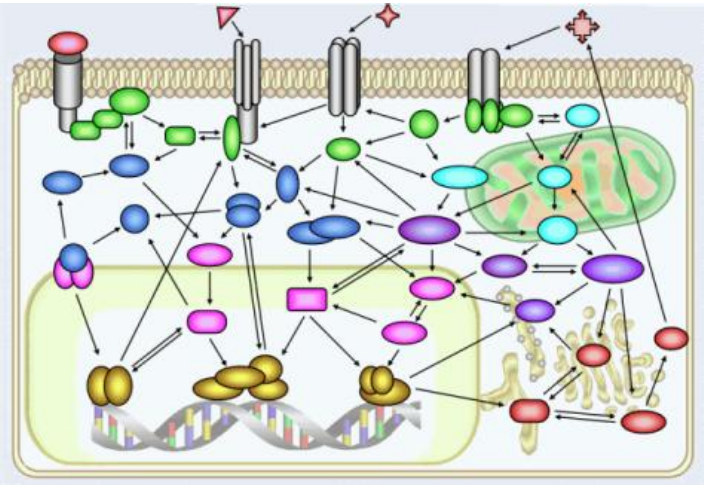
A Bayesian network integration framework for modeling biomedical data



Olivier Gevaert

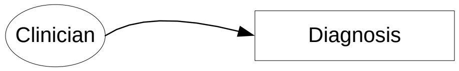
PhD defense

Overview

- Motivation
- Bayesian networks
- Results
  - Aim 1: modeling primary data
  - Aim 2: integrating primary data
  - Aim 3: integrating secondary data
- Conclusions
- Future work

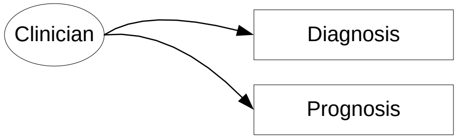
Motivation

- Clinicians have to make many decisions concerning the therapy of their patients e.g.:
  - Diagnosis



Motivation

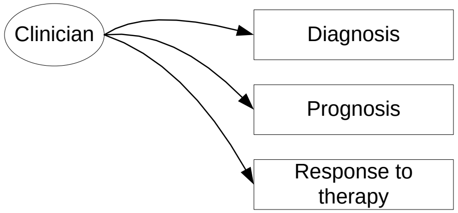
- Clinicians have to make many decisions concerning the therapy of their patients e.g.:
  - Diagnosis
  - Prognosis



Motivation

- Clinicians have to make many decisions concerning the therapy of their patients e.g.:
  - Diagnosis
  - Prognosis
  - Therapy response

- Based on **expertise**
- But often the clinician has



Clinician

Diagnosis

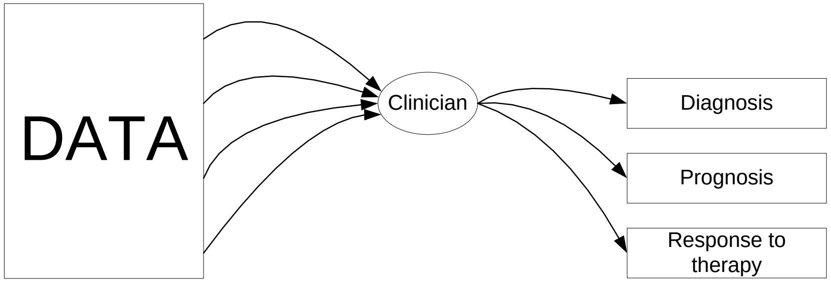
Prognosis

Response to therapy

Motivation

- Clinicians have to make many decisions concerning the therapy of their patients e.g.:
  - Diagnosis
  - Prognosis
  - Therapy response

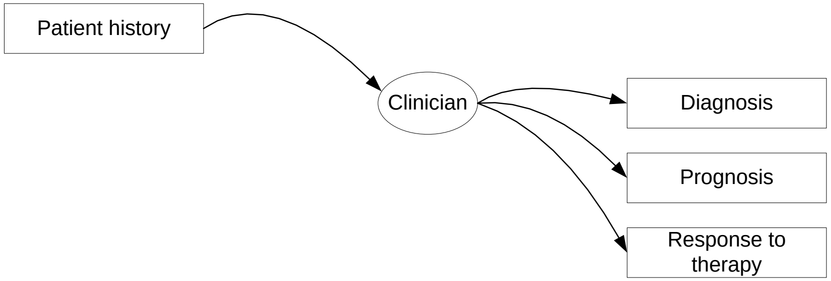
- Based on **expertise**
- But often the clinician has
  - Patient Data



Motivation

- Clinicians have to make many decisions concerning the therapy of their patients e.g.:
  - Diagnosis
  - Prognosis
  - Therapy response

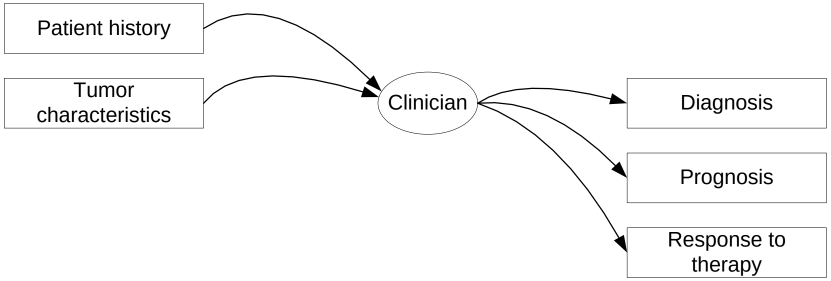
- Based on **expertise**
- But often the clinician has
  - Patient Data
    - Patient history



Motivation

- Clinicians have to make many decisions concerning the therapy of their patients e.g.:
  - Diagnosis
  - Prognosis
  - Therapy response

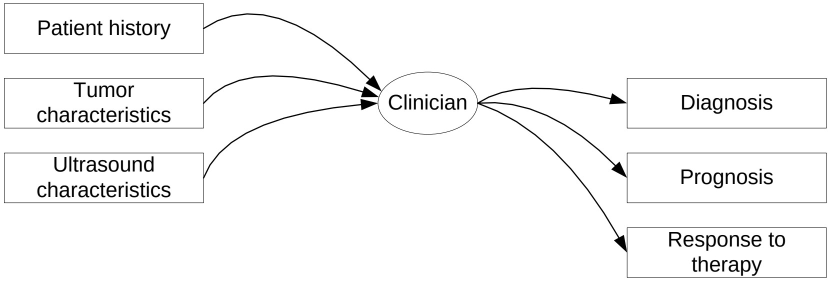
- Based on **expertise**
- But often the clinician has
  - Patient Data
    - Patient history
    - Tumor characteristics



Motivation

- Clinicians have to make many decisions concerning the therapy of their patients e.g.:
  - Diagnosis
  - Prognosis
  - Therapy response

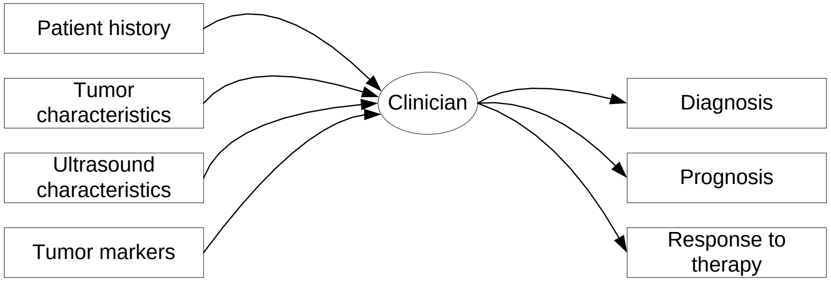
- Based on **expertise**
- But often the clinician has
  - Patient Data
    - Patient history
    - Tumor characteristics
    - Ultrasound characteristics



Motivation

- Clinicians have to make many decisions concerning the therapy of their patients e.g.:
  - Diagnosis
  - Prognosis
  - Therapy response

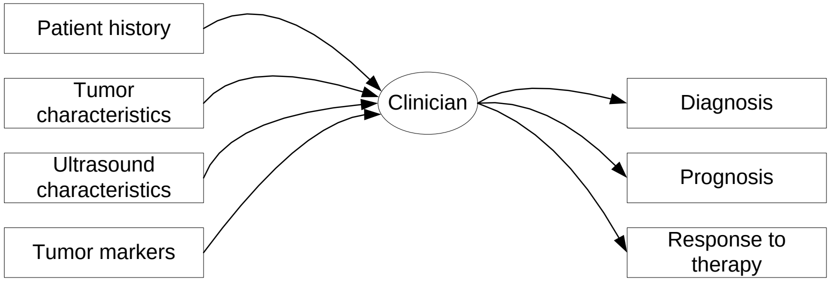
- Based on **expertise**
- But often the clinician has
  - Patient Data
    - Patient history
    - Tumor characteristics
    - Ultrasound characteristics
    - Tumor markers



Motivation

- Not all these data types are relevant for every disease
- But for example for the diagnosis of ovarian masses

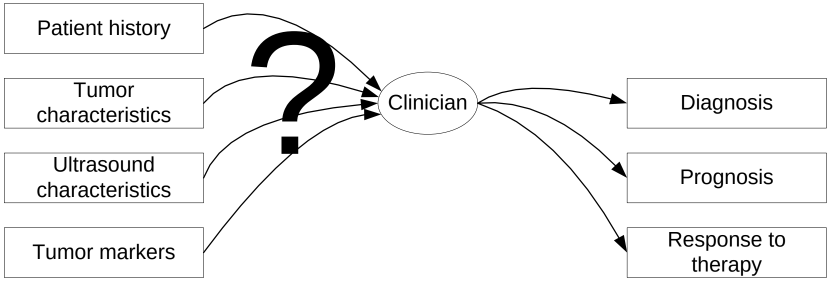
  many data types are suspected to be relevant
- And for many other diseases this is also the case



Motivation

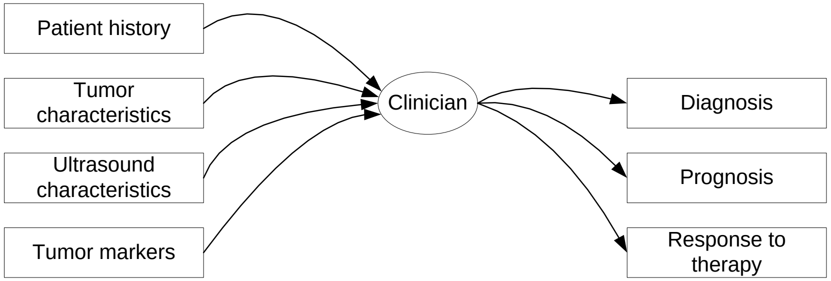
- Not all these data types are relevant for every disease
- But for example for the diagnosis of ovarian masses

  <u>many data types</u> are suspected to be relevant
- And for many other diseases this is also the case
- Problem
  - In many cases it is difficult for the clinician to interpret <u>all data</u> manually



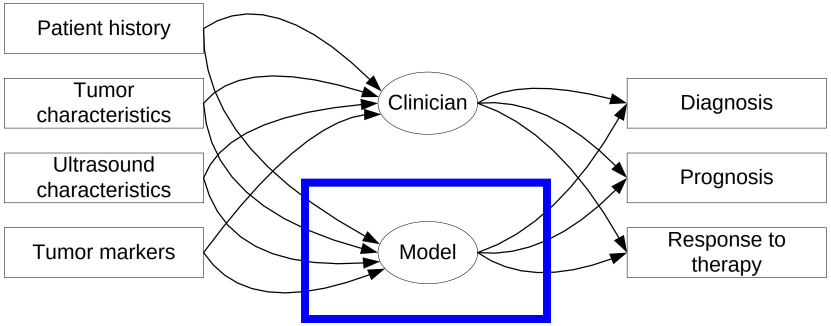
Motivation

- Solution:



Motivation

- Solution:
  - Medical decision support modeling
  - Building a mathematical model on the data
  - Use this model to predict patient outcome
    - Diagnosis
    - Prognosis
    - Therapy response

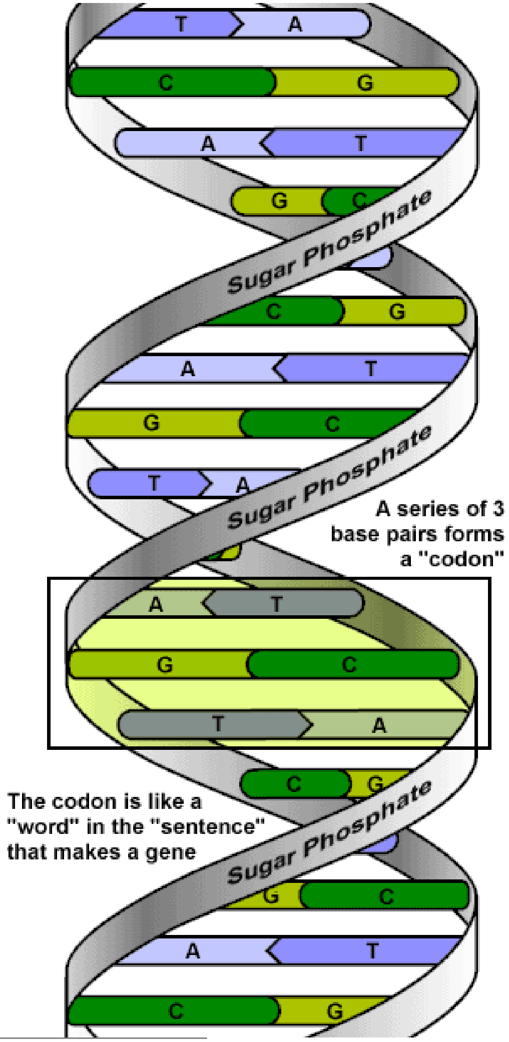


Medical decision support modeling

- History of more than 30 years
- Many different methods exist
  - Logistic regression
  - Artificial Neural networks
  - Support vector machines
  - Bayesian networks
  - …
- The general idea is the same
  - Assist clinicians when making decisions

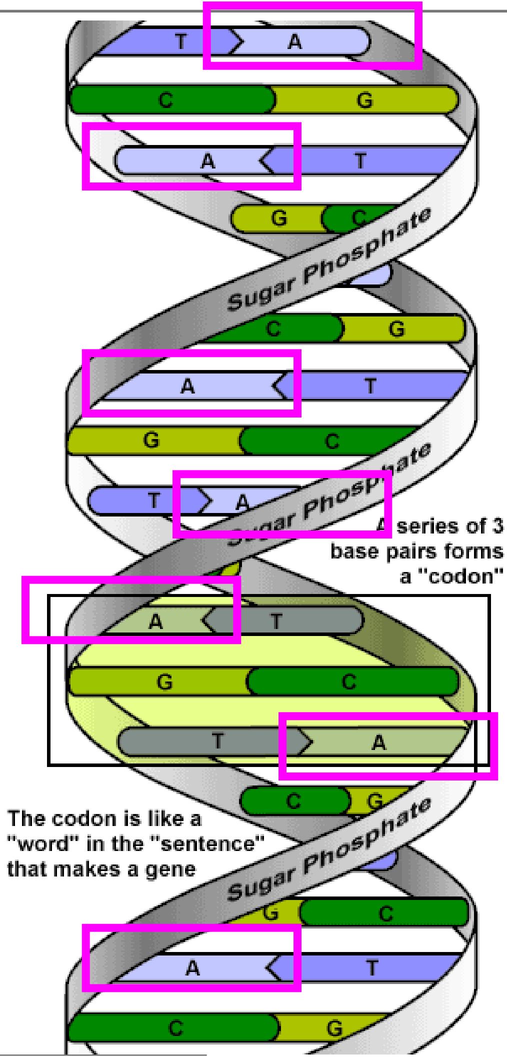
Microarray technology

- The rise of new technology changed medical decision support into **bio**medical decision support
- New technologies allow to gather biological data
- When studying **cancer**, this has particular advantages
  - Biological
  - Individualized
  - Genome-scale


T
A
C
G
A
T
G
C
Sugar Phosphate
C
G
A
T
G
C
T
A
Sugar Phosphate
A series of 3 base pairs forms a "codon"
A
T
G
C
T
A
C
G
The codon is like a "word" in the "sentence" that makes a gene
Sugar Phosphate
G
C
A
T
C
G

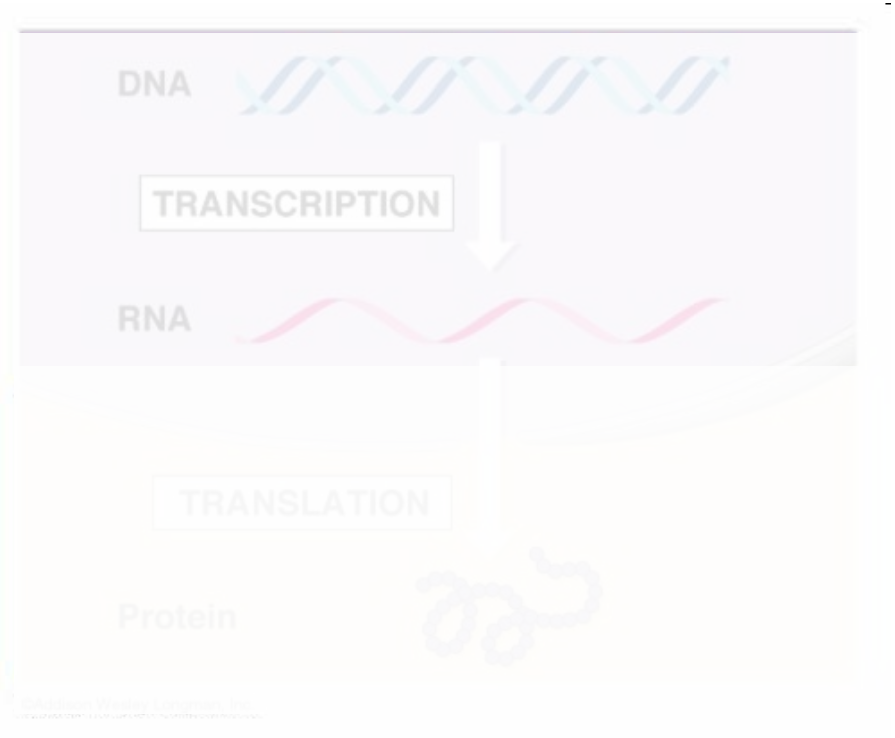
Molecular biology

- Short introduction in molecular biology
  - DNA consists of 4 bases
    - Adenine ⇒ A
    - Guanine ⇒ G
    - Cytosine ⇒ C
    - Thymine ⇒ T
  - Human DNA consists of a sequence of two times 3 billion of these bases


T
A
C
G
A
T
G
C
Sugar Phosphate
C
G
A
T
G
C
T
A
Sugar Phosphate
A series of 3 base pairs forms a "codon"
A
T
G
C
T
A
C
G
The codon is like a "word" in the "sentence" that makes a gene
Sugar Phosphate
G
C
A
T
C
G

Molecular biology

- DNA stores the genetic information in the form of genes
- Gene is a small piece of DNA
- Central dogma of molecular biology
  - Transcription
    - Gene ⇒ mRNA
  - Translation
    - mRNA ⇒ protein

