separately using RMA [14]. Secondly, the probe sets were mapped on Entrez Gene Ids by taking the median of all probe sets that matched on the same gene. Probe sets that matched on multiple genes were excluded and unknown probe sets were given an arbitrary Entrez Gene Id. This reduces the number of features from 54613 probe sets to 27650 genes. Next, a pre-filtering without reference to phenotype was used to reduce the number of genes by only retaining the genes with a variance in the top 25%. This reduces the number of microarray features at each time point to 6913 genes.
The proteomics data consist of 96 proteins, previously known to be involved in cancer, measured for all patients in a Luminex 100 instrument. Proteins that had absolute values above the detection limit in less than 20% of the samples were excluded for each time point separately. This results in the exclusion of six proteins at T0, four at T1 and six at T2. The proteomics expression values of transforming growth factor alpha (TGFα), which had too many values below the detection limit, were replaced by the results of ELISA tests performed at the Department of Experimental Oncology in Leuven. For the remaining proteins the missing values were replaced by half of the minimum detected for each protein over all samples, and values exceeding the upper limit were replaced by the upper limit value. Because most of the proteins had a positively skewed distribution, a log transformation (base 2) was performed. In this paper, only the data sets at T0 and T1 were used because our goal is to predict the RCRG before the start of therapy or early in therapy .

2.5 Integration framework

To integrate data sources, we built further on the partial integration method that we developed earlier [7]. Previously, each data source was modeled separately by a Bayesian network. Then, these networks were connected and the parameters of the full network were learned. In this paper we extend this idea by developing it in a more Bayesian manner. Figure 1 shows the different steps of the algorithm in more detail. The data sets are represented with matrices where rows correspond to patients and columns to genes or proteins. Note that the microarray data sets are represented with larger matrices than the proteomics data

sets to stress their different dimensionality. There are four steps involving our data integration metholodyg using Bayesian networks. First, instead of learning one network per data source, a distribution of Bayesian networks is constructed (Step 1 in Figure 1). This is done by repeating the model building process a fixed number of times (e.g. 100). Each edge then receives a probability depending on the number of times it occurs in all runs. Next, the information from each of these Bayesian network distributions is encoded in a structure prior modeling all data sources (Step 2 in Figure 1). The probability of having an edge between two variables is represented with colors ranging from black (i.e. probalitity=0) to white (i.e. probability=1). This structure prior only contains prior probabilities between variables within each data source, between data sources nothing is known based on Step 1 of the algorithm. However, some information is included here, the black regions between data sources have been forced to zero such that links that go back in time are penalized. The orange regions correspond to links between data sources going forward in time and have maximum uncertainty (i.e. probability=0.5). The following step consists of learning a network with all data sources using the previously constructed structure prior (Step 3). Finally, this structure can be used for predicting new cases (Step 4 in Figure 1). To accomplish this we use the junction tree algorithm [15]. In our case we focus on the prediction of the RCRG of a patient which is represented by the yellow variable in Figure 1.

2.6 Model evaluation

The Area Under the ROC curve (AUC) was used to assess the prediction performance of the integration framework. Due to the limited data set size, a Leave-One-Out Cross Validation (LOOCV) approach was used to estimate the generalization performance. In each LOOCV run, not using any information of the left-out sample, the top 50 genes and the top 10 proteins with the largest differential expression between the RCRG groups are selected (Wilcoxon rank sum test). Then, Step 1 to 3 of the integration method are performed. Finally the model resulting from step 3 is used for predicting the RCRG of the left-out sample (Step 4).

Table 1. Area Under the ROC curve for all models when predicting RCRG.

	ALL	MPT0	MPT1	MT0T1	PT0T1	MT0	MT1	PT0	PT1
AUC	0.73	0.23	0.67	0.54	0.55	0.41	0.55	0.49	0.57

3 Results

Two data sources were available at two timepoints, therefore we examined the following combinations: all data sources (ALL), microarray and proteomics data at T0 (MPT0), microarray and proteomics data at T1 (MPT1), microarray data at T0 and T1 (MT0T1), proteomics data at T0 and T1 (PT0T1). In addition, all data sources were also modeled separately: microarray data at T0 (MT0), microarray data at T1 (MT1), proteomics data at T0 (PT0) and proteomics data at T1 (PT1). Table 2.6 shows the AUC for all models. The ALL model, which contains all data sources performs best followed by MPT1. After choosing an operating point optimizing accuracy, this results in a sensitivity of 92.3%, a specificity of 30% and an accuracy of 75%.

Next, we gathered the networks from all LOOCV iterations and represented them as a combined network where each edge receives a weight depending on the number of times this edge occurs in all LOOCV iterations (see Figure 3, left panel). Edges with high weight are represented with tick lines whereas edges with low weight are thin. Figure 3 shows that there is a ring of genes and proteins which have low edge weights and are most likely spuriously related. Next, there are two big clusters in the network corresponding to the genes at T0 and at T1. The RCRG is in between these two large clusters. The proteins are scattered throughout the network and have strong links with RCRG. Table 3 shows the links with the highest weights, for example for the protein Ferritin there is a strong link between the expression at both time points. Additionally, the proteins IL6 and Ferritin have strong links with the RCRG suggesting that they are important for prediction of response to therapy and more importantly that they are robust in the LOOCV iterations.

