# International Workshop on Probabilistic Modelling in Computational Biology

## Probabilistic Methods for Active Learning and Data Integration in Computational Biology

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

Olivier Gevaert
Dept. Electrical Engineering/ESAT-Sista-BioI

# Introduction

- Microarray technology has had a great impact on cancer research
- In the past decade many studies have been published applying microarray data to breast cancer, ovarian cancer, lung cancer, …
- Pubmed: cancer AND microarrays
  - 6325 articles
  - First article in 1996 Nature Genetics

# Introduction

- However, most cancer studies focus only on microarray data …
- … while these data suffer from some disadvantages:
  - High dimensional and "much" data, however many variables and few observations (i.e. patients)
  - Low signal-to-noise ratio: e.g. accidental differential expression
  - Influence and difficulty of pre-processing: assumptions
  - Sample heterogenity

# Introduction

- In our opinion integration of other sources of information could alleviate these disadvantages
- Recently there has been a significant increase of publicly available databases:
  - Reactome
  - Transfac
  - IntAct
  - Biocarta
  - KEGG
- However still many knowledge is contained in publications in unstructured form
- … and not deposited in public databases where it can be easily used by algorithms

# Introduction

- Goal:
  - Mine the vast resource of literature abstracts
  - Transform it to the gene domain
  - Combine it with expression data
- How:
  - Probabilistic models provide a natural way to integrate prior information by using a prior over model space
  - More specifically:
    - Text information incorporated in the structure prior of a Bayesian network
    - Applied to predict the outcome of cancer patients

# Overview

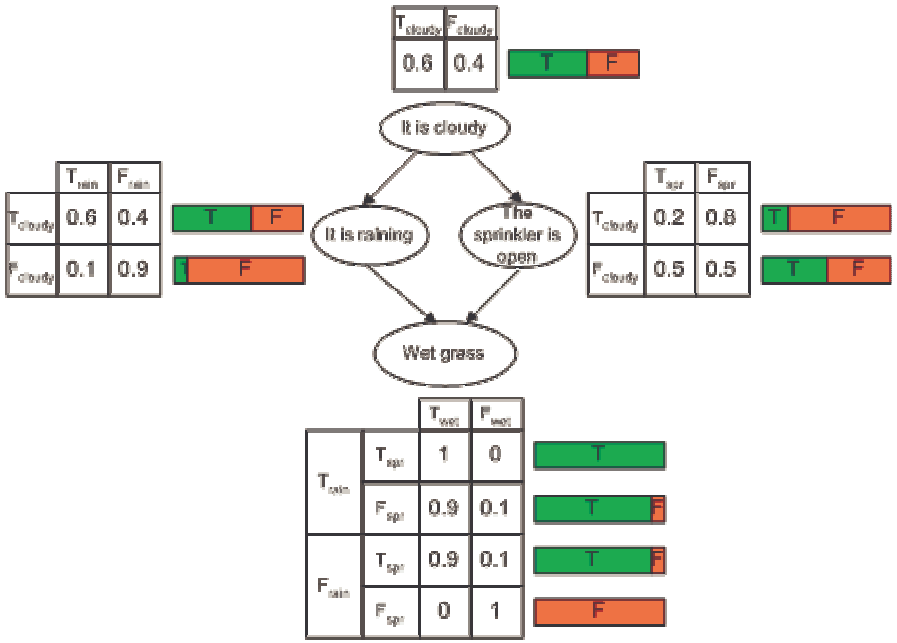
- Introduction
- Bayesian networks
- Structure prior
- Data
- Results
- Conclusions

# Overview

- Introduction
- Bayesian networks
- Structure prior
- Data
- Results
- Conclusions

# Bayesian networks

- Probabilistic model that consists of two parts:
  - Directed acyclic graph
  - Local probability models




$$p(x_1,\ldots,x_n)=\prod_{i=1}^{n} p(x_i \mid Pa(x_i))$$

# Bayesian networks

- Discrete or continuous variables
- Different local probability models
  - Discrete variables:
    - Conditional probability tables *Heckerman et al. Machine Learning 1995*
    - Noisy OR
    - Decision trees
  - Continuous variables:
    - Gaussian *Heckerman et al. Machine Learning 1995*
    - Non-parametric regression *Imoto et al. Journal of bioinformatics and computational biology 2003*
    - Neural networks