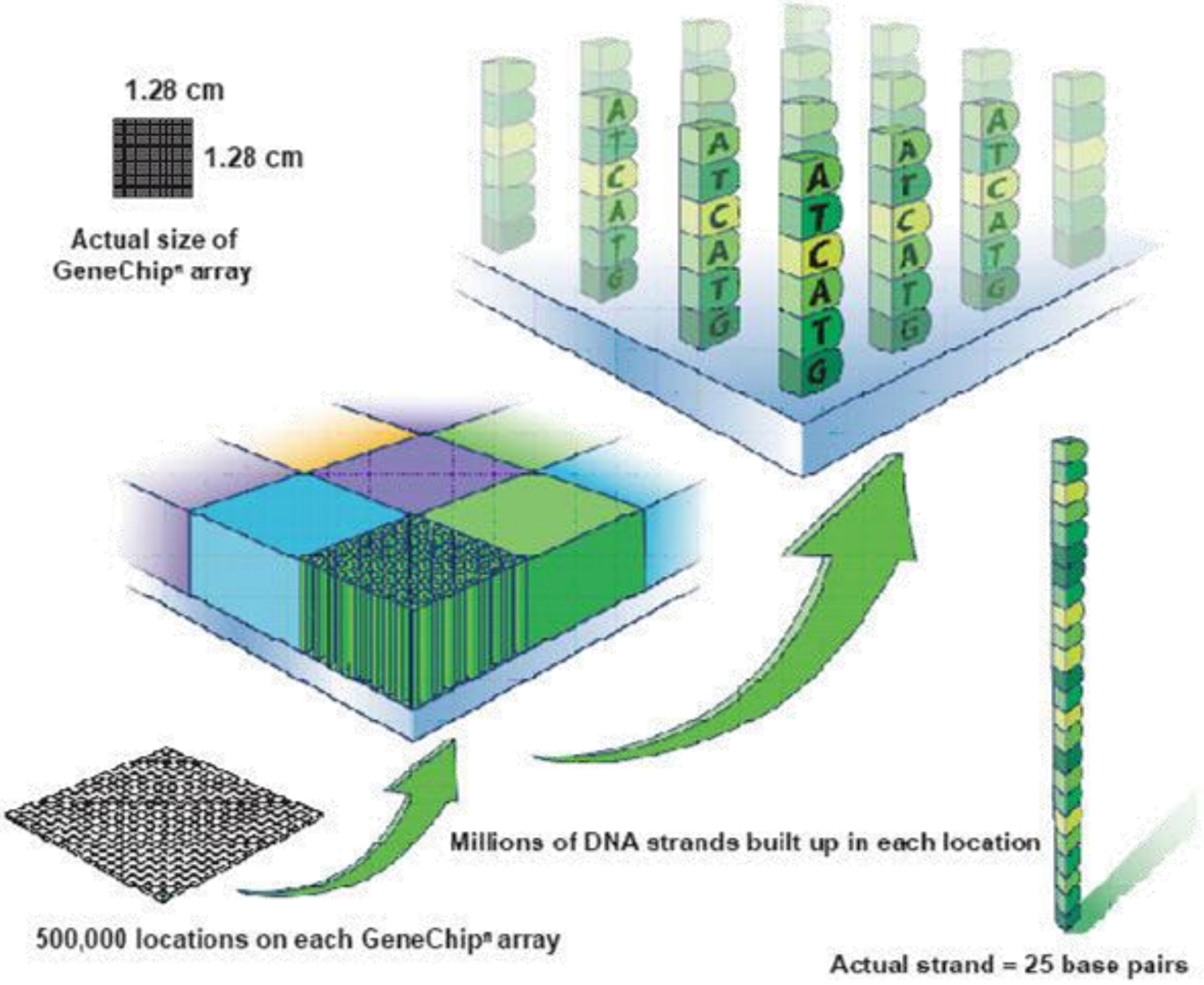


©Addison Wesley Longman, Inc.

The human genome project

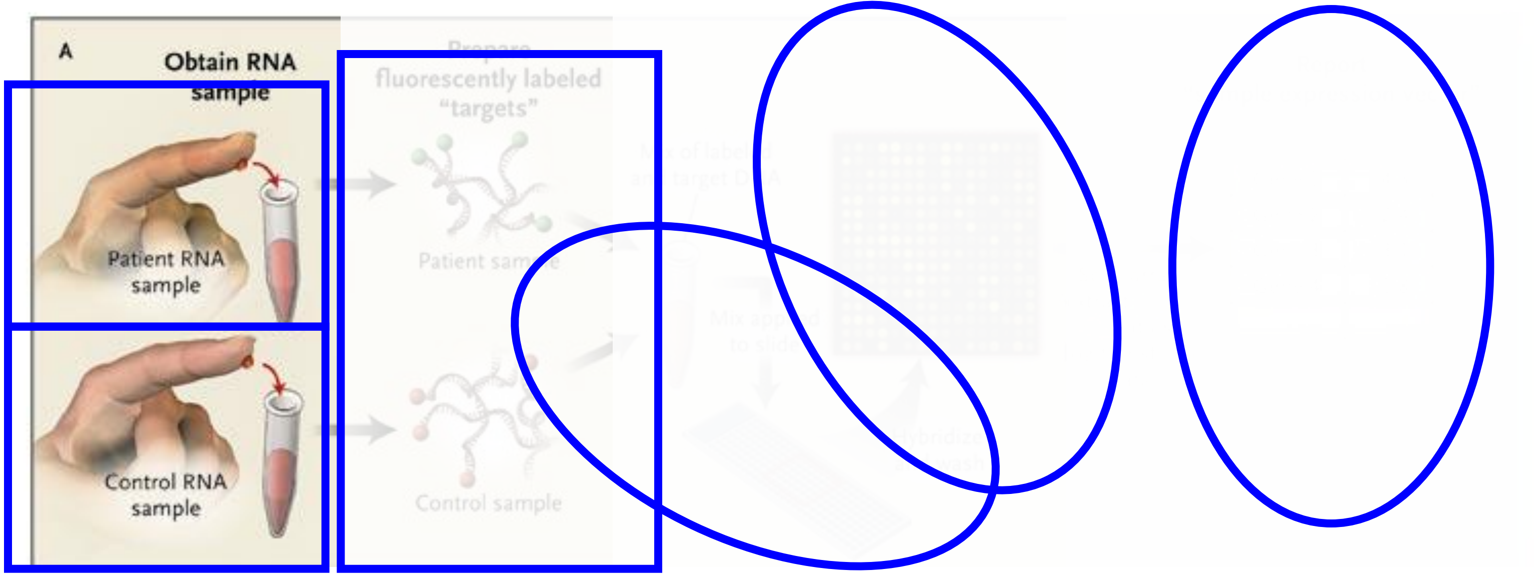
- In 2001 the first draft sequence of the human genome was published
  - DNA sequence of 3 billion A,C,T and G unraveled
- This resulted in a more consistent map of all the genes in the human genome (~25000)
- Concurrently a technology to measure the mRNA activity of all genes was developed: **microarray technology**
  - Chip
  - Probes representing all 25000 genes
  - Measure mRNA activity of all genes in the genome

Microarray technology


1.28 cm
1.28 cm
Actual size of GeneChip® array
ATCATG
Millions of DNA strands built up in each location
500,000 locations on each GeneChip® array
Actual strand = 25 base pairs

Microarray technology

Tumor sample

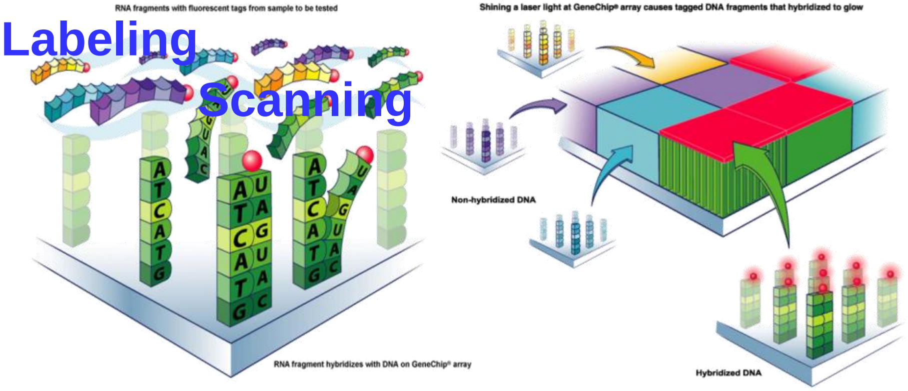


Control sample

Labeling

Scanning

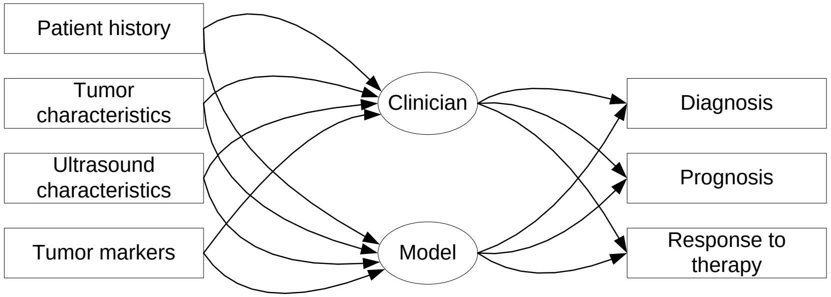
Hybridization


RNA fragments with fluorescent tags from sample to be tested
Shining a laser light at GeneChip® array causes tagged DNA fragments that hybridized to glow
Non-hybridized DNA
RNA fragment hybridizes with DNA on GeneChip® array
Hybridized DNA

DATA

Microarray data

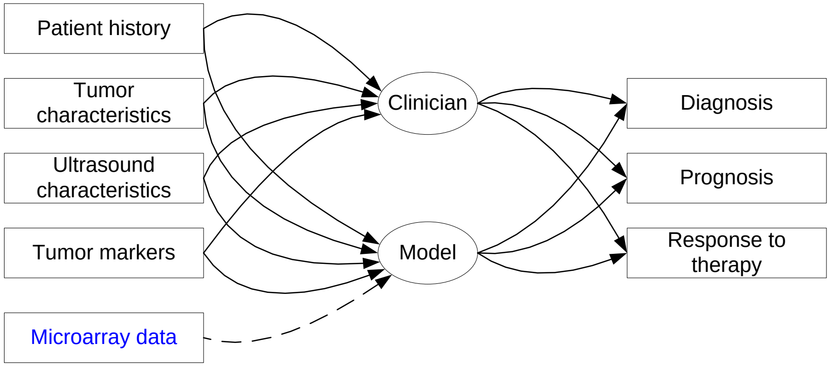
- Microarray technology produces huge quantities of data
  - ~ 25000 values per patient
- This data can also be used for decision support
- Virtually impossible for a clinician to interpret the data directly



Microarray data

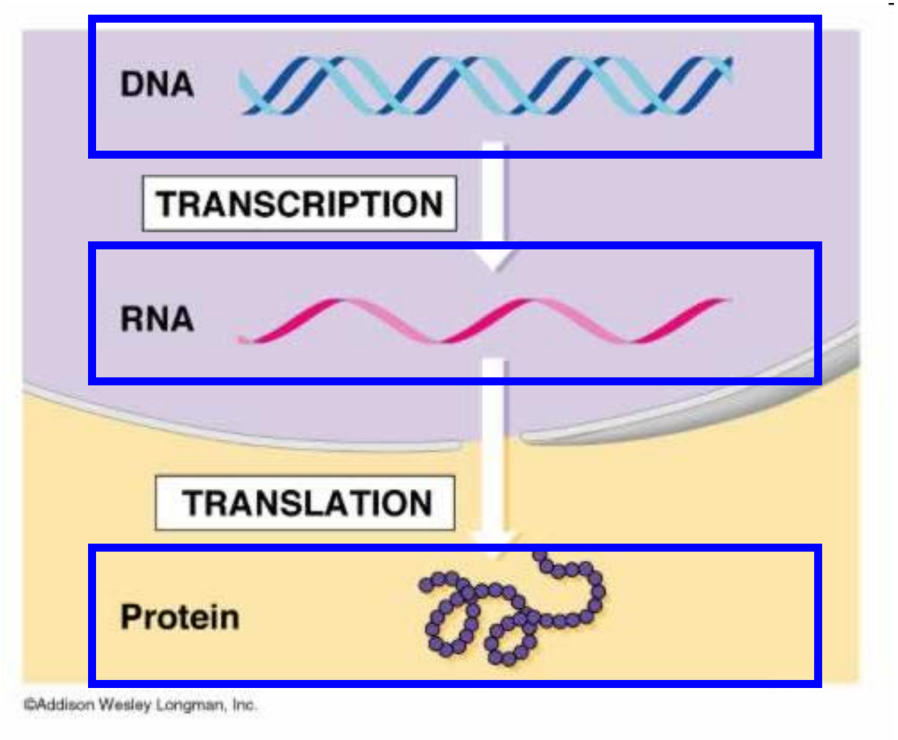
- Microarray technology produces huge quantities of data
  - ~ 25000 values per patient
- This data can also be used for decision support
- Virtually impossible for a clinician to interpret the data directly

- **Biomedical decision support modeling** is the only option



Omics

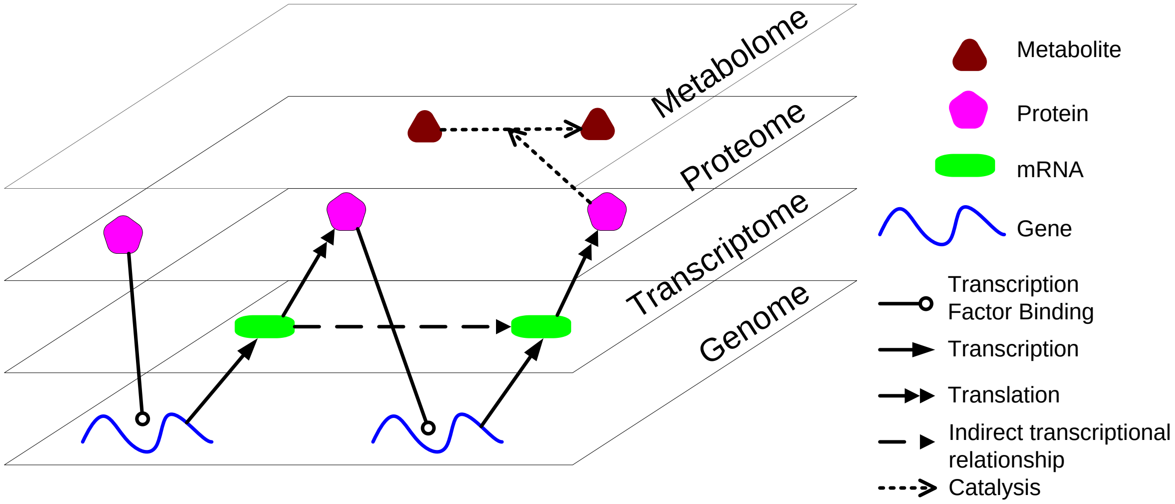
- Microarray technology only measures **mRNA** or the **transcriptome**
- Other levels of molecular biology exist such as
  - Genome
  - Proteome
- These levels are often called **omics**



©Addison Wesley Longman, Inc.

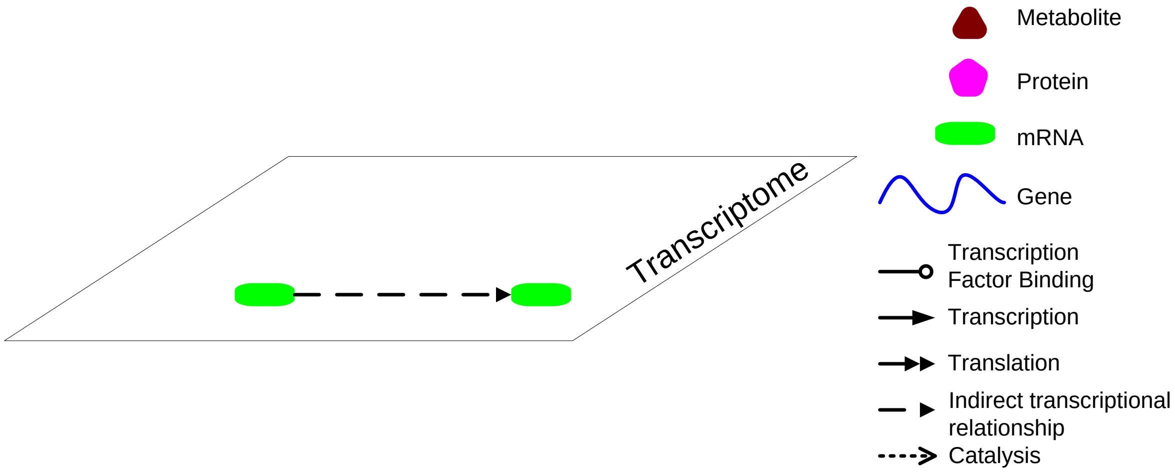
Omics

- Microarray technology is not the only "**omics**" technology
- Other technologies have emerged that profile different levels of molecular biology



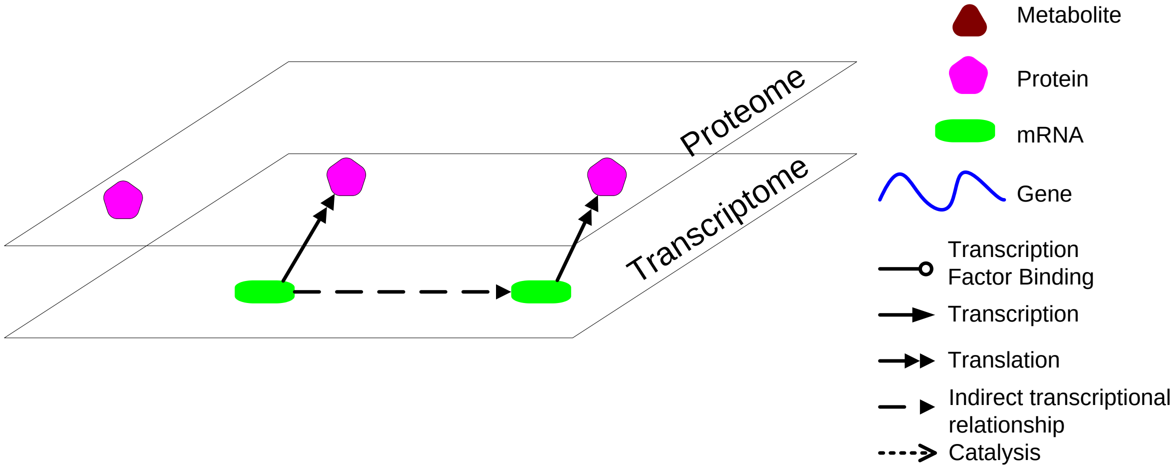
Omics

- Microarray technology only studies the **transcriptome**
- Only **indirect** relationships can be found



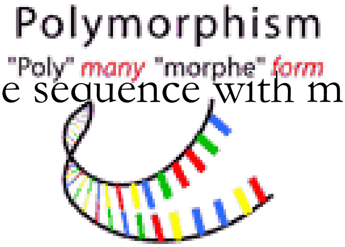
Omics

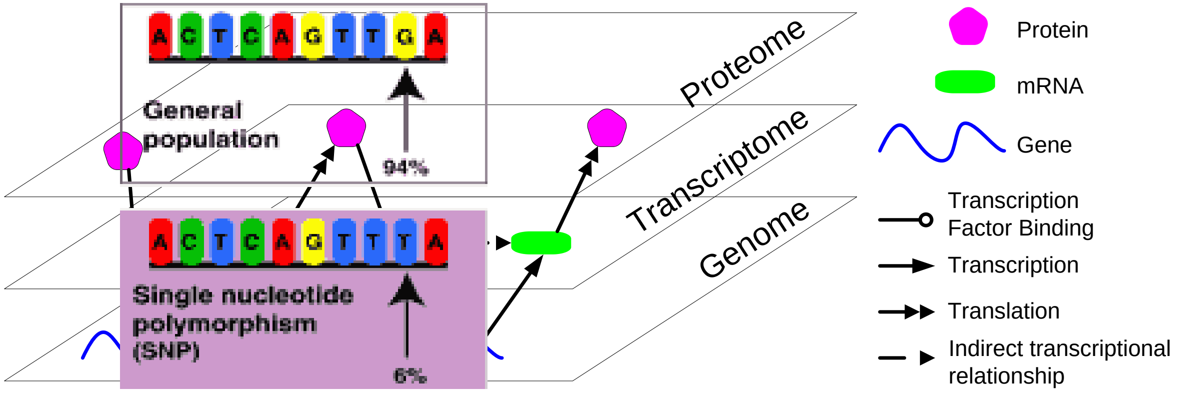
- **Mass spectrometry** based proteomics allows to target the proteome