Finally, we investigated the variables that are important for prediction of RCRG

Table 2. Edges with the highest weights from the two networks in Figure 3. The left column shows the 10 edges which occurred the most in all LOOCV iterations. The right column shows the 10 edges which are linked with the RCRG and have the highest posterior probability based on the full data set.

LOOCV edges	RCRG edges
T2 Ferritin ↔ T1 Ferritin	T1 Ferritin
T2 232144_at ↔ T2 KCNJ2	T1 TNFα
T1 ETV6 ↔ T1 NFKB1	T2 IL6
T2 240084_at ↔ T2 RGS18	T1 GPR35
T2 IL6 ↔ RCRG	T1 LGR6
T2 Ferritin ↔ RCRG	T2 Ferritin
T1 ETV6 ↔ T1 238787_at	T2 MICA
T2 237622_at ↔ T2 USP42	T1 231037_at
T2 ST8SIA4 ↔ T2 RGS18	T2 NLRC5
T1 C8orf55 ↔ T1 KIFC2	T2 Apolipoprotein H

in the network using the complete data set. Instead of looking only at one network, we used the posterior Bayesian network distribution to estimate the posterior probability that a gene or protein is related to outcome. This was done by sampling 100 networks from this distribution and this network is represented in Figure 3 right panel. Table 3 shows the top 10 links with RCRG from this network.

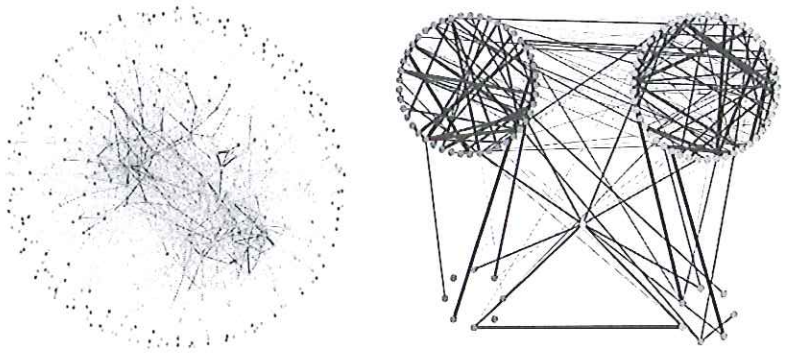


Fig. 2. Integrated network representing the networks from all LOOCV iterations on the left and RCRG network on the full data set on the right. See Table 3 for the important edges in these networks. The edge thickness refers to the uncertainty of each edge. The colors refer to the different data sources: blue = microarray T0, green = microarray T1, purple = proteomics T0, red = proteomics T1, yellow= RCRG.

4 Discussion

In this paper, we have developed a Bayesian framework for integrating high dimensional and heterogeneous data sources focussing on the prediction of therapy response of rectal cancer patients. By integrating microarray and proteomics data at different time points, a better prediction could be made than any data source alone or a combination of a few data sources. Additionally, when comparing the AUC of MPT0 and MPT1 (see Table 2.6), the results showed that data at the second time point, was more informative to predict RCRG. This indicates that administering a loading dose of cetuximab causes changes in gene and protein expression that reflect the RCRG.

Table 2 showed the genes and proteins linked with RCRG of which many have strong links in the literature. Several of the proteins have known associations with rectal and colon cancer, such as ferritin, TNFα and IL6. Ferritin, the major intracellular iron storage protein, is an indicator for iron deficiency anemia. This disease is recognized as a presenting feature of right-sided colon cancer and increases in men significantly the risk of having colon cancer [16]. Next,

the tumour necrosis factor TNFα has important roles in immunity and cellular remodeling and influences apoptosis and cell survival. Deregulation and especially overproduction of TNFα have been observed to occur in colorectal cancer [17]. Finally, IL-6 regulates the immune response, modulates normal and cancer cell growth, differentiation and cell survival [18]. It causes increased steady-state levels of TGFα mRNA in macrophage-like cells [19].
Additionally, several of the genes in Table 2 have known associations with tumorigenesis. MICA is part of the MHC class I complex which plays a major role in immunity. Many cancers suffer from immune attacks due to these membrane proteins. They subsequently develop immuno-evasion strategies and hide the stress by repressing NKG2D ligands such as MICA. Otherwise these ligands would be expressed by cells undergoing neoplastic transformation and recognized by natural killer cells [20]. Next, 2 of the 4 genes, LGR6 and GPR35, are G protein-coupled receptors. Although nothing is known yet about these two individual genes, this class of membrane proteins has proven to induce tumorigenesis. They stimulate a large class of G proteins that function as mitogenic and anti-mitogenic signals and deregulation of this signalling has been observed in certain types of cancer cells (e.g. small cell lung carcinoma) [20].
These examples provide strong evidence that our methodology is able to find biologically related genes and proteins from a wealth of high throughput data gathered at different time points during therapy. A limitation of our current approach is that much data (i.e. many patients) is needed to learn complex models such as our Bayesian integration framework. This is due to the fact that our methodology not only attempts to classify new patients but also provides a probabilistic underlying model that can be interpreted and verified. However, a possible advantage of our method is that in its first step, where each data source is modeled separately, other data sets can be used studying a similar patient population. Only in the third step data is needed where all data sources are available for each individual patient. In this manner publicly available data sources can be used to refine the construction of the structure prior in step 2 of our algorithm.
Finally, our research on the incorporation of other sources of prior information (e.g. protein-protein interactions) easily fits in our Bayesian framework. Recently, we showed how the use of PUBMED abstracts in the structure prior of a Bayesian network can improve outcome prediction [21] which is highly complimentary to this work and may further improve predictive performance.