# Bayesian networks

- All these local probability models have different properties and (dis)advantages
- We chose discrete valued Bayesian networks because:
  - Exact computation
  - Non-linear (i.e. arbitrary discrete distributions can be represented)
  - Space of arbitrary non-linear continous distributions is very large
  - Limited data set size may not allow to infer non-linear continuously valued relations

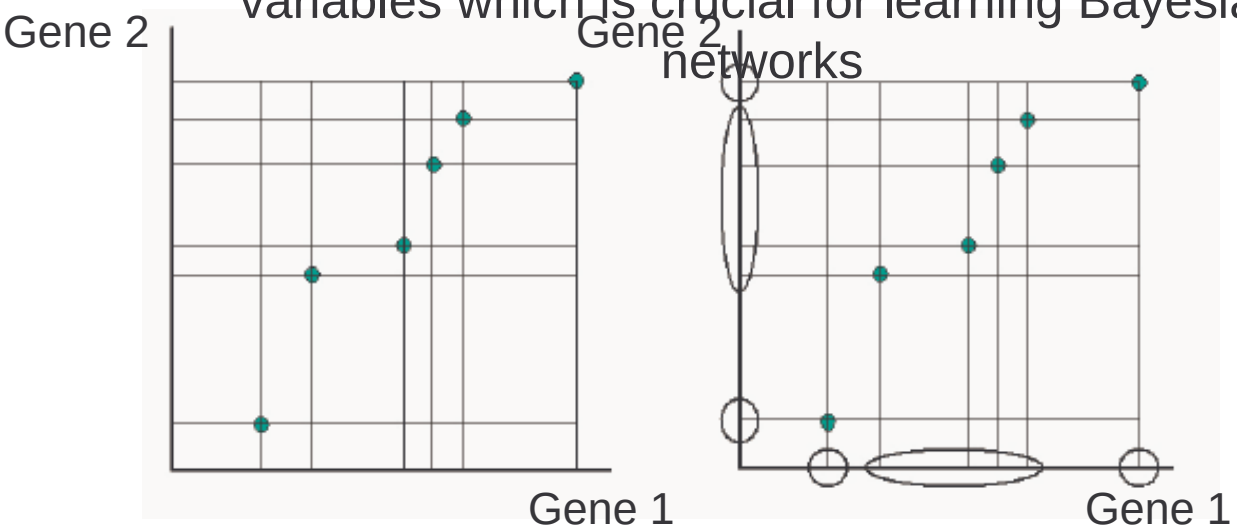
*Hartemink PhD thesis 2001*

# Discretization

Univariate discretization


Multivariate discretization

Problem: loose relationship between the variables which is crucial for learning Bayesian networks



# Discretization

- Multivariate discretization in three bins by:
  - First simple discretization method with a large number of bins (interval discretization or quantile discretization)
  - Join bins where Mutual information decreases the least
  - Iterate algorithm untill each gene has three bins