Omics

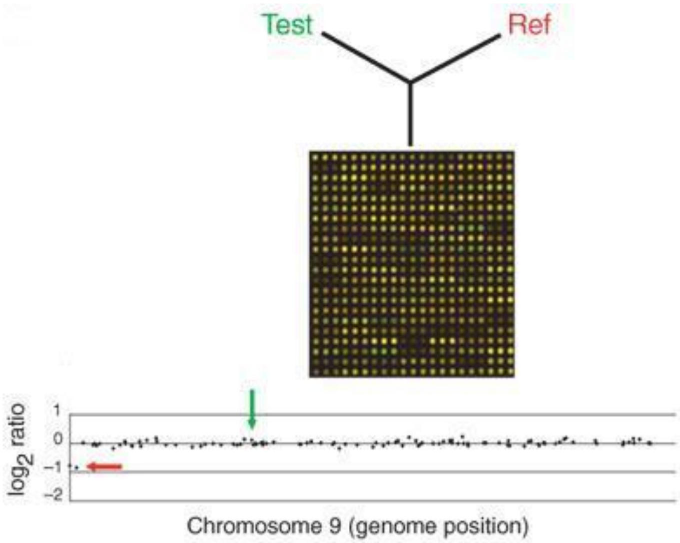
- Also the genome is more variable than previously thought
  - Single base differences between individuals (SNPs)
  - Copy number variations
    - Large pieces of genome sequence with more or less copies





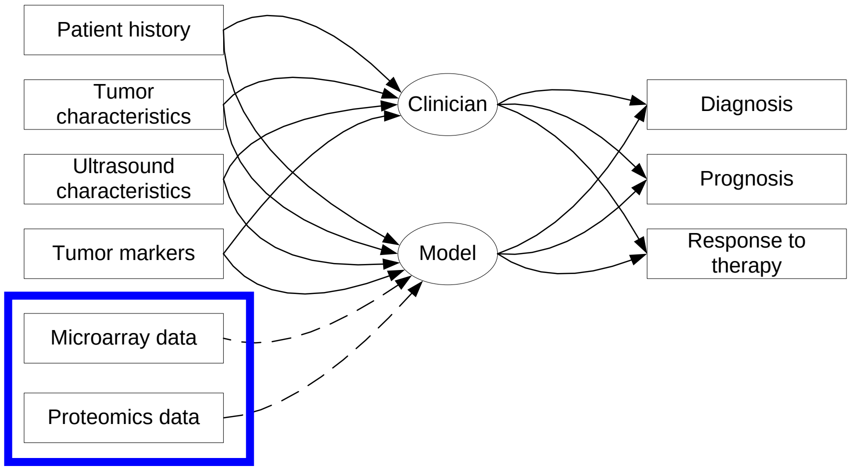
ArrayCGH

- Also the genome is more variable than previously thought
  - Single base differences between individuals (SNPs)
  - Copy number variations
    - Large pieces of genome sequence with more or less copies
    - Array Comparative Genomic Hybridization (arrayCGH)



Motivation

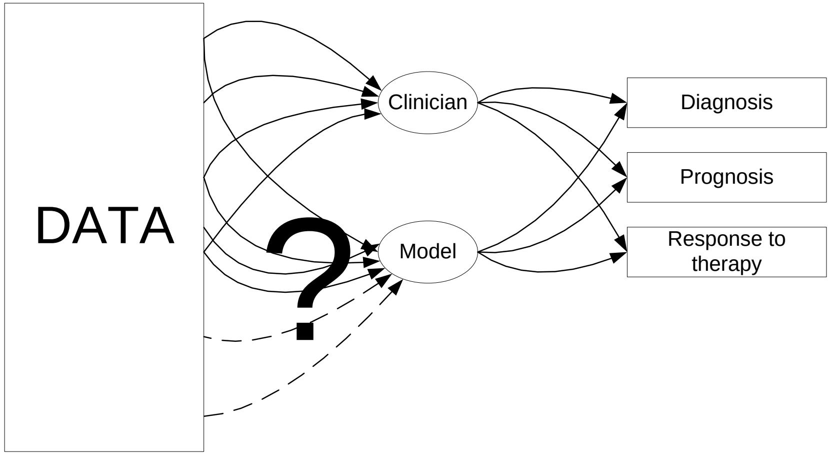
- All these omics technologies have in common that they provide data at a genome scale level:
  - Many variables per patient
  - Not possible to interpret the data manually



Motivation

- All these omics technologies have in common that they provide data at a genome scale level
  - Many variables per patient
  - Not possible to interpret the data manually

- Methods needed to model all these data

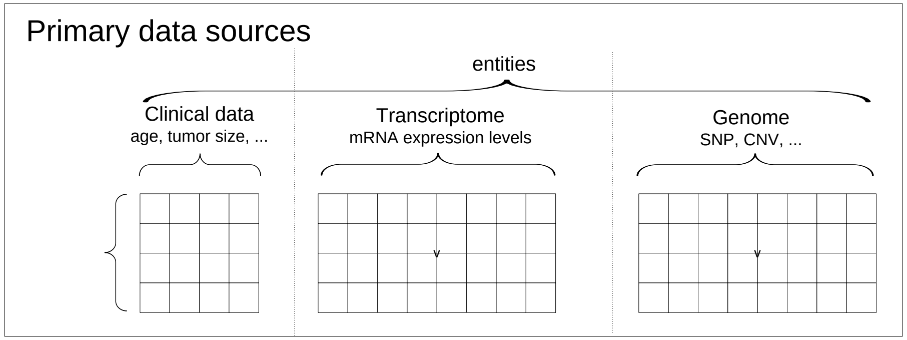


Motivation

- **Our aim** is to investigate if integrating these heterogeneous and high-dimensional data using Bayesian networks improves **predictive performance**
- To support the clinician in making decisions related to the clinical management of diseases:
  - Diagnosis
  - Prognosis
  - Therapy response
- We have defined two types of data

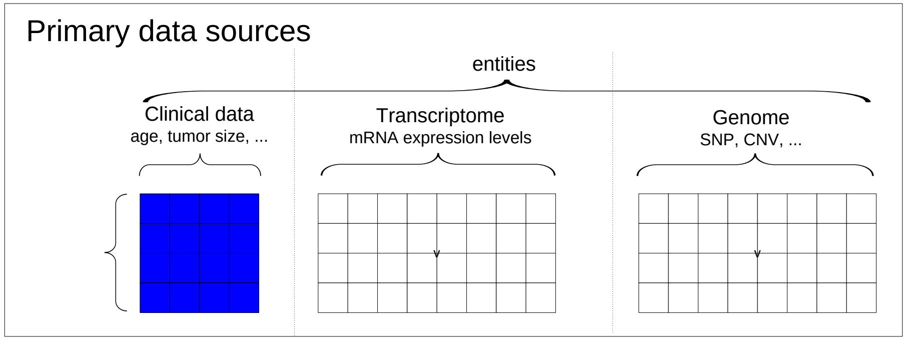
Primary vs. secondary data

- Primary data is patient specific



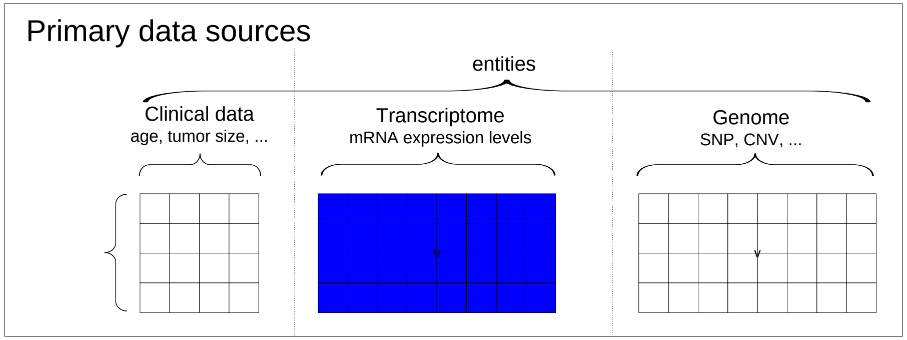
Primary vs. secondary data

- Primary data is patient specific



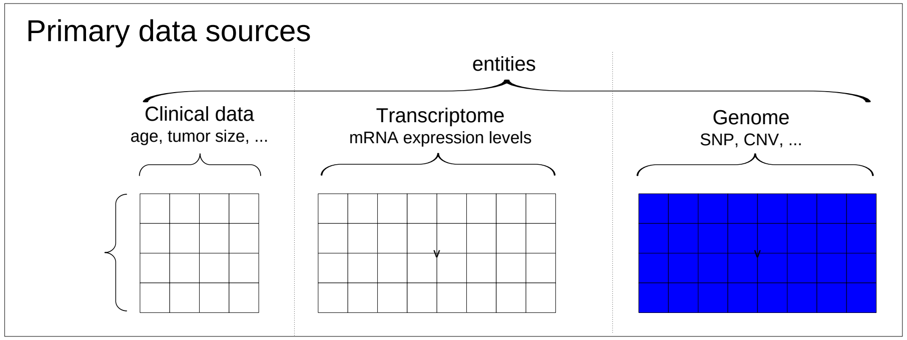
Primary vs. secondary data

- Primary data is patient specific



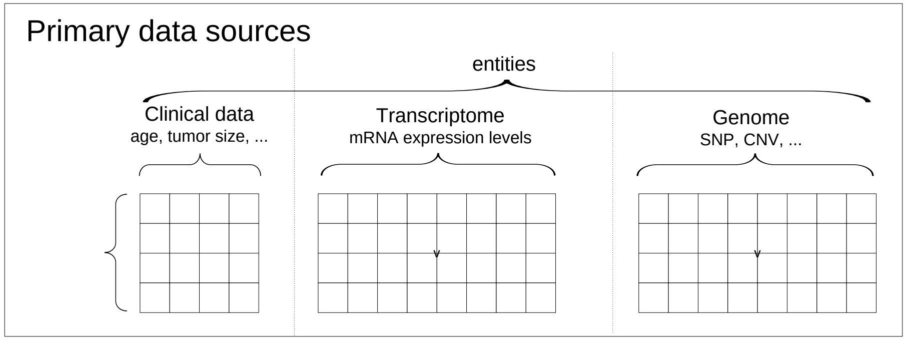
Primary vs. secondary data

- Primary data is patient specific



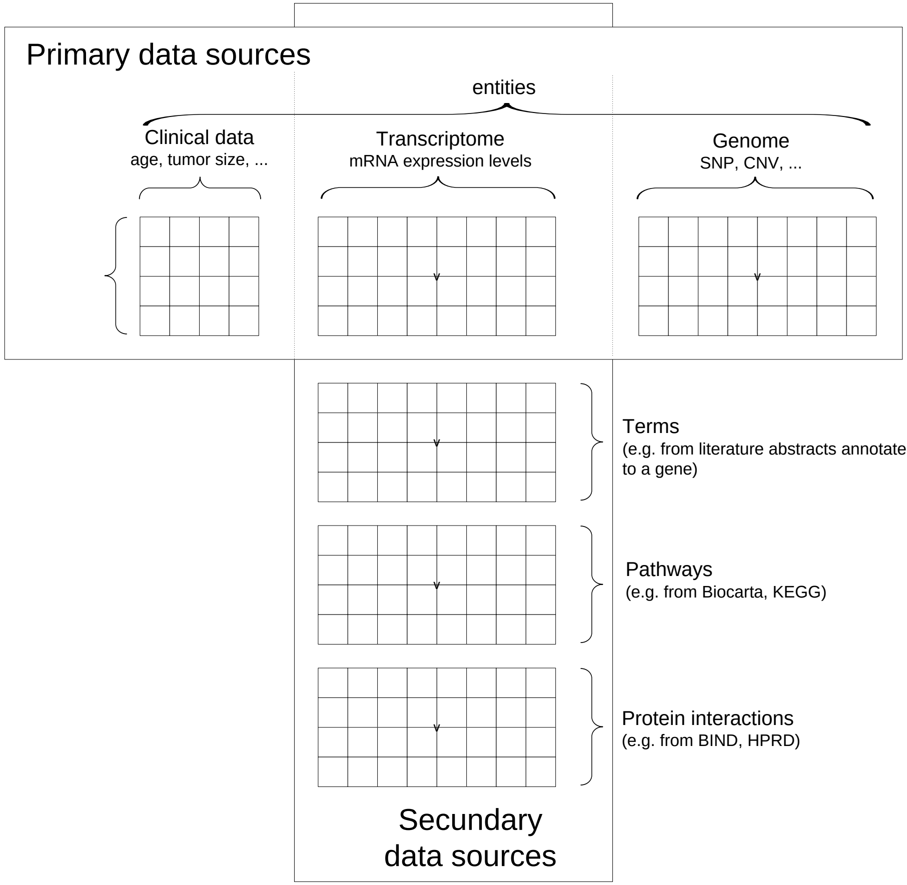
Primary vs. secondary data

- Primary data is patient specific
- Secondary data is entity specific
  - Gene in genome
  - mRNA in transcriptome
  - Protein in proteome



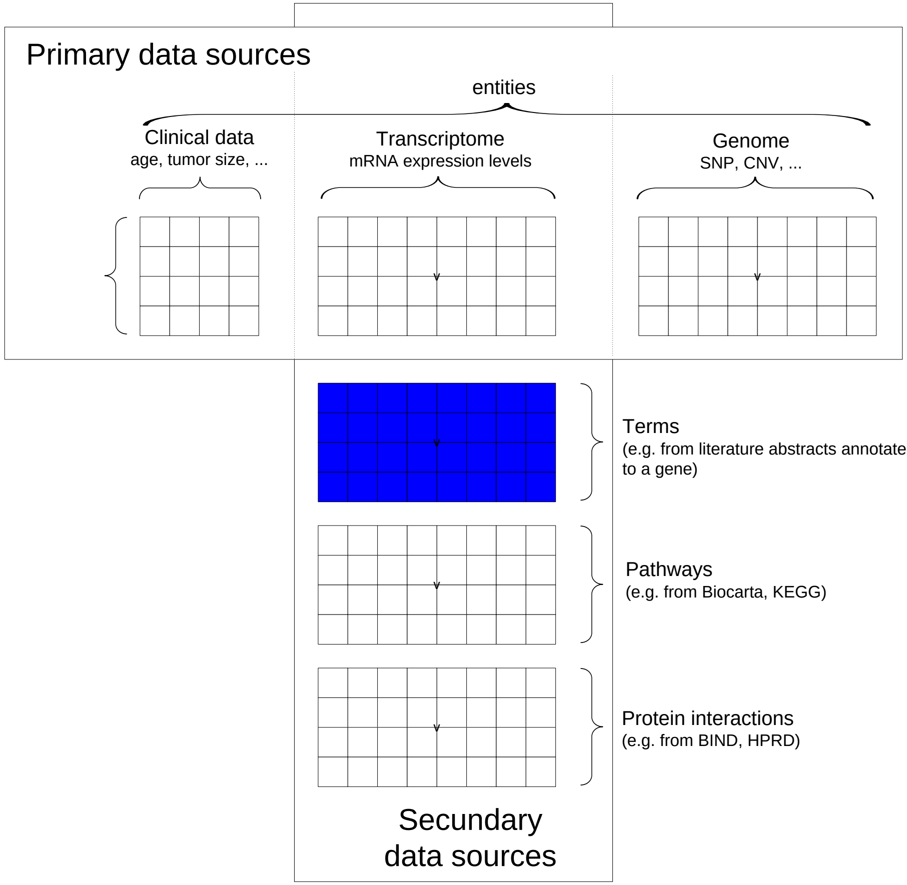
Primary vs. secondary data

- Primary data is patient specific
- Secondary data is entity specific
  - Gene in genome
  - mRNA in transcriptome
  - Protein in proteome



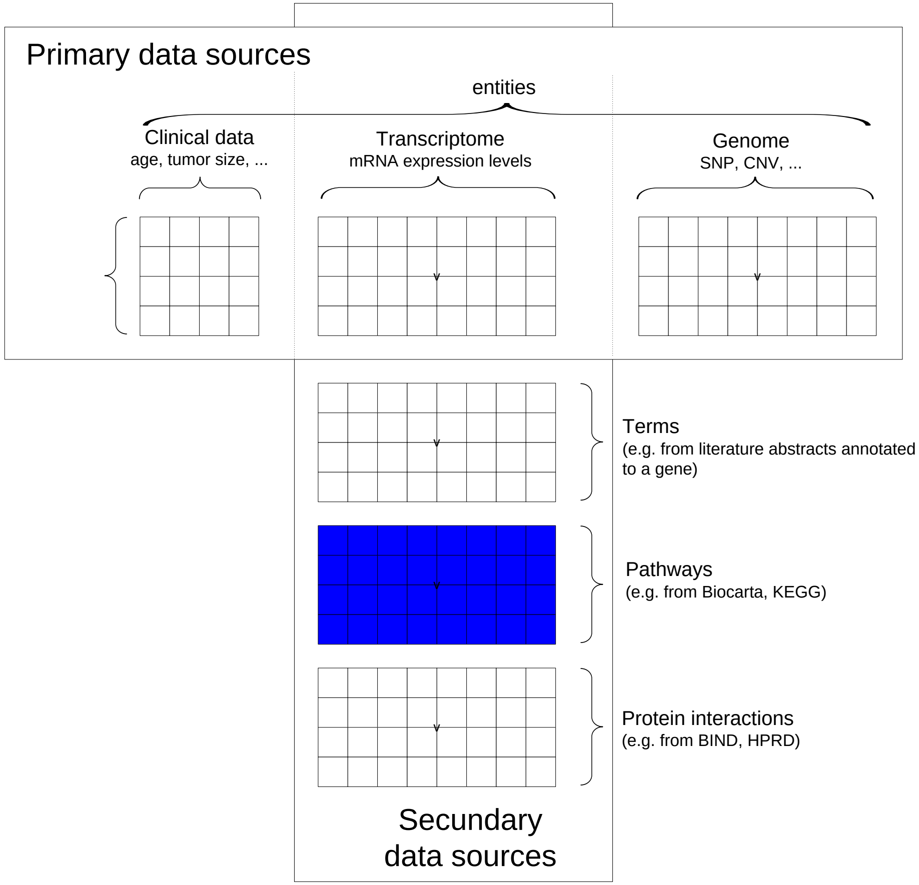
Primary vs. secondary data

- Primary data is patient specific
- Secondary data is entity specific
  - Gene in genome
  - mRNA in transcriptome
  - Protein in proteome