Acknowledgments. This research is supported by the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen) and GOA AMBioRICS, 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. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci U S A **96**(12) (June 1999) 6745–6750
2. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286** (1999) 531–7
3. Eschrich, S., Yang, I., Bloom, G., Kwong, K.Y., Boulware, D., Cantor, A., Coppola, D., Kruhffer, M., Aaltonen, L., Orntoft, T.F., Quackenbush, J., Yeatman, T.J.: Molecular staging for survival prediction of colorectal cancer patients. J Clin Oncol **23**(15) (May 2005) 3526–3535
4. van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., Friend, S.H.: Gene expression profiling predicts clinical outcome of breast cancer. Nature **415**(6871) (January 2002) 530–536
5. van de Vijver, M.J., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E.T., Friend, S.H., Bernards, R.: A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med **347**(25) (December 2002) 1999–2009
6. Pecorino, L.: Molecular biology of cancer. Oxford university press, New York (2005)
7. Gevaert, O., Smet, F.D., Timmerman, D., Moreau, Y., Moor, B.D.: Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. Bioinformatics **22**(14) (July 2006)
8. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers, San Matteo, California (1988)
9. Neapolitan, R.E.: Learning Bayesian networks. Prentice Hall, Upper Saddle River, NJ (2004)
10. Cooper, G.F., Herskovits, E.: A bayesian method for the induction of probabilistic networks from data. Machine Learning **9** (1992)
11. Heckerman, D., Geiger, D., Chickering, D.M.: Learning bayesian networks: The combination of knowledge and statistical data. Machine Learning **20** (1995) 197–243
12. Antal, P., Fannes, G., Timmerman, D., Moreau, Y., De Moor, B.: Using literature and data to learn bayesian networks as clinical models of ovarian tumours. Artif Intell Med **30** (2004) 257–81
13. Machiels, J.P., Sempoux, C., Scalliet, P., Coche, J.C., Humblet, Y., Van Cutsem, E., Kerger, J., Canon, J.L., Peeters, M., Aydin, S., Laurent, S., Kartheuser, A., Coster, B., Roels, S., Daisne, J.F., Honhon, B., Duck, L., Kirkove, C., Bonny, M.A., Haustermans, K.: Phase i/ii study of preoperative cetuximab, capecitabine, and external beam radiotherapy in patients with rectal cancer. Ann Oncol **18**(4) (April 2007) 738–744
14. Irizarry, R.A., Hobbs, B., Collin, F., Yasmin, Antonellis, K.J., Scherf, U., Speed, T.P.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostat **4**(2) (2003) 249–264

15. Huang, C., Darwiche, A.: Inference in belief networks: A procedural guide. International Journal of Approximate Reasoning **15**(3) (1996) 225–263
16. Raje, Durgesh, Mukhtar, Hasan, Oshowo, Ayo, Clark, I., Celia: What proportion of patients referred to secondary care with iron deficiency anemia have colon cancer? Diseases of the Colon and Rectum **50**(8) (August 2007) 1211–1214
17. Zins, K., Abraham, D., Sioud, M., Aharinejad, S.: Colon cancer cell-derived tumor necrosis factor-alpha mediates the tumor growth-promoting response in macrophages by up-regulating the colony-stimulating factor-1 pathway. Cancer Res **67**(3) (February 2007) 1038–1045
18. Lee, S.O., Chun, J.Y., Nadiminty, N., Lou, W., Gao, A.C.: Interleukin-6 undergoes transition from growth inhibitor associated with neuroendocrine differentiation to stimulator accompanied by androgen receptor activation during lncap prostate cancer cell progression. Prostate **67**(7) (May 2007) 764–773
19. Hallbeck, A.L., Walz, T.M., Wasteson, A.: Interleukin-6 enhances transforming growth factor-alpha mrna expression in macrophage-like human monocytoid (u-937-1) cells. Biosci Rep **21**(3) (June 2001) 325–339
20. Weinberg, R.A.: The biology of cancer. Garland Science, Taylor and Francis group. (2007)
21. Gevaert, O., Van Vooren, S., De Moor, B.: Integration of microarray and textual data improves the prognosis prediction of breast, lung and ovarian cancer patients. Pac Symp Biocomput (2008) 1–12