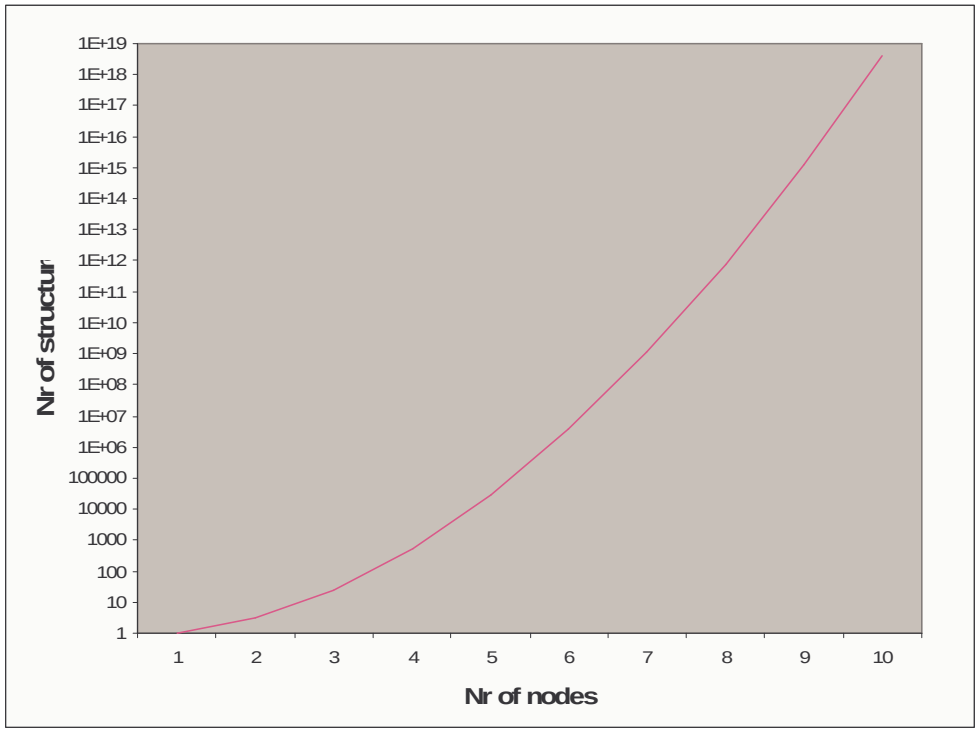
*Hartemink PhD thesis 2001*

# Bayesian networks

- Bayesian network consists of two parts a DAG and CPTs
- … thus model estimation in two steps:
  - Structure learning
  - Parameter learning

# Bayesian networks

- Mostly the structure is unknown and has to be learned from data
- Exhaustively searching for all structures is impossible
- As number of nodes increases, the number of structures to evaluate increases super-exponentially:



# Bayesian networks

- K2 algorithm *Cooper & Herskovits Machine learning 1992*
  - Greedy search
    - ordering to restrict possible structures
    - suboptimal
  - Scoring metric
    - Scores a specific structure that was chosen by the search procedure
    - Bayesian Dirichlet score

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

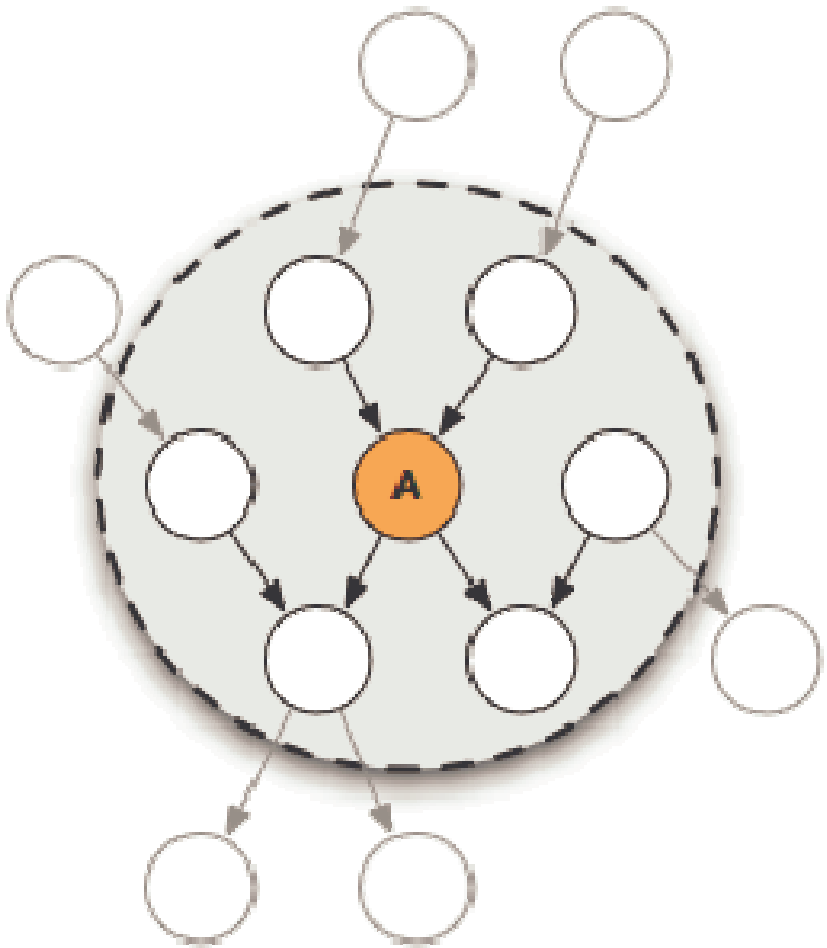
# Bayesian networks

- Parameter learning
  - Straightforward updating the dirichlet priors
  - i.e. counting the number of times a specific situation occurs