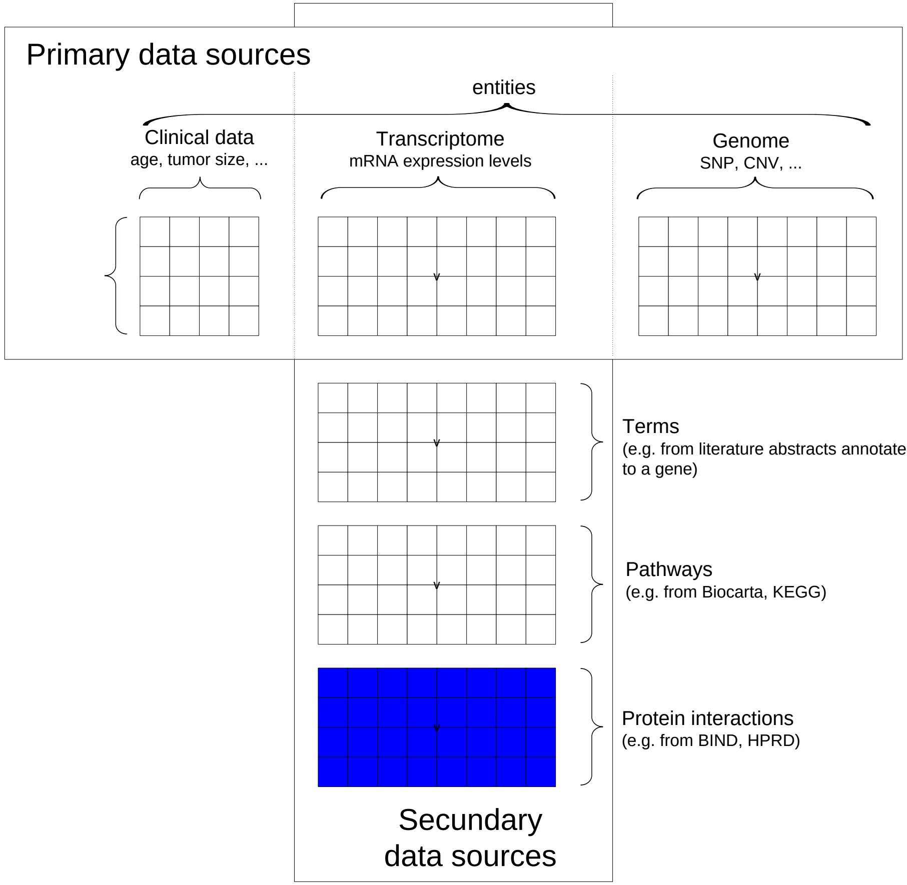
Primary vs. secondary data

- Primary data is patient specific
- Secondary data is entity specific
  - Gene in genome
  - mRNA in transcriptome
  - Protein in proteome



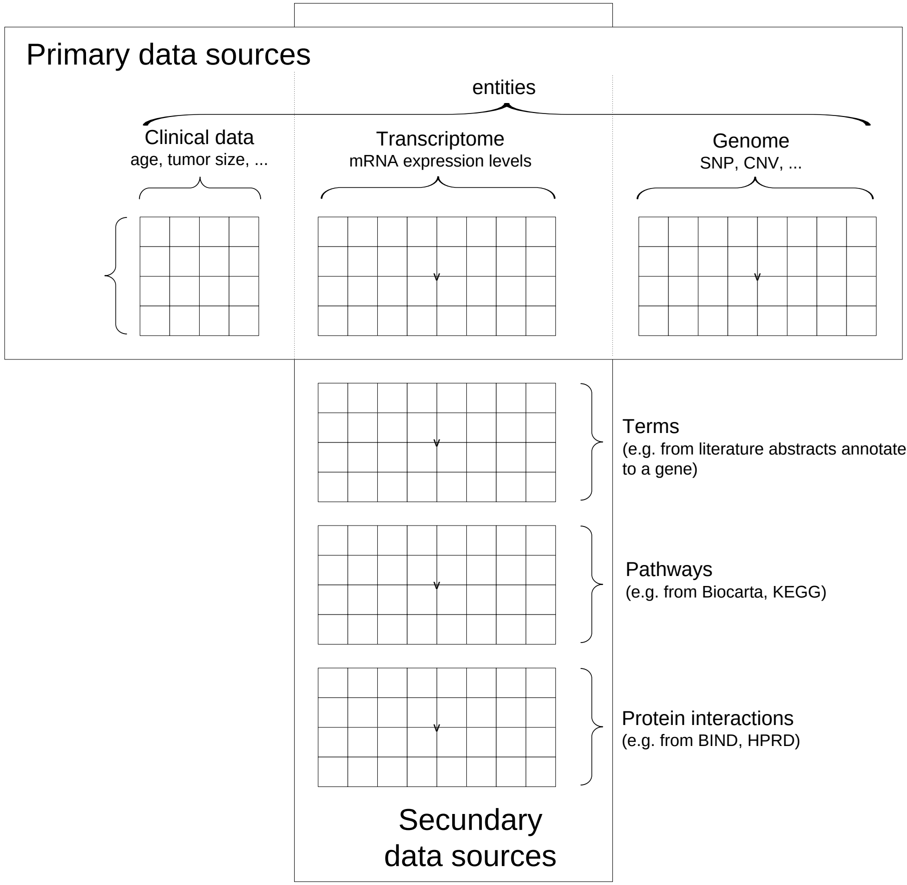
Primary vs. secondary data

- Primary data is patient specific
- Secondary data is entity specific
  - Gene in genome
  - mRNA in transcriptome
  - Protein in proteome



Primary vs. secondary data

- Primary data is patient specific
- Secondary data is entity specific
  - Gene in genome
  - mRNA in transcriptome
  - Protein in proteome
- Secondary data integration is motivated by its availability in publicly available databases
  - IntAct
  - Reactome
  - KEGG
  - TRANSFAC



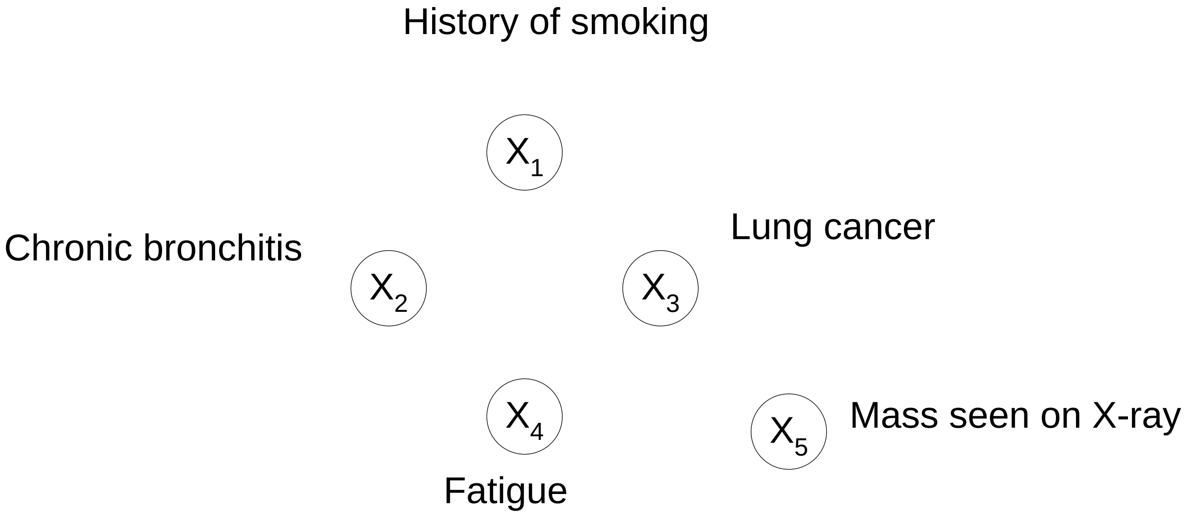
Aims

1. Modeling separate primary data sources
   - Clinical data – modeling ovarian masses with Bayesian networks
   - Genomic data – modeling CNAs using a special class of Bayesian networks on BRCA1-mutated and sporadic ovarian cancers
2. Integration of primary data
   - Breast cancer
   - Rectal cancer
3. Integration of secondary data

Bayesian networks

Toy example

- What is a Bayesian network?
  - 5 variables related to lung cancer: X_1, X_2, X_3, X_4 and X_5
  - All variables can have two values: Yes/No



History of smoking

X_1

Lung cancer

Chronic bronchitis

X_2

X_3

X_4

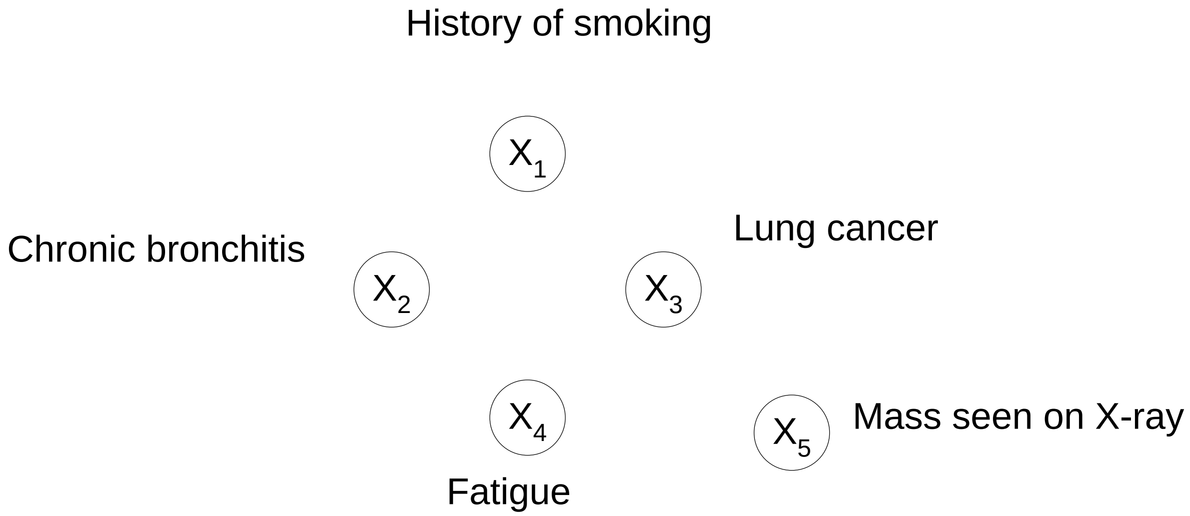
X_5

Mass seen on X-ray

Fatigue

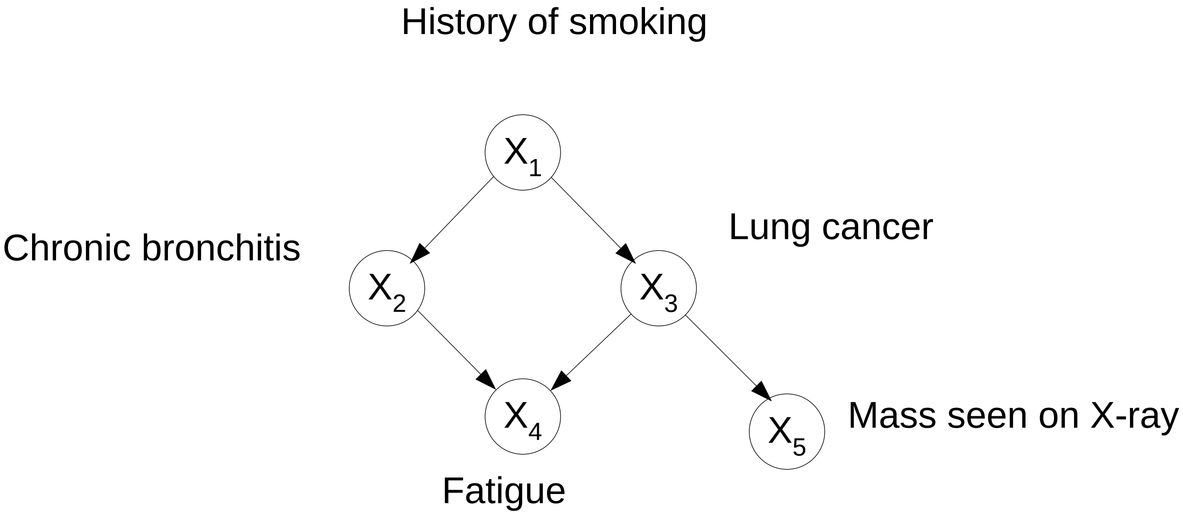
Definition

- A Bayesian network consists of two parts



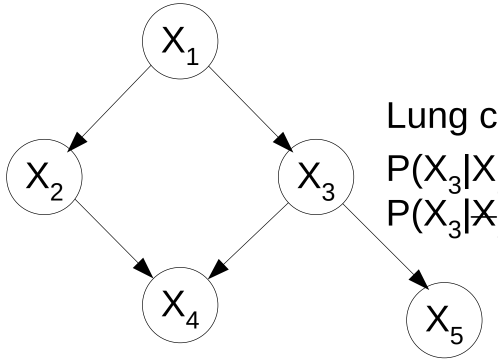
Definition

- A Bayesian network consists of two parts
  - Structure: directed acyclic graph



Definition

- A Bayesian network consists of two parts
  - Structure: directed acyclic graph
  - Parameters: conditional probability tables (CPT)



History of smoking

$P(X_1)= 20\%$

Chronic bronchitis

$P(X_2|X_1)= 25\%$

$P(X_2|\cancel{X}_1)= 5\%$

Lung cancer

$P(X_3|X_1)= 0.3\ \%$

$P(X_3|\cancel{X}_1)= 0.005\ \%$

Fatigue

$P(X_4|X_2,X_3)= 75\%$

$P(X_4|X_2,\cancel{X}_3)= 10\%$

$P(X_4|\cancel{X}_2,X_3)= 50\%$

$P(X_4|\cancel{X}_2,\cancel{X}_3)= 5\%$

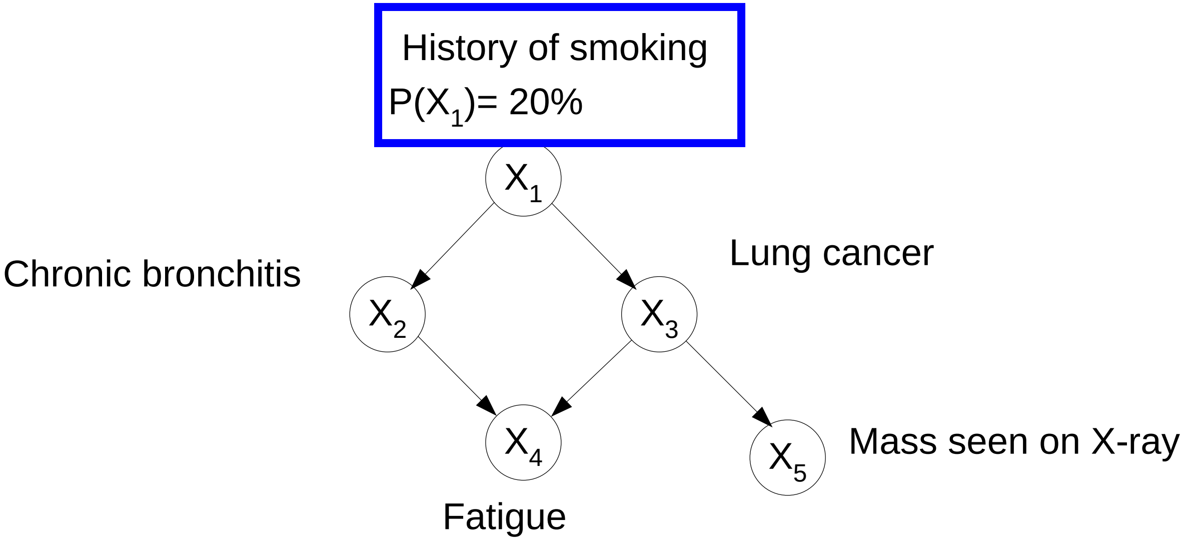
Mass seen on X-ray

$P(X_5|X_3)= 60\ \%$

$P(X_5|\cancel{X}_3)= 2\ \%$

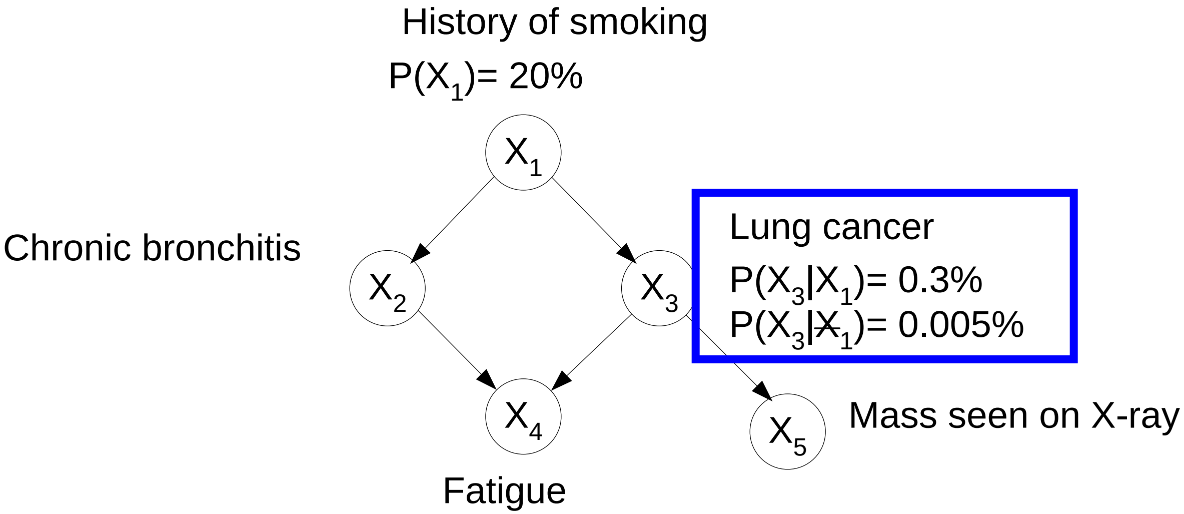
Definition

- A Bayesian network consists of two parts
  - Structure: directed acyclic graph
  - Parameters: conditional probability tables (CPT)



Definition

- Two parts
  - Structure: directed acyclic graph
  - Parameters: conditional probability tables (CPT)