$$p(\theta_{ij}|S) = Dir(\theta_{ij}|N'_{ij1},\ldots,N'_{ijr_i})$$

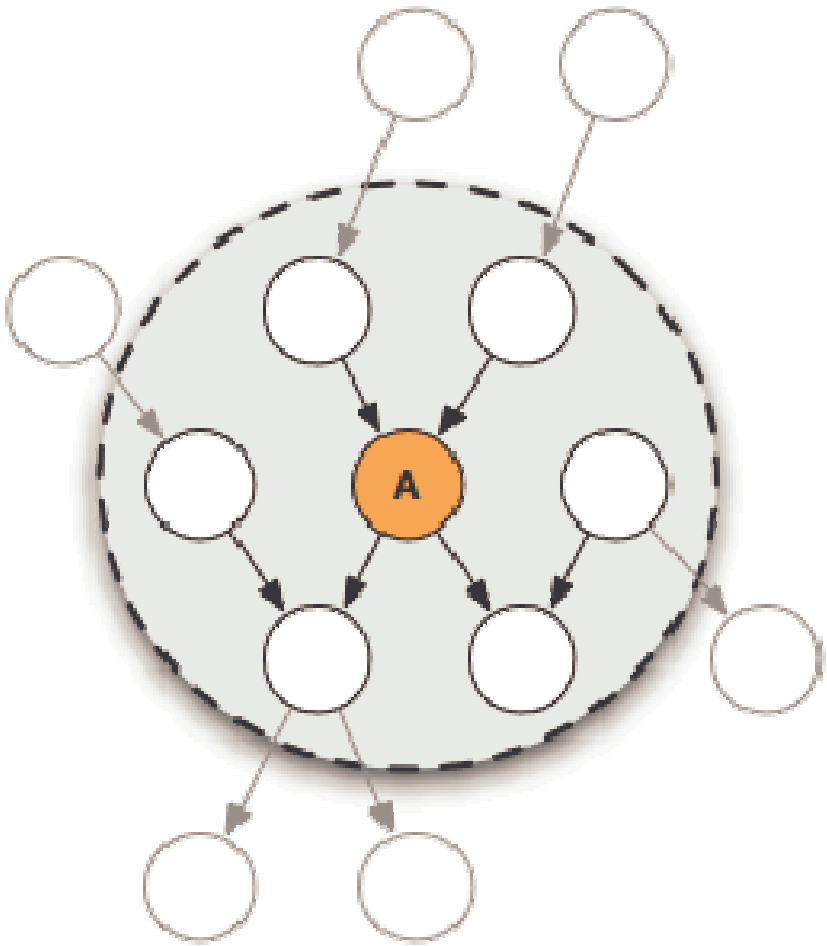
$$p(\theta_{ij}|D,S) = Dir(\theta_{ij}|N'_{ij1}+N_{ij1},\ldots,N'_{ijr_i}+N_{ijr_i})$$

# Markov blanket




- The set of variables that completely shields off a specific variable from the rest of the network
- Defined as
  - Its parents
  - Its children
  - Its children's other parents.

# Markov blanket




- Bayesian networks perform feature selection
- The Markov blanket variables influence the outcome directly …
- … and block the influence of other variables

# Overview

- Introduction
- Bayesian networks
- Structure prior
- Data
- Results
- Conclusions

# Structure prior

- Bayesian model building allows integration of prior information:
  - Structure prior
  - Parameter prior (not used here, uninformative prior)

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

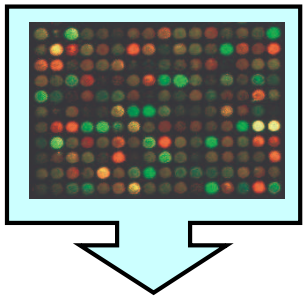
Structure prior

Parameter prior

Heckerman, Machine Learning, Vol. 20 (1995), pp. 197-243.

# Structure prior

**Microarray data**



**Prior**

PubMed

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

Posterior — Likelihood — Prior

# Structure prior

*How do we get the structure prior?*

- Two approaches have been used to define structure priors:
  - Penalization methods
    - Score structure based on difference with prior structure
  - *Pairwise methods*
    - Being a parent of a variable is independent of any other parental relation
- Our information is in the form of pairwise (gene-gene) similarities therefore we chose a pairwise method:
  - Structure prior then decomposes as:

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

# Structure prior

- The probability of a local structure is then calculated by:

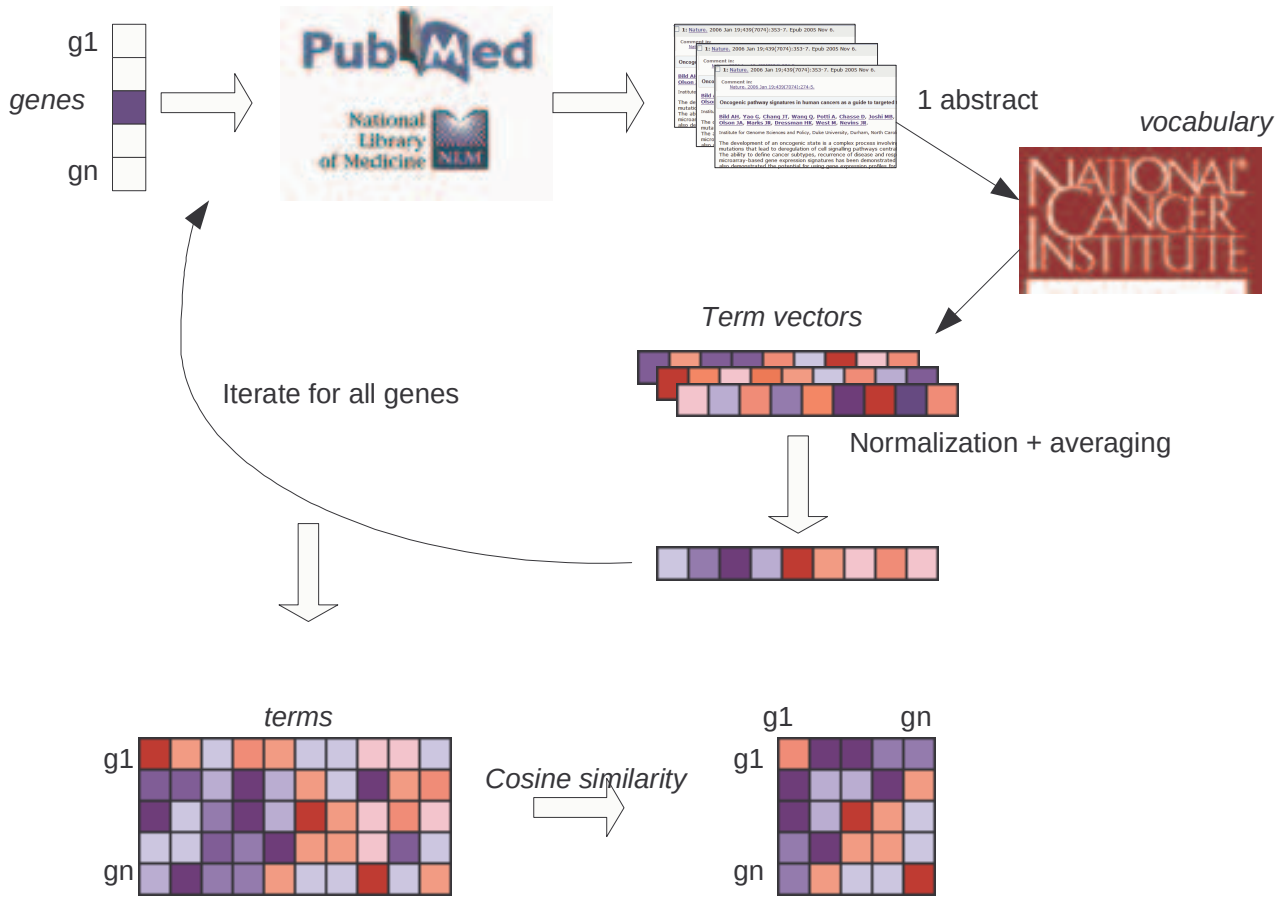
$$p(Pa(x_i) \rightarrow x_i) = \prod_{y \in Pa(x_i)} p(y \rightarrow x_i) \prod_{y \notin Pa(x_i)} p(y \otimes x_i)$$

- How do we get the $p(y \rightarrow x_i)$ and the $p(y \otimes x_i)$ ?
- … from PubMed

# Structure prior

- Genes $x_i$ are represented in the Vector Space Model
  - Each $x_{ij}$ corresponds to a term or phrase in a controlled vocabulary
  - We used the national cancer institute thesaurus
  - Using a fixed vocabulary has several advantages:
    - Simply using all terms would result in very large vectors, whereas use of only a small number of terms improves the **quality** of gene-gene similarities
    - Use of **phrases** reduces noise in the data set, as genes will only be compared from a domain specific view
    - Use of multi-word phrases without having to resort to **co-occurrence statistics** on the corpus to detect them
    - No need to filter **stop words**, only cancer specific terms are considered