History of smoking

$P(X_1)= 20\%$

Chronic bronchitis

Lung cancer

$P(X_3|X_1)= 0.3\%$

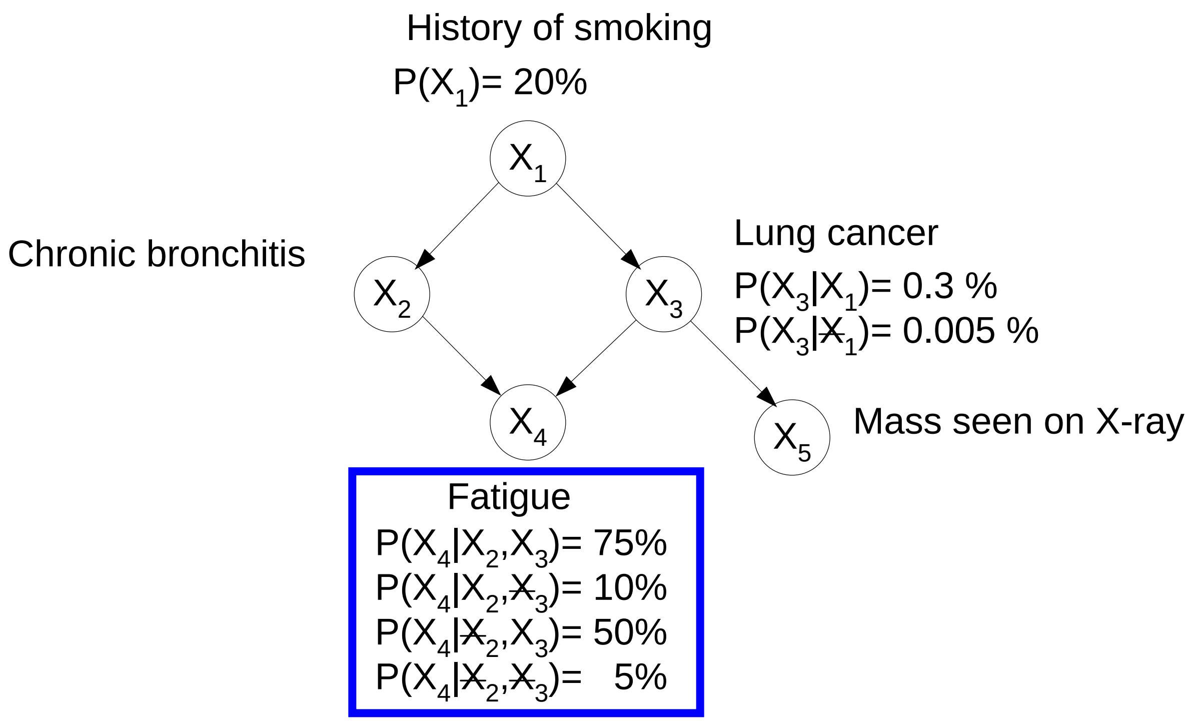
$P(X_3|\cancel{X}_1)= 0.005\%$

Mass seen on X-ray

Fatigue

Definition

- Two parts
  - Structure: directed acyclic graph
  - Parameters: conditional probability tables (CPT)



History of smoking

$P(X_1)= 20\%$

Chronic bronchitis

Lung cancer

$P(X_3|X_1)= 0.3\ \%$

$P(X_3|\cancel{X}_1)= 0.005\ \%$

Mass seen on X-ray

Fatigue

$P(X_4|X_2,X_3)= 75\%$

$P(X_4|X_2,\cancel{X}_3)= 10\%$

$P(X_4|\cancel{X}_2,X_3)= 50\%$

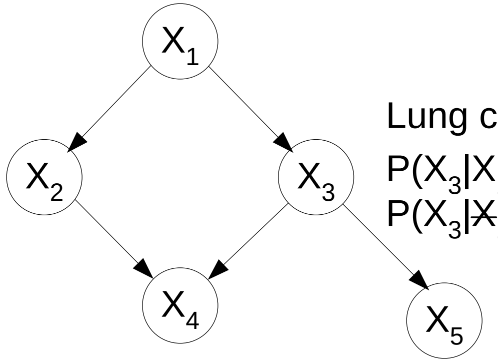
$P(X_4|\cancel{X}_2,\cancel{X}_3)= 5\%$

Definition

- Two parts
  - Structure: directed acyclic graph
  - Parameters: conditional probability tables (CPT)

History of smoking

$P(X_1)= 20\%$



Chronic bronchitis

$P(X_2|X_1)= 25\%$
$P(X_2|\cancel{X_1})= 5\%$

Lung cancer

$P(X_3|X_1)= 0.3\ \%$
$P(X_3|\cancel{X_1})= 0.005\ \%$

Mass seen on X-ray

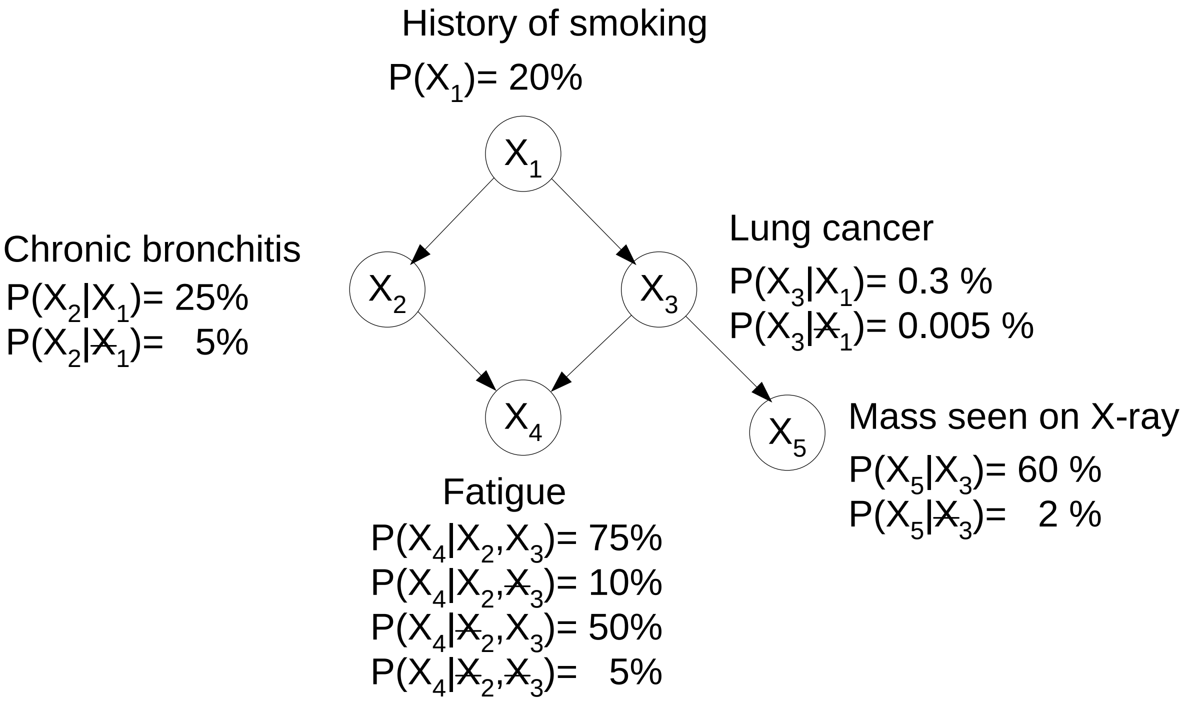
$P(X_5|X_3)= 60\ \%$
$P(X_5|\cancel{X_3})= 2\ \%$

Fatigue

$P(X_4|X_2,X_3)= 75\%$
$P(X_4|X_2,\cancel{X_3})= 10\%$
$P(X_4|\cancel{X_2},X_3)= 50\%$
$P(X_4|\cancel{X_2},\cancel{X_3})= 5\%$

Bayesian networks

- In most cases both the structure and the parameters are not known



History of smoking

$P(X_1)$= 20%

Chronic bronchitis

$P(X_2|X_1)$= 25%
$P(X_2|\cancel{X}_1)$= 5%

Lung cancer

$P(X_3|X_1)$= 0.3 %
$P(X_3|\cancel{X}_1)$= 0.005 %

Fatigue

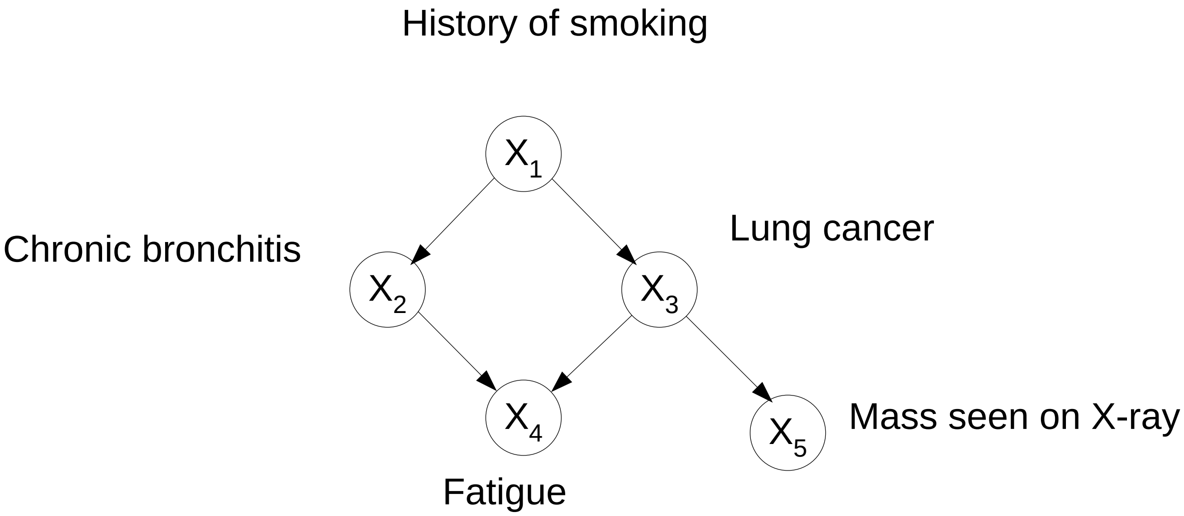
$P(X_4|X_2,X_3)$= 75%
$P(X_4|X_2,\cancel{X}_3)$= 10%
$P(X_4|\cancel{X}_2,X_3)$= 50%
$P(X_4|\cancel{X}_2,\cancel{X}_3)$= 5%

Mass seen on X-ray

$P(X_5|X_3)$= 60 %
$P(X_5|\cancel{X}_3)$= 2 %

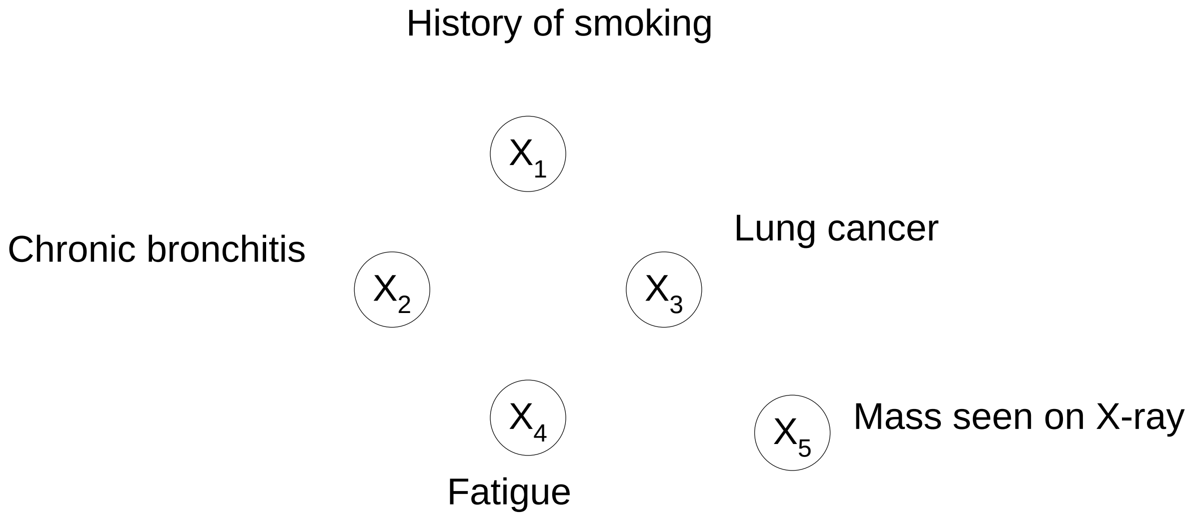
Bayesian networks

- In most cases both the structure and the parameters are not known



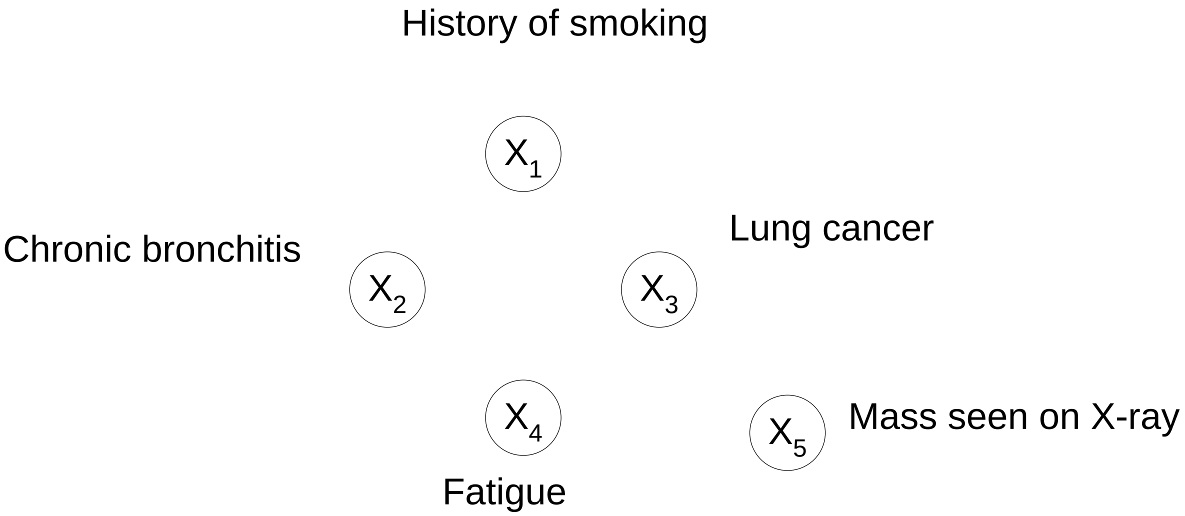
Bayesian networks

- In most cases both the structure and the parameters are not known
- And have to be learned from data



Bayesian networks

- In most cases both the structure and the parameters are not known
- And have to be learned from data
- Bayesian network learning
  - Structure learning
  - Parameter learning



Structure learning

- Greedy search with Bayesian Dirichlet scoring metric
- Reflects how well a structure has produced the data

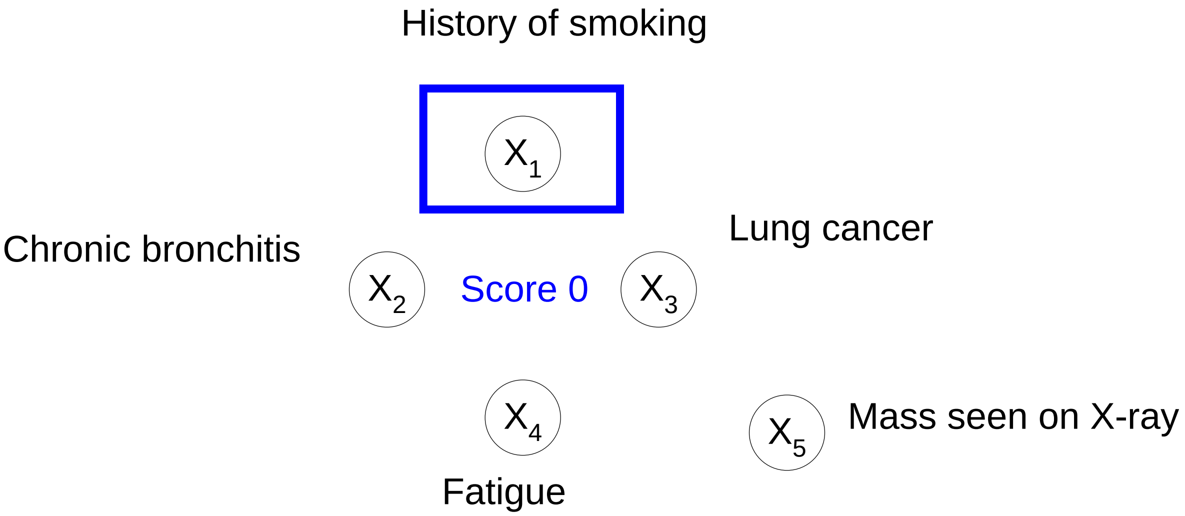
$$p\left(S\middle|D\right)\propto\prod_{i=1}^{n}\prod_{j=1}^{q_i}\left[\frac{\Gamma\left(N_{ij}^{'}\right)}{\Gamma\left(N_{ij}^{'}+N_{ij}\right)}\right]\prod_{k=1}^{r_i}\frac{\Gamma\left(N_{ijk}^{'}+N_{ijk}\right)}{\Gamma\left(N_{ijk}^{'}\right)}\quad P\left(S\right)$$