# Structure prior



# Structure prior

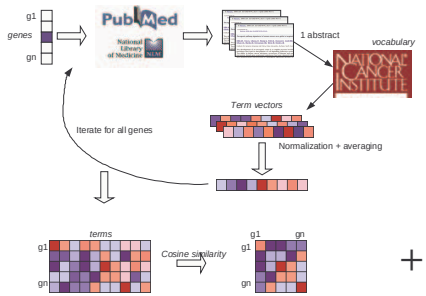
- Our goal is to predict the outcome of cancer patients
- One extra variable: outcome of the patient, e.g. survival in months, prognosis (good/poor), metastasis (yes/no)
- Therefore we need also a prior for the relationship gene ⇔ outcome
- Based on average relation between specific terms (outcome, survival, metastasis, recurrence, prognosis) and gene

# Structure prior

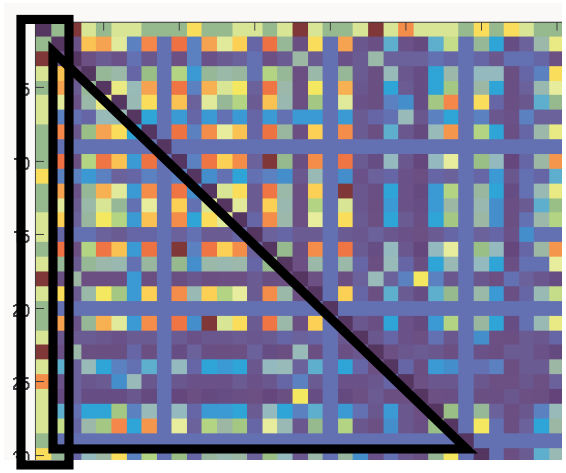
- Scaling
  - A fully connected Bayesian network can explain any data set but we want simple models
  - The prior contains many gene-gene similarities however we will not use them directly
    - We will introduce an extra parameter: mean density
    - "the average number of parents per variable"
    - Structure prior will be scaled according to this mean density
- Low mean density ⇒ less edges ⇒ less complex networks

# Structure prior

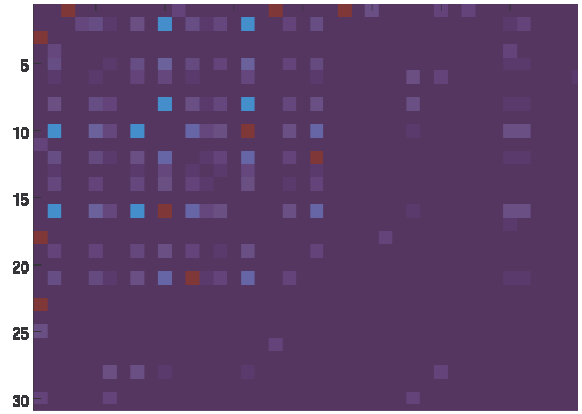
Summary



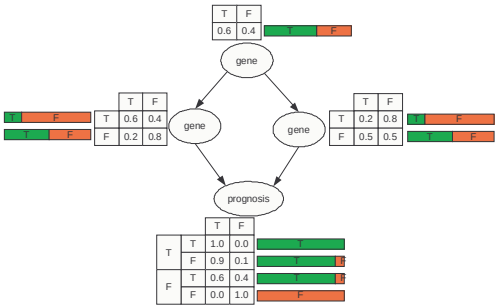

+ gene-outcome relationship



Scaling by mean density




Text prior



# Overview

- Introduction
- Bayesian networks
- Structure prior
- Data
- Results
- Conclusions

- Veer data:
  - 97 breast cancer patients belonging to two groups: poor and good prognosis
  - Preprocessing similar to original publication
  - 232 genes selected which correlated with outcome
- Bild data:
  - 3 data sets on breast, ovarian and lung cancer
    - 171 breast cancer patients
    - 147 ovarian cancer patients
    - 91 lung cancer patients
  - Outcome: survival of patients in months

# Evaluation of models

- 100 randomizations of the data with and without the text prior
  - 70% for training the model
  - 30% for estimating the generalization performance
- Area under the ROC curve is used as performance measure
- Wilcoxon rank sum test to assess statistical significance
  - P-value < 0.05 is considered statistically significant

# Overview

- Introduction
- Bayesian networks
- Structure prior
- Data
- Results
- Conclusions