⇩

Scoring structures based on data

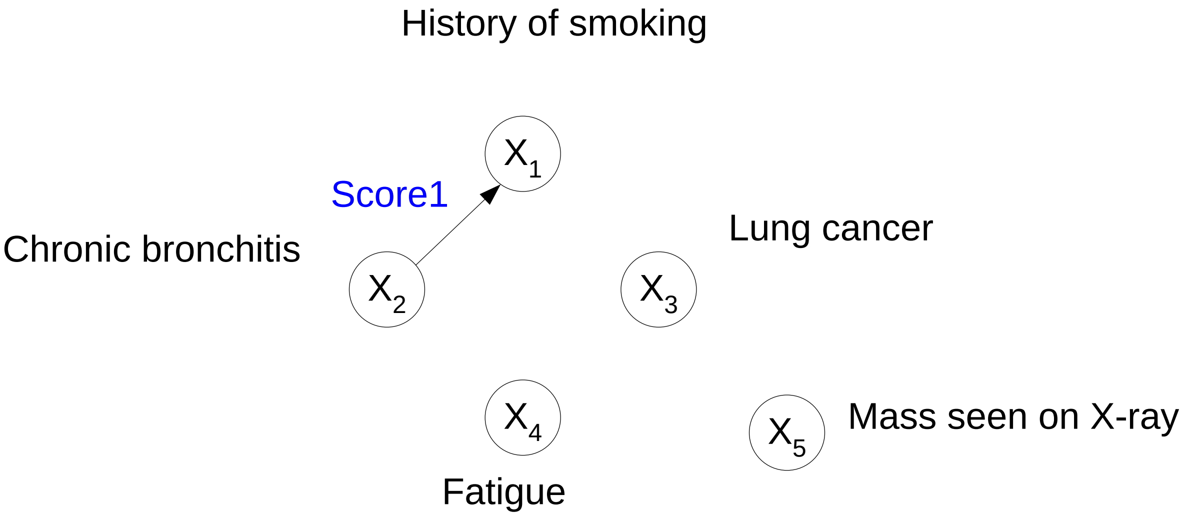
Structure learning

- Greedy search
  - Model 0



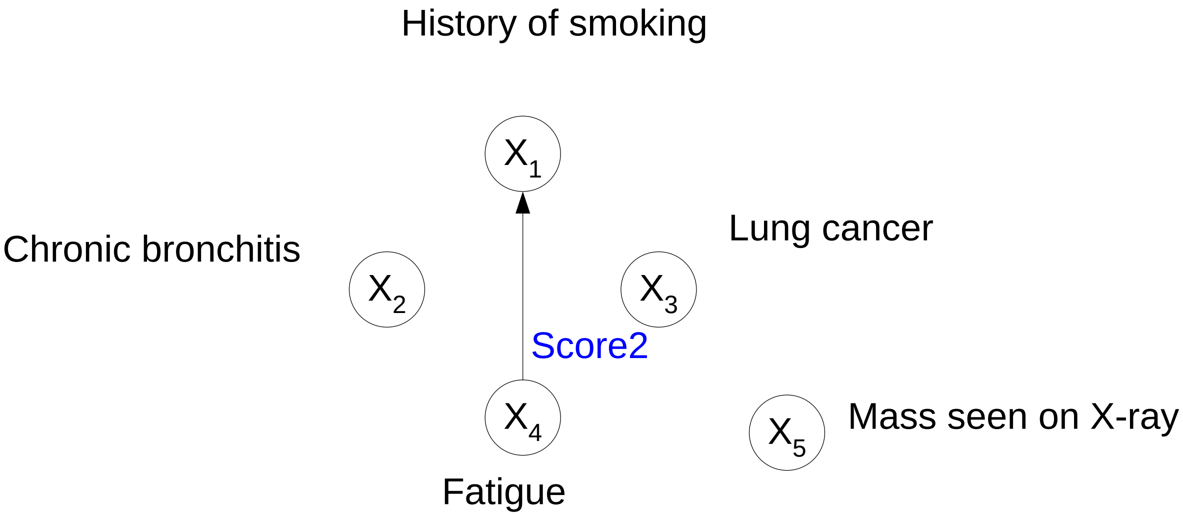
Structure learning

- Greedy search
  - Model 0
  - Model 1



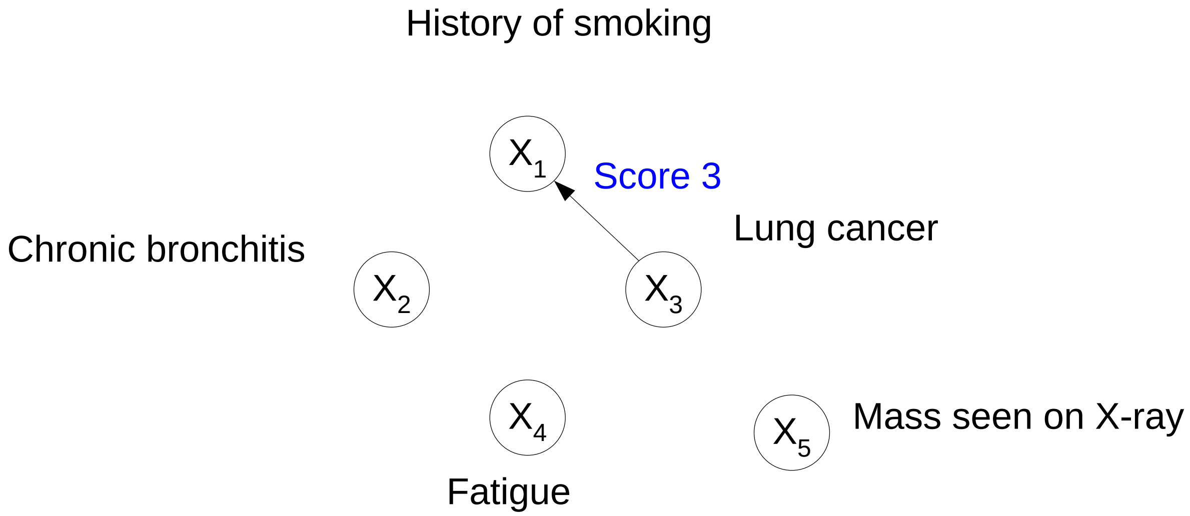
Structure learning

- Greedy search
  - Model 0
  - Model 1
  - Model 2



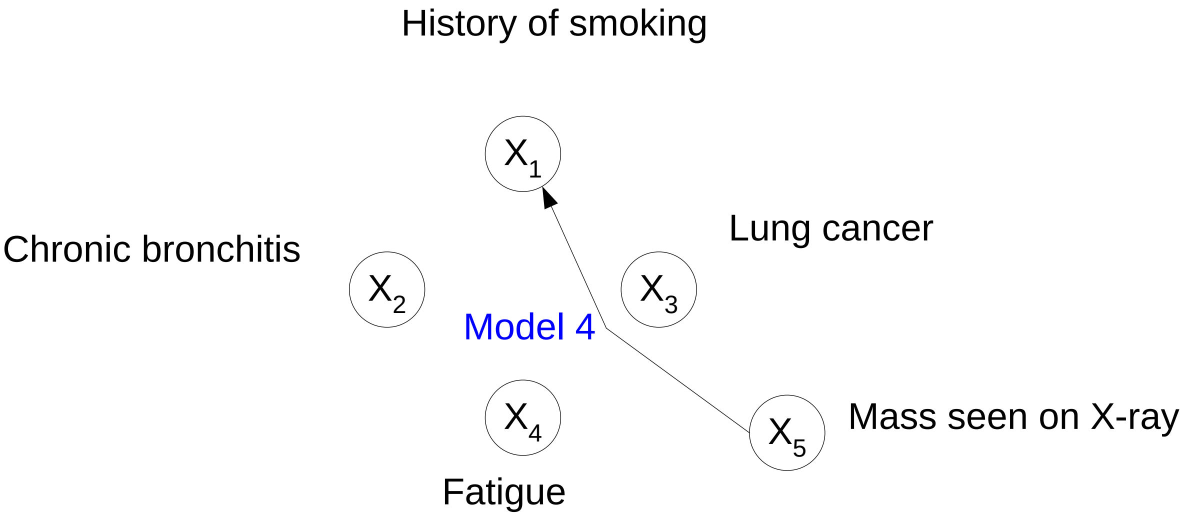
Structure learning

- Greedy search
  - Model 0
  - Model 1
  - Model 2
  - Model 3



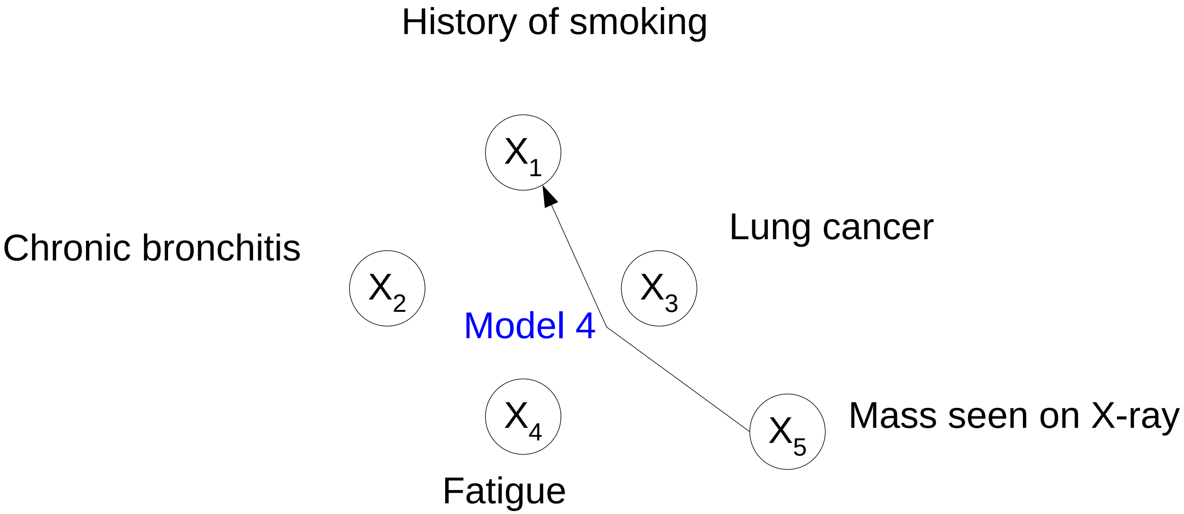
Structure learning

- Greedy search
  - Model 0
  - Model 1
  - Model 2
  - Model 3
  - Model 4



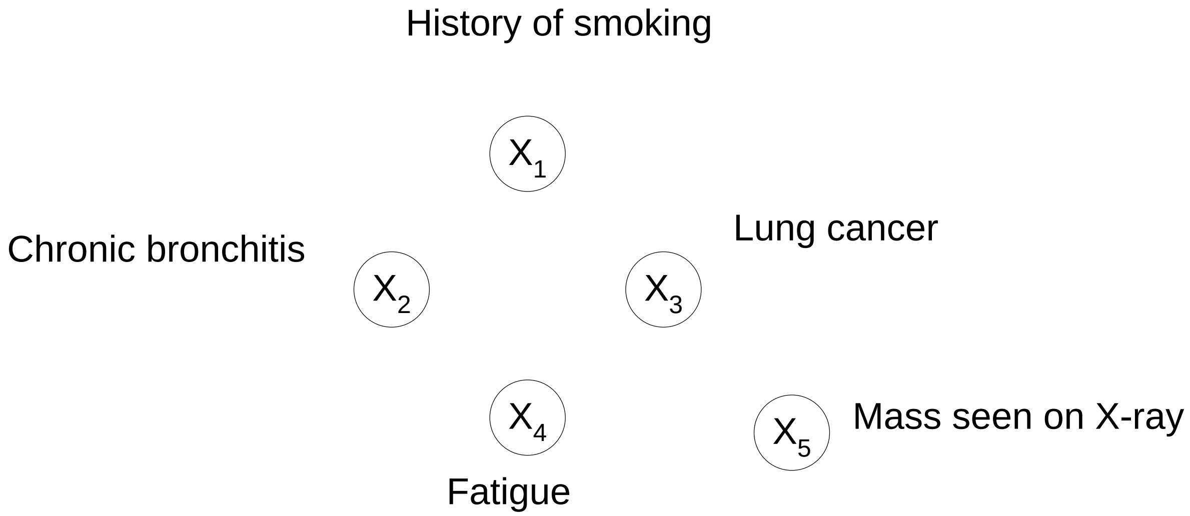
Structure learning

- Greedy search
  - **Model 0: best model**
  - Model 1
  - Model 2
  - Model 3
  - Model 4



Structure learning

- Best of these models is chosen
  - Model 0 with no edges
  - No edges added ⇒ move to next variable



History of smoking

X_1

Chronic bronchitis

X_2

Lung cancer

X_3

X_4

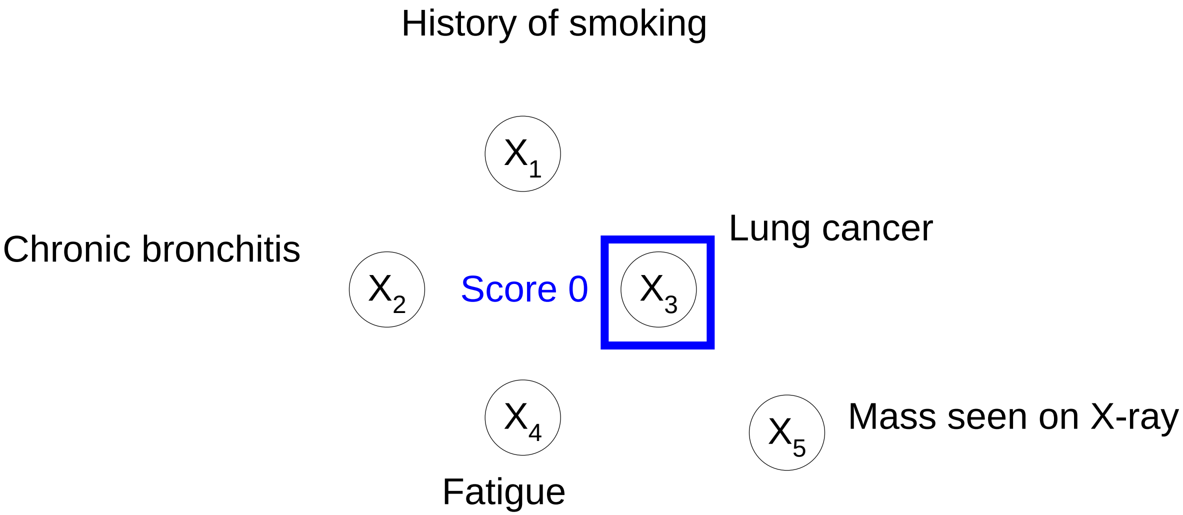
Fatigue

X_5

Mass seen on X-ray

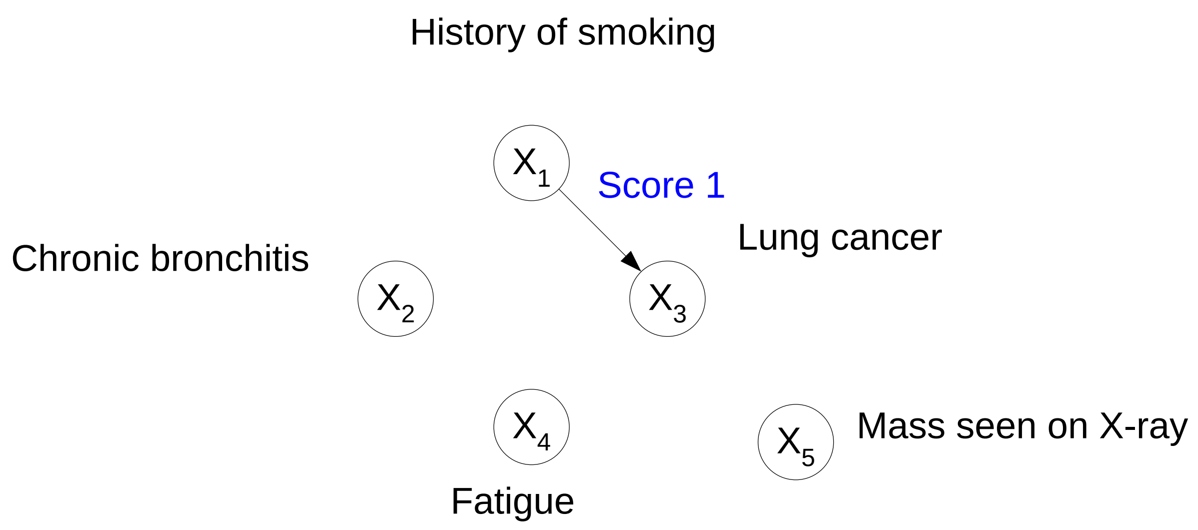
Structure learning

- Suppose X_3 is next variable
- Start greedy search for X_3
  - Model 0



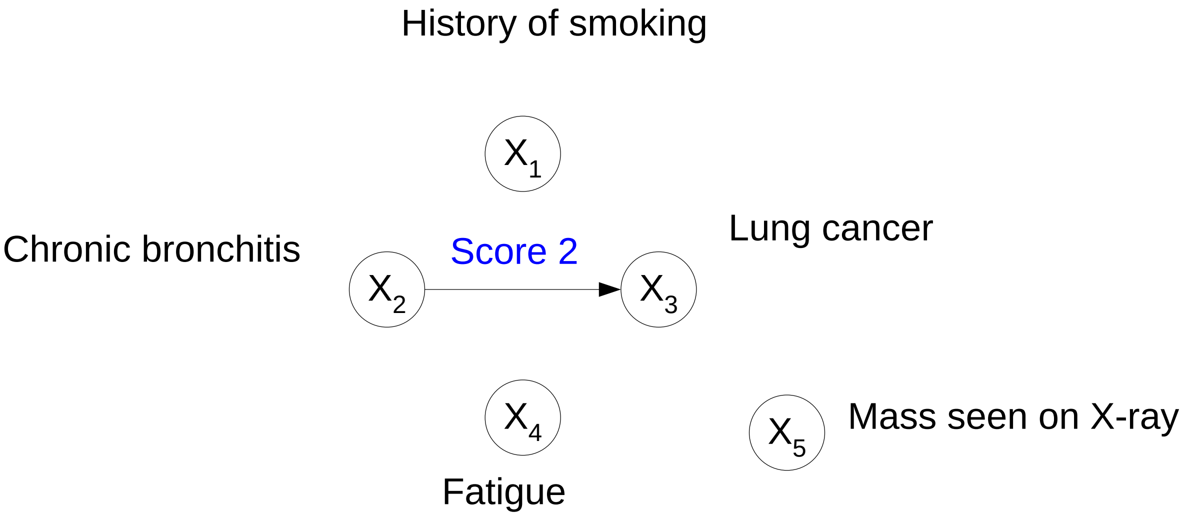
Structure learning

- Suppose X_3 is next variable
- Start greedy search for X_3
  - Model 0
  - Model 1



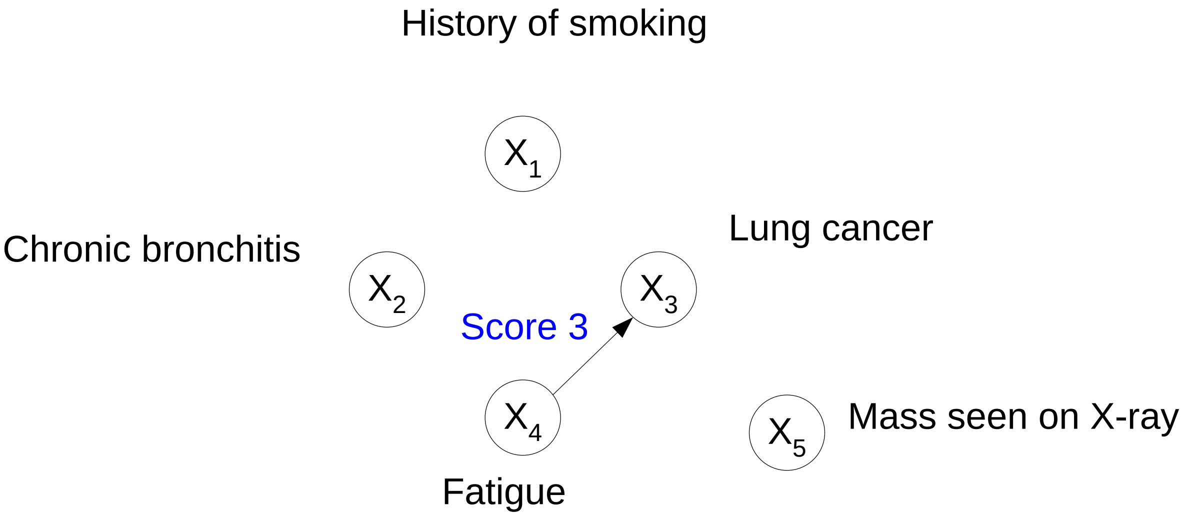
Structure learning

- Suppose X_3 is next variable
- Start greedy search for X_3
  - Model 0
  - Model 1
  - Model 2



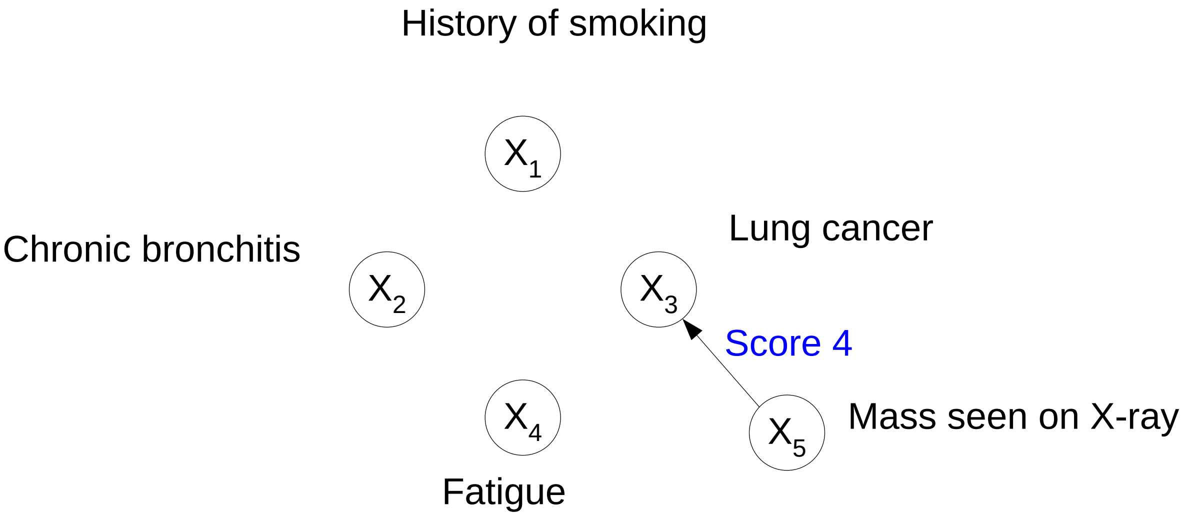
Structure learning

- Suppose X_3 is next variable
- Start greedy search for X_3
  - Model 0
  - Model 1
  - Model 2
  - Model 3



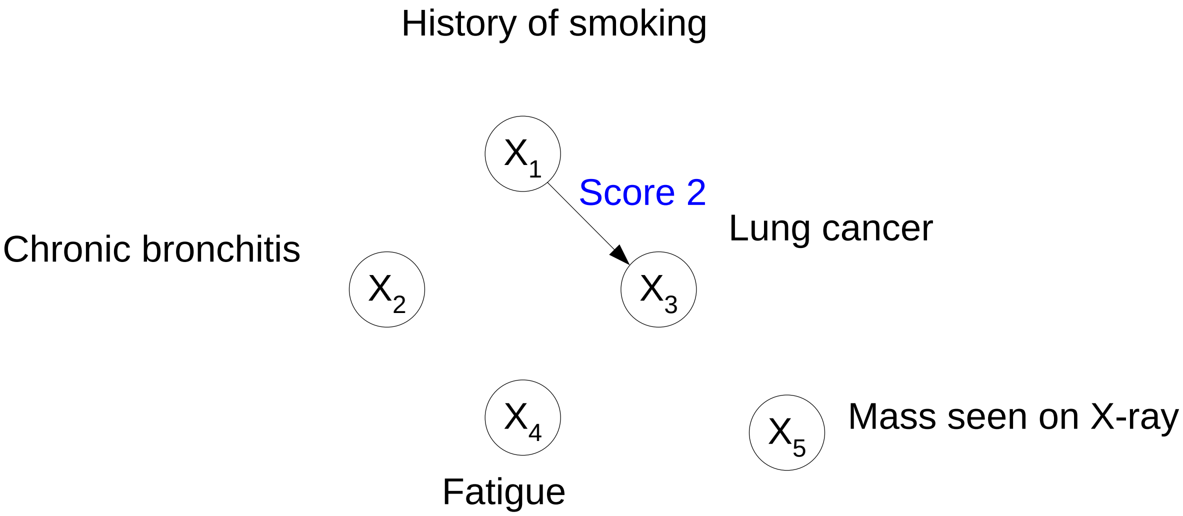
Structure learning

- Suppose X_3 is next variable
- Start greedy search for X_3
  - Model 0
  - Model 1
  - Model 2
  - Model 3
  - Model 4



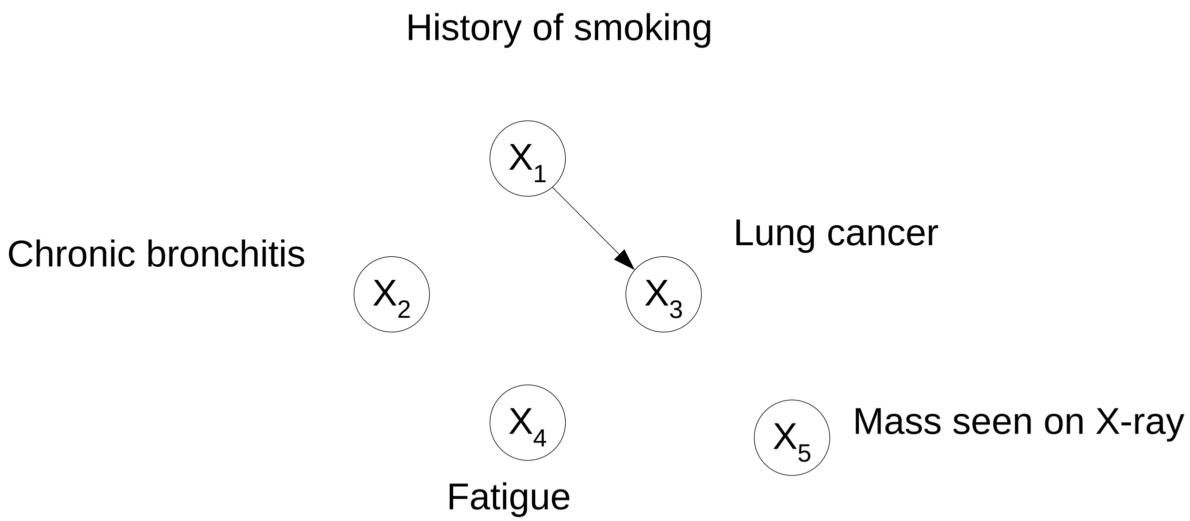
Structure learning

- Suppose X_3 is next variable
- Start greedy search for X_3
  - Model 0
  - Model 1
  - **Model 2**
  - Model 3
  - Model 4



Structure learning

- Model 2
  - Add second edge if score is improved upon



History of smoking

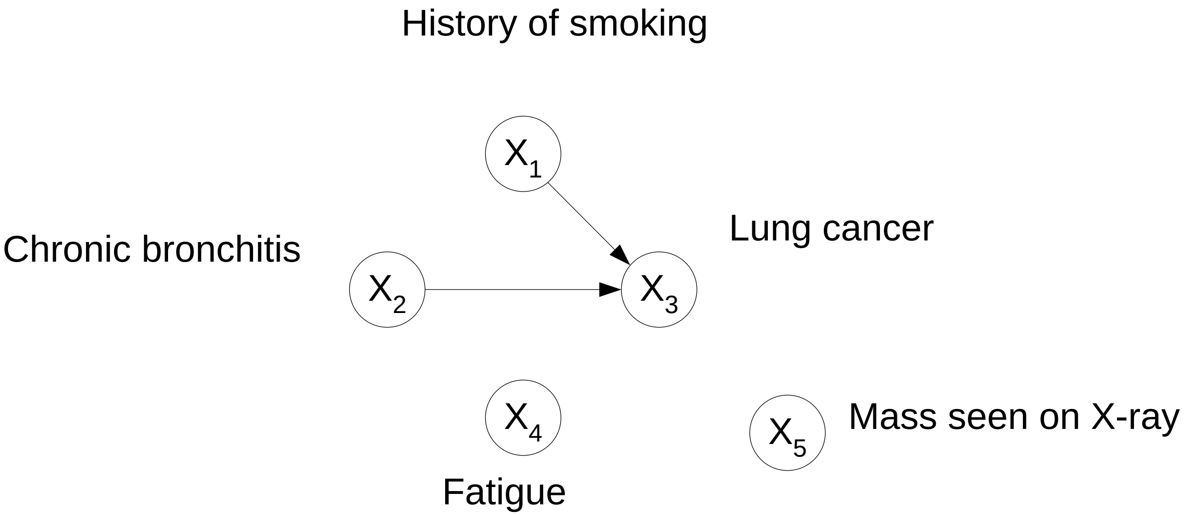
Chronic bronchitis

Lung cancer

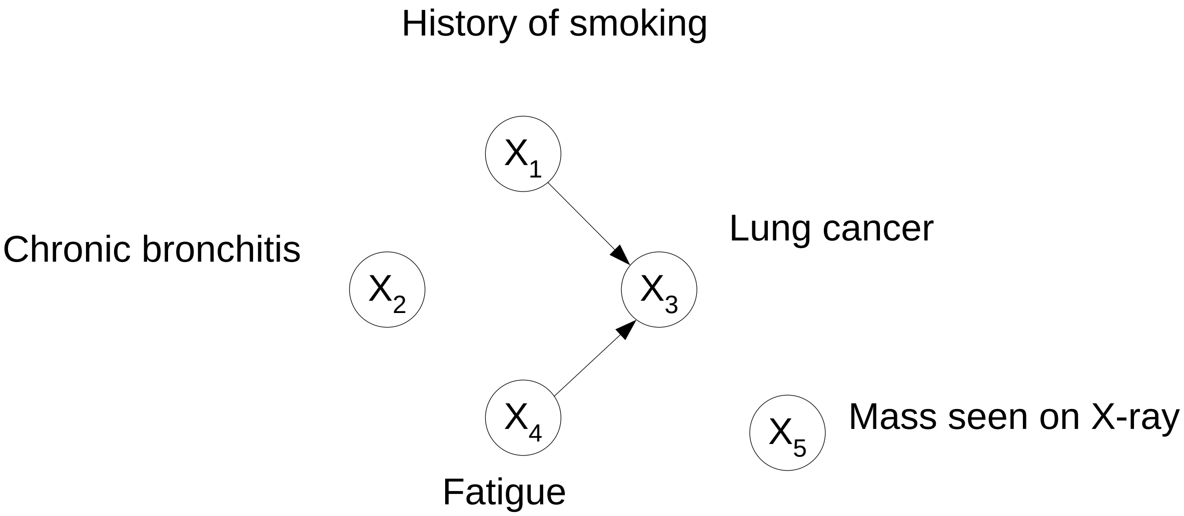
Mass seen on X-ray

Fatigue

Structure learning

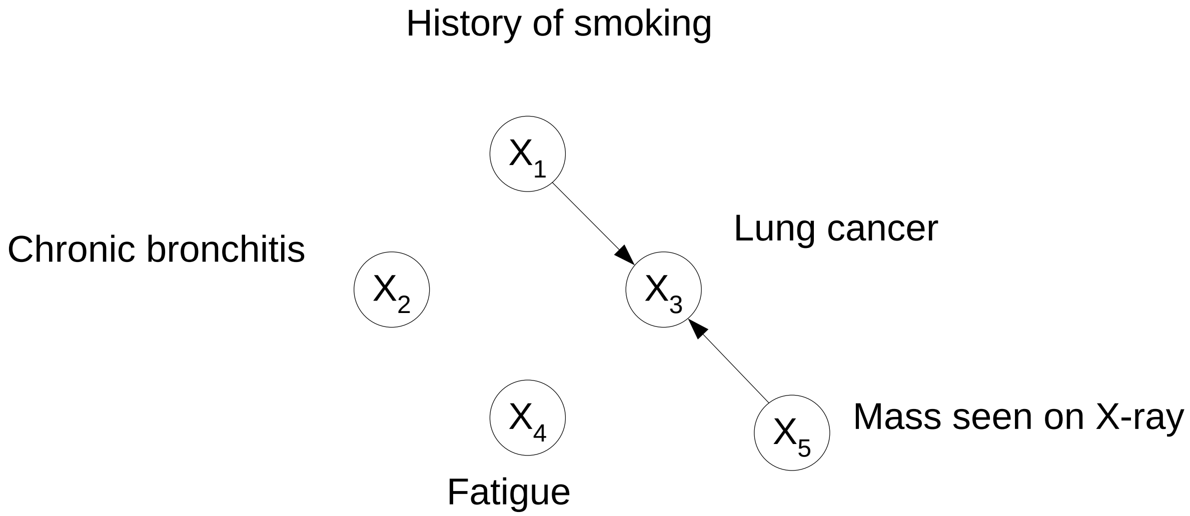


Structure learning



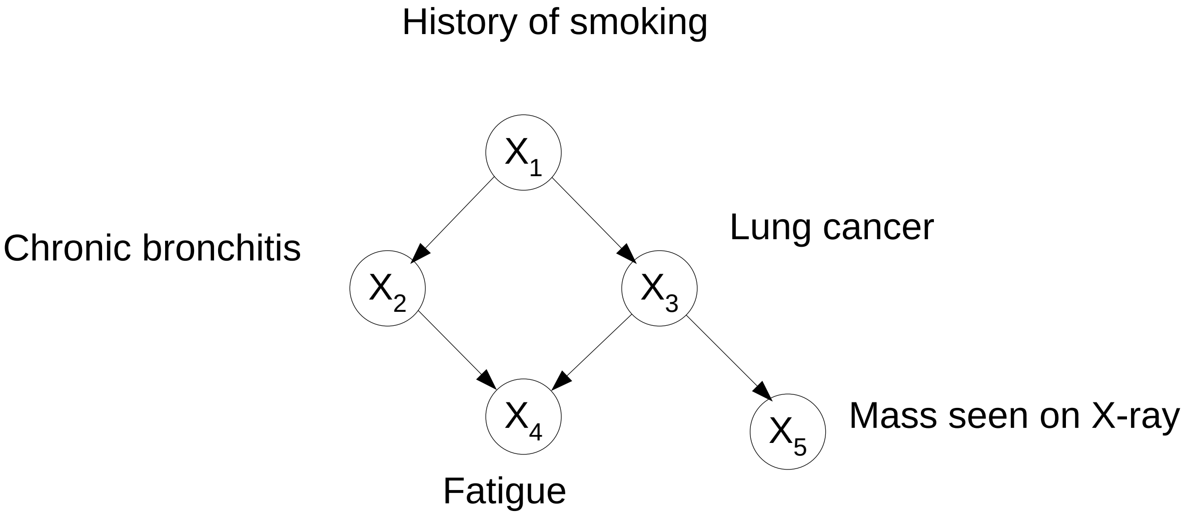
Structure learning

- Second edge does not improve model
- Repeat this for all variables



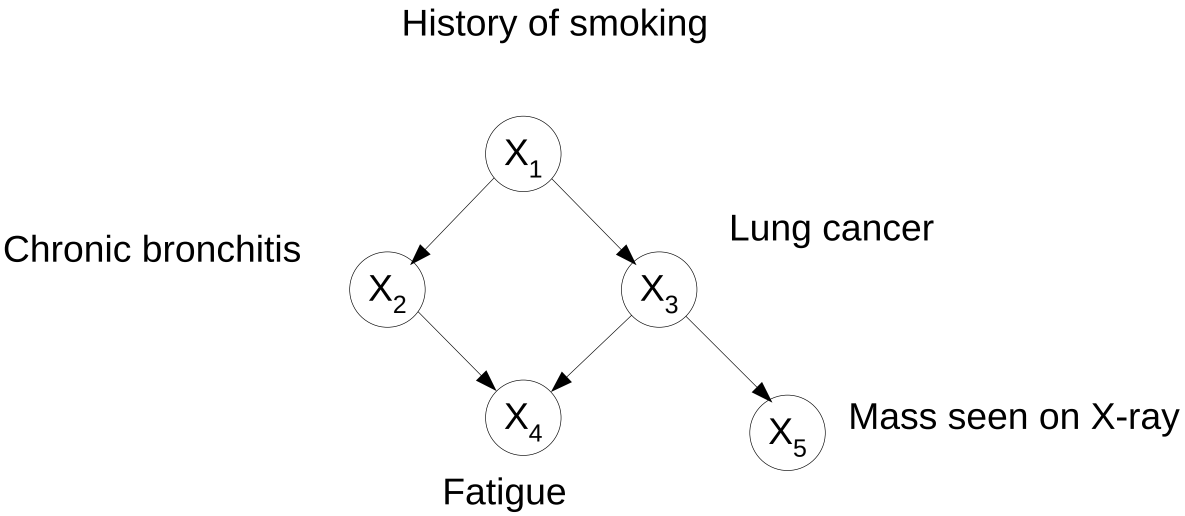
Structure learning

- Second edge does not improve model
- Repeat this for all variables
- Final structure



Parameter learning

- Counting the number of times each situation occurs
- Conditioned on the parents

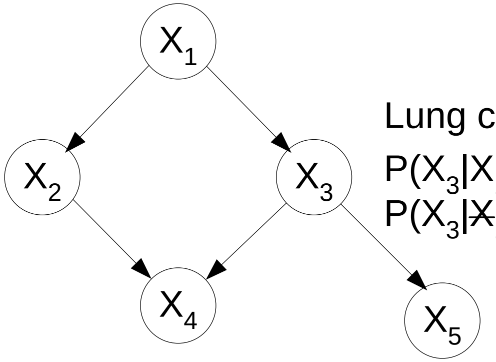


Parameter learning

- Counting the number of times each situation occurs
- Conditioned on the parents

History of smoking

$P(X_1)= 20\%$



Chronic bronchitis

$P(X_2|X_1)= 25\%$
$P(X_2|\cancel{X}_1)= 5\%$

Lung cancer

$P(X_3|X_1)= 0.3\ \%$
$P(X_3|\cancel{X}_1)= 0.005\ \%$

Mass seen on X-ray

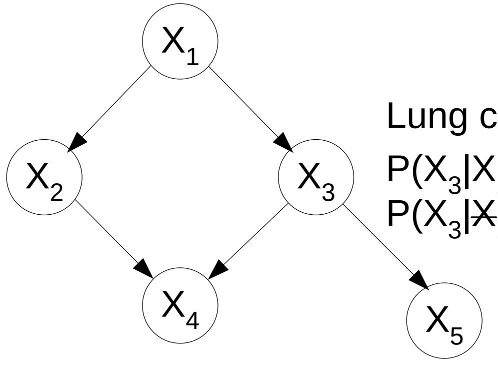
$P(X_5|X_3)= 60\ \%$
$P(X_5|\cancel{X}_3)= 2\ \%$

Fatigue

$P(X_4|X_2,X_3)= 75\%$
$P(X_4|X_2,\cancel{X}_3)= 10\%$
$P(X_4|\cancel{X}_2,X_3)= 50\%$
$P(X_4|\cancel{X}_2,\cancel{X}_3)= 5\%$

Prediction

- Predict the presence of lung cancer on new patients



History of smoking

$P(X_1)= 20\%$

Chronic bronchitis

$P(X_2|X_1)= 25\%$

$P(X_2|\cancel{X}_1)= 5\%$

Lung cancer

$P(X_3|X_1)= 0.3\ \%$

$P(X_3|\cancel{X}_1)= 0.005\ \%$

Fatigue

$P(X_4|X_2,X_3)= 75\%$

$P(X_4|X_2,\cancel{X}_3)= 10\%$

$P(X_4|\cancel{X}_2,X_3)= 50\%$

$P(X_4|\cancel{X}_2,\cancel{X}_3)= 5\%$

Mass seen on X-ray

$P(X_5|X_3)= 60\ \%$

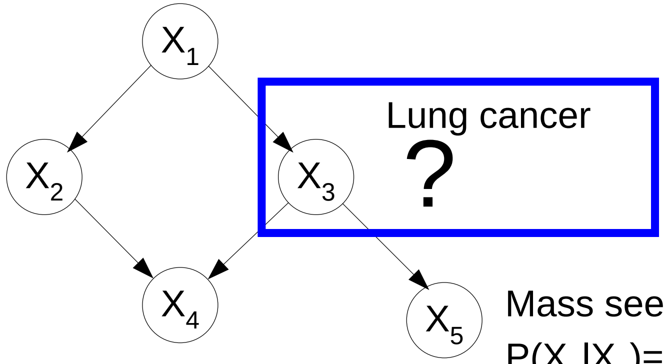
$P(X_5|\cancel{X}_3)= 2\ \%$

Prediction

- Predict the presence of lung cancer on new patients
- New data where the presence of lung cancer is not known

History of smoking

$P(X_1)= 20\%$



Chronic bronchitis

$P(X_2|X_1)= 25\%$
$P(X_2|\cancel{X}_1)= 5\%$

Mass seen on X-ray

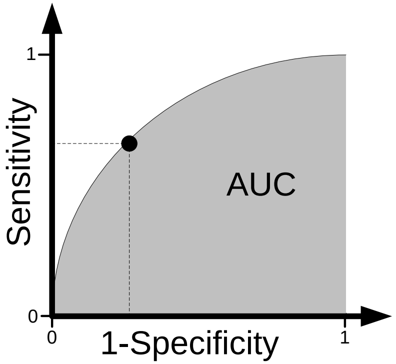
$P(X_5|X_3)= 60\ \%$
$P(X_5|\cancel{X}_3)= 2\ \%$