# Results

- Veer data:

| Mean density | Text prior mean AUC | Uniform prior mean AUC | P-value |
|---|---|---|---|
| 1 | 0.80 (0.08) | 0.75(0.08) | 0.000396[§] |
| 2 | 0.80 (0.08) | 0.75(0.07) | <2e-06[§] |
| 3 | 0.79 (0.08) | 0.75(0.08) | 0.00577[§] |
| 4 | 0.79 (0.07) | 0.74(0.08) | <6e-06[§] |

Average number of parents per variable

# Markov blanket

- Next, we build a model with and without the text prior called TXTmodel and UNImodel resp.
- We investigated the Markov blanket of the outcome variable

# Results

- TXTmodel
  - Genes implicated in breast cancer
    - TP53, VEGF, MMP9, BIRC5, ADM, CA9
  - Weaker link
    - ACADS, NEO1, IHPK2
  - No association
    - MYLIP
- UNImodel
  - Breast cancer related
    - WISP1, FBXO31, IGFBP5, TP53
  - Other genes
    - Unknown or not related

| **TXTmodel** | | **UNImodel** | |
|---|---|---|---|
| Gene name | Text score | Gene name | Text Score |
| MYLIP | 0.58 | PEX12 | 0.58 |
| *TP53* | 1 | LOC643007 | 0.5 |
| ACADS | 0.58 | WISP1 | 0.75 |
| VEGF | 1 | SERF1A | 0.58 |
| ADM | 0.83 | QSER1 | 0.5 |
| NEO1 | 0.67 | ARL17P1 | 0.5 |
| *IHPK2* | 0.5 | LGP2 | 0.58 |
| CA9 | 1 | *IHPK2* | 0.5 |
| MMP9 | 1 | TSPYL5 | 0.5 |
| BIRC5 | 1 | FBXO31 | 0.58 |
| | | LAGE3 | 0.5 |
| | | IGFBP5 | 0.58 |
| | | AYTL2 | 0.5 |
| | | *TP53* | 1 |
| | | PIB5PA | 0.58 |
| Average text score | 0.85 | Average text score | 0.58 |

# Results

- Average text score of TXTmodel (0.85) is higher than UNImodel score (0.58) as expected
- TP53 and IHBK2 appear in both sets

| **TXTmodel** | | **UNImodel** | |
|---|---|---|---|
| Gene name | Text score | Gene name | Text Score |
| MYLIP | 0.58 | PEX12 | 0.58 |
| *TP53* | 1 | LOC643007 | 0.5 |
| ACADS | 0.58 | WISP1 | 0.75 |
| VEGF | 1 | SERF1A | 0.58 |
| ADM | 0.83 | QSER1 | 0.5 |
| NEO1 | 0.67 | ARL17P1 | 0.5 |
| *IHPK2* | 0.5 | LGP2 | 0.58 |
| CA9 | 1 | *IHPK2* | 0.5 |
| MMP9 | 1 | TSPYL5 | 0.5 |
| BIRC5 | 1 | FBXO31 | 0.58 |
| | | LAGE3 | 0.5 |
| | | IGFBP5 | 0.58 |
| | | AYTL2 | 0.5 |
| | | *TP53* | 1 |
| | | PIB5PA | 0.58 |
| Average text score | 0.85 | Average text score | 0.58 |

# Results

- Bild data
- Mean density is set to 1

| Data set | Text prior mean AUC | Uniform prior mean AUC | P-value |
|---|---|---|---|
| **Breast** | 0.79 | 0.75 | 0.00020 |
| **Ovarian** | 0.69 | 0.63 | 0.00002 |
| **Lung** | 0.76 | 0.74 | 0.02540 |

# Overview

- Introduction
- Bayesian networks
- Structure prior
- Data
- Results
- Conclusions

# Conclusions

- Verified the actual influence of the text prior:
  - Improves outcome prediction of cancer compared to not using a prior
    - Both on the initial data set and the validation data sets
  - Allows to select a set of genes (cfr. Markov blanket) based on both gene expression data and knowledge available in the literature related to cancer outcome

# Limitations

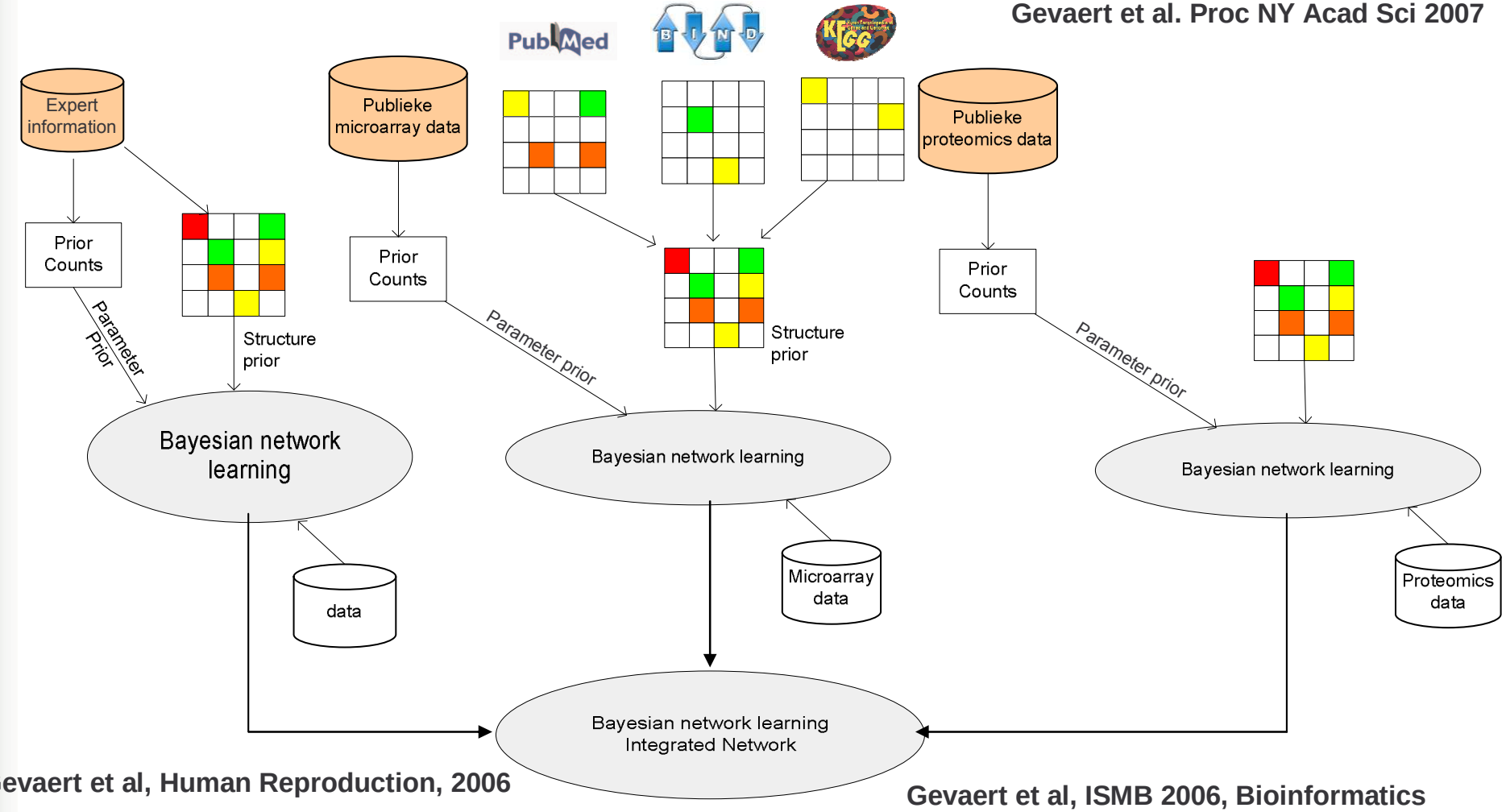
- Making the connection between the outcome and the genes in the prior is currently arbitrary
  - Investigating ways to automize it
  - E.g. Based on terms characterizing well known cancer genes
- No validation yet of the Markov blanket of important genes in the posterior network
  - No ground truth

# Future work

- Continually developing text prior
  - Gene name recognition in abstracts instead of manually curated references
  - Reduction of the literature to cancer related journals or abstracts mentioning "cancer"
- Adding other sources of information
  - Protein-DNA interactions (TRANSFAC)
  - Pathway information (KEGG, Biocarta)
- Long term goal:
  - Developing a framework for modeling regulatory networks behind cancer outcomes

# Future work

Gevaert et al. Proc NY Acad Sci 2007




Gevaert et al, Human Reproduction, 2006

Gevaert et al, ISMB 2006, Bioinformatics