Fatigue

$P(X_4|X_2,X_3)= 75\%$
$P(X_4|X_2,\cancel{X}_3)= 10\%$
$P(X_4|\cancel{X}_2,X_3)= 50\%$
$P(X_4|\cancel{X}_2,\cancel{X}_3)= 5\%$

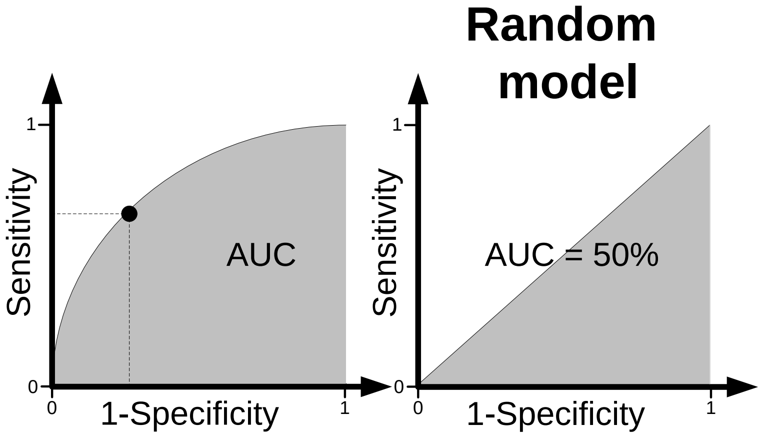
Performance evaluation

- By comparing the predictions with the true value we can evaluate if the model has a good performance
- Area Under the ROC curve



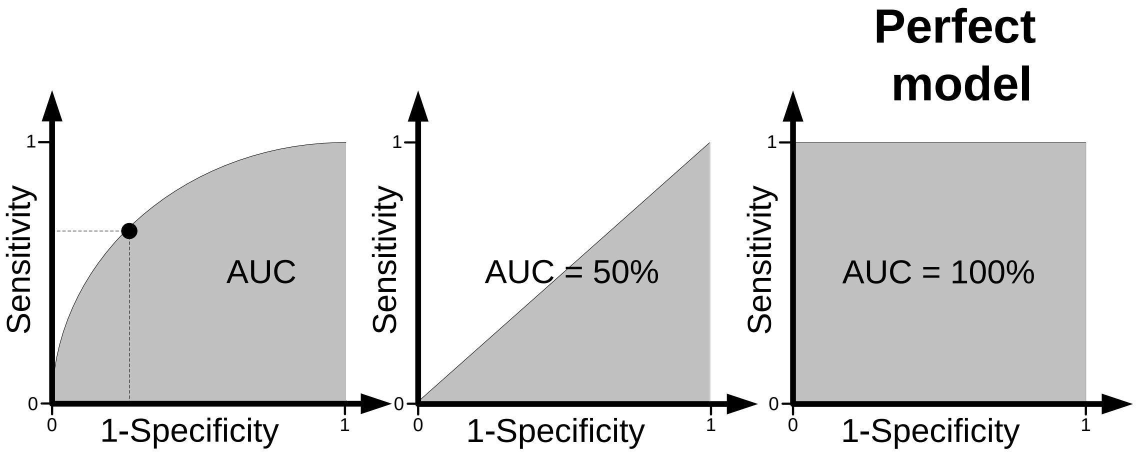
Performance evaluation

- By comparing the predictions with the true value we can evaluate if the model has a good performance
- Area Under the ROC curve
  - AUC of a random model is 50%



Performance evaluation

- By comparing the predictions with the true value we can evaluate if the model has a good performance
- Area Under the ROC curve
  – AUC of a random model is 50%
  – AUC of a perfect model is 100%



Overview

- Motivation
- Bayesian networks
- Results
  - Aim 1: modeling primary data
    - Case 1: Clinical data
    - Case 2: Genomic data
  - Aim 2: integrating primary data
    - Case 1: integrating clinical and microarray data
    - Case 2: integrating microarray and proteomics data
  - Aim 3: integrating secondary data
- Conclusions
- Future work

Clinical data

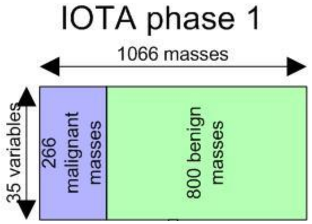
The IOTA project

Benign vs. malignant ovarian masses

Clinical Data

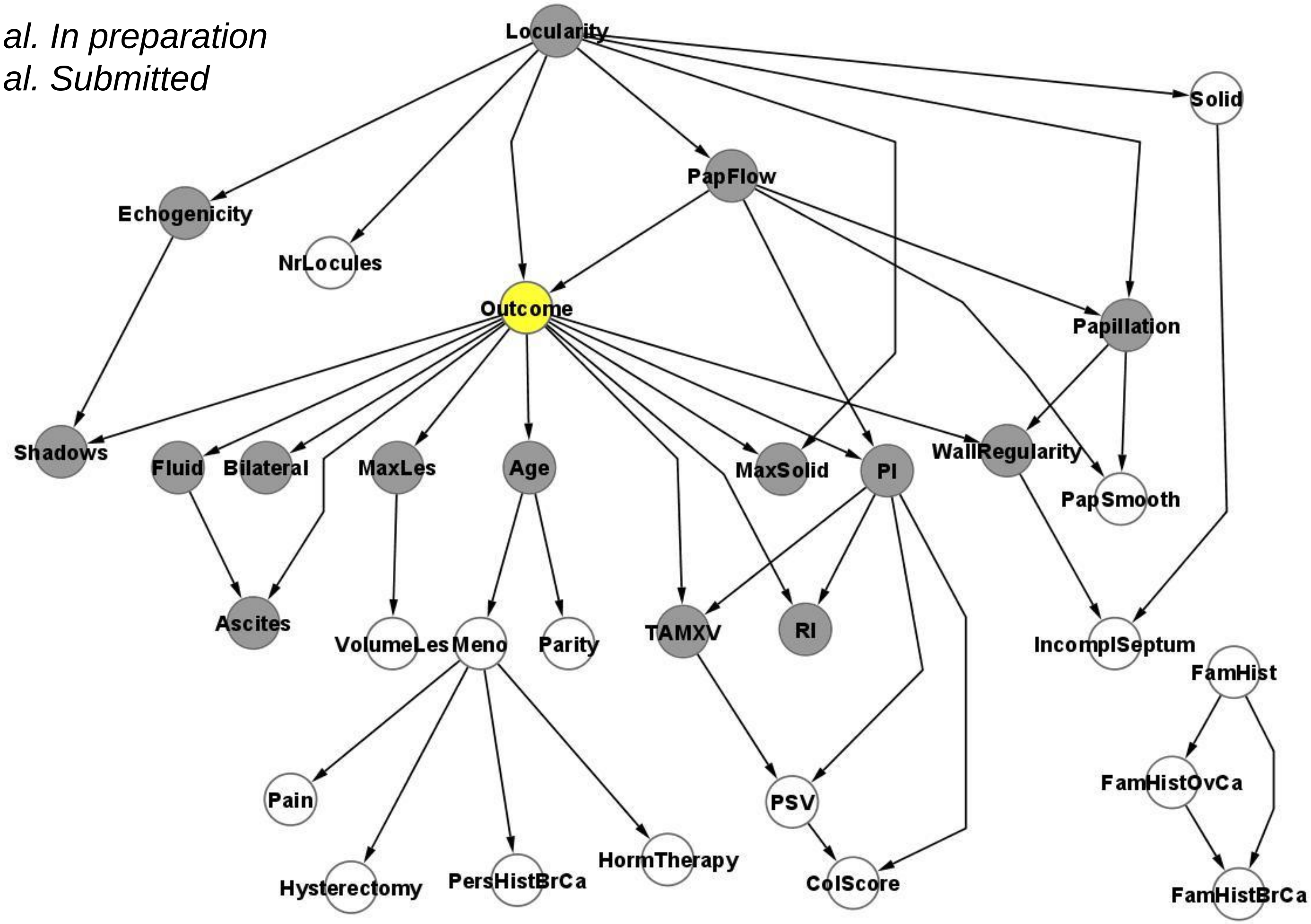
- Data gathered by the International Ovarian Tumor Analysis consortium (IOTA)
  - Standardized multi-centric collection of clinical data
  - Aim predict malignancy of ovarian masses based on clinical data
  - > 60 variables collected, 32 selected relevant for prediction
- Data gathered in three phases:
  - Phase 1: 1066 patients in 9 European centers
  - Phase 1b: 507 patients in 3 centers (internal validation)
  - Phase 2: 1938 patients in 19 International centers (old and new).

Overview

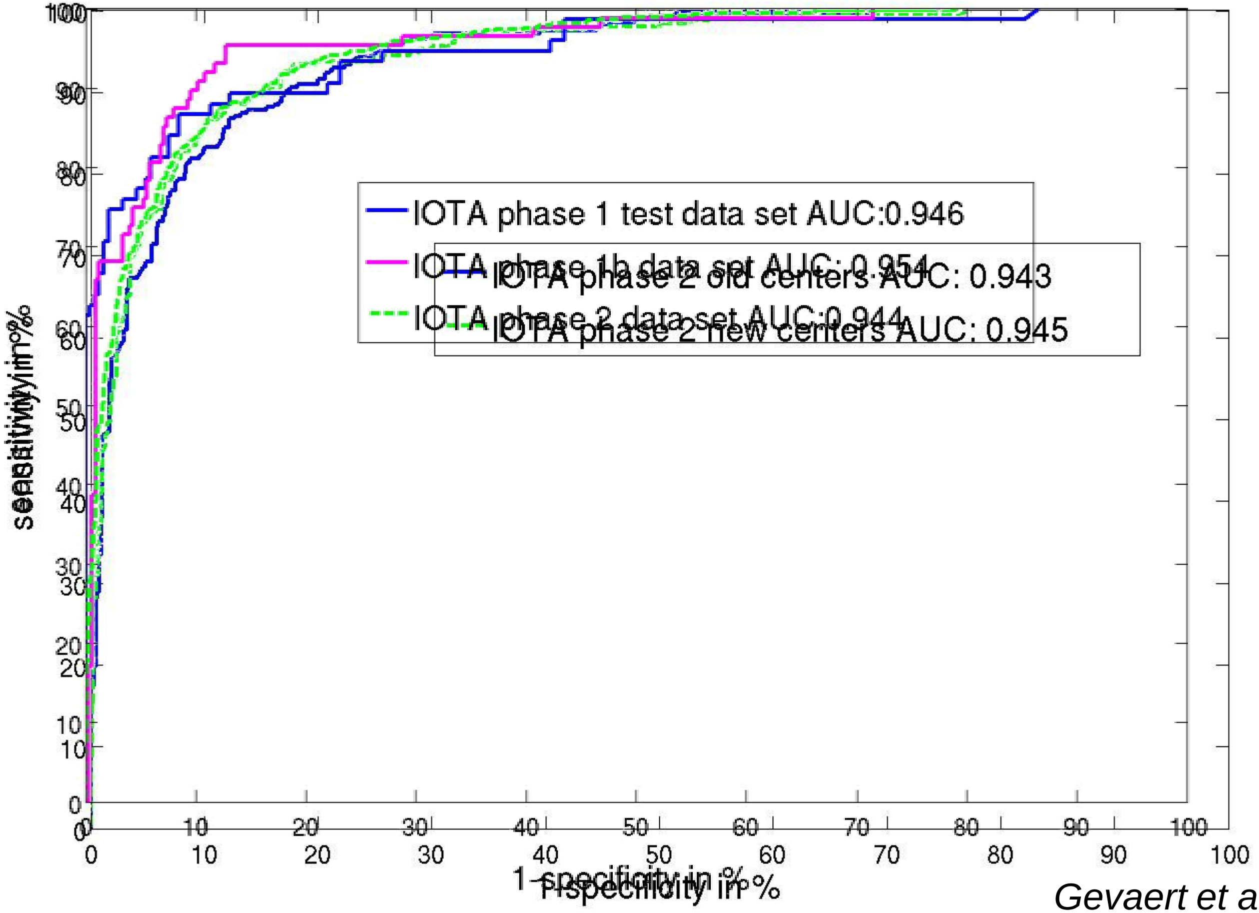

IOTA phase 1
1066 masses
35 variables
266 malignant masses
800 benign masses

Results

Gevaert et al. In preparation
Gevaert et al. Submitted



Results



Gevaert et al. In preparation
Gevaert et al. Submitted

Comparison with Logistic regression

Data set	BN1	LR1	LR2
IOTA phase 1 test data	0.946	0.942	0.920
IOTA phase 1b	0.954	0.950	0.950
IOTA phase 2	0.944	0.951	0.934
IOTA phase 2 old	0.943	0.945	0.918
IOTA phase 2 new	0.945	0.956	0.949

BN1 Bayesian network
LR1 Logistic regression model with 12 variables
LR2 Logistic regression model with 6 variables

Gevaert et al. In preparation
Gevaert et al. Submitted

Conclusion

- Bayesian networks are an alternative for more traditional modeling of clinical data
- Similar performance compared to logistic regression
- Network allows analysis of relationships between variables

Genomic data

BRCA1-mutated vs. sporadic ovarian cancers

Introduction

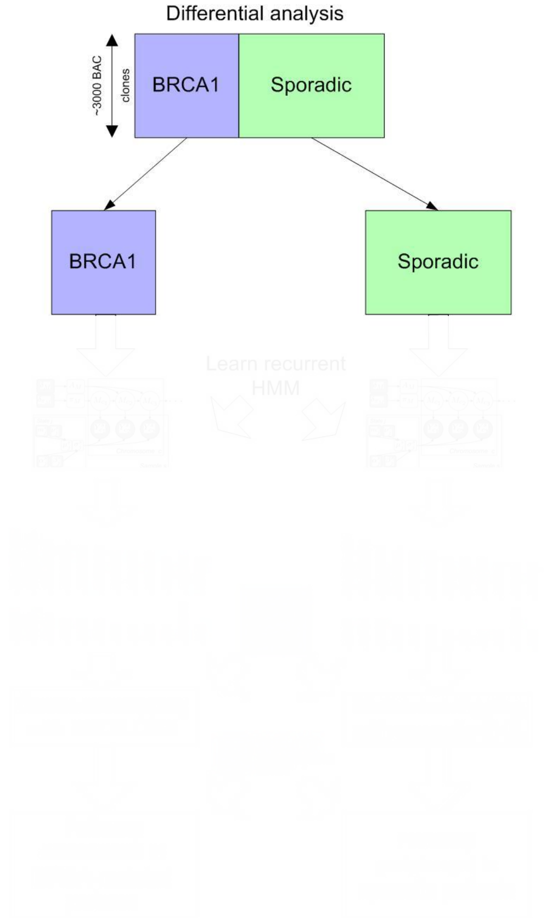
- Approximately 5%-10% of ovarian cancers are caused by inheriting mutations in the BRCA1 or BRCA2 gene
- These BRCA-mutated tumors behave differently compared to the sporadic ovarian cancers
- We investigated if there are differences in the genomes of **BRCA1-mutated** vs. **sporadic ovarian cancers**

Overview

- Tumor samples gathered at the University Hospitals Leuven:
  - 5 BRCA1-mutated ovarian cancers
  - 8 sporadic ovarian cancers
- All 13 samples subjected to arrayCGH technology
- ArrayCGH data model:
  - Subclass of Bayesian networks
  - Recurrent Hidden Markov model (RHMM)
  - To discover recurrent Copy Number Alterations (CNA)

Overview

- RHMM modeling both groups separately
- This results in the identification of recurrent CNA genome wide
- Extract genes from Ensembl database
- Pathway enrichment

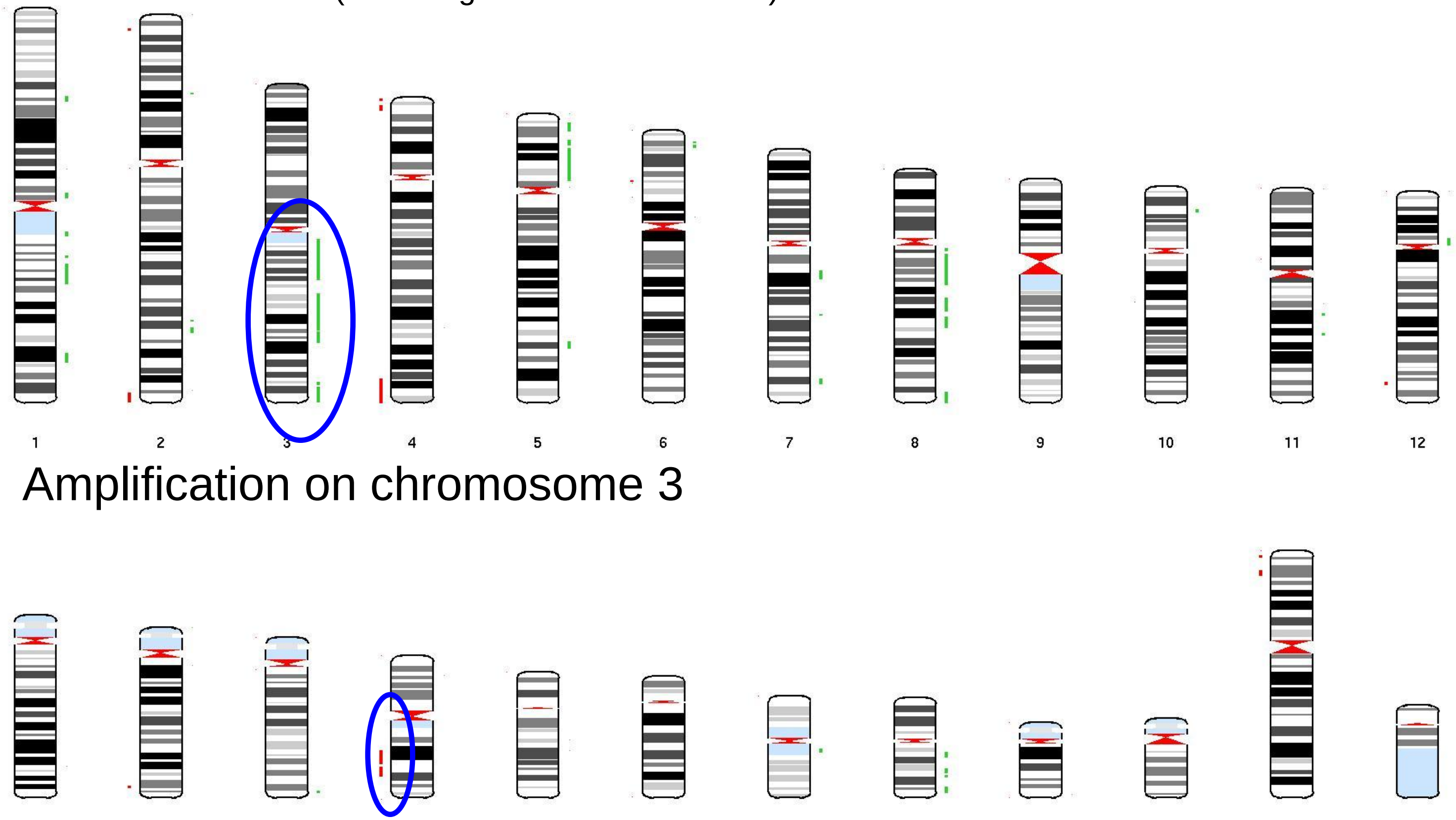


Results: sporadic genome

~ 475 Mb aberrated or 15%

(384 Mb gained and 91 Mb lost)

Leunen, Gevaert et al. Submitted



Amplification on chromosome 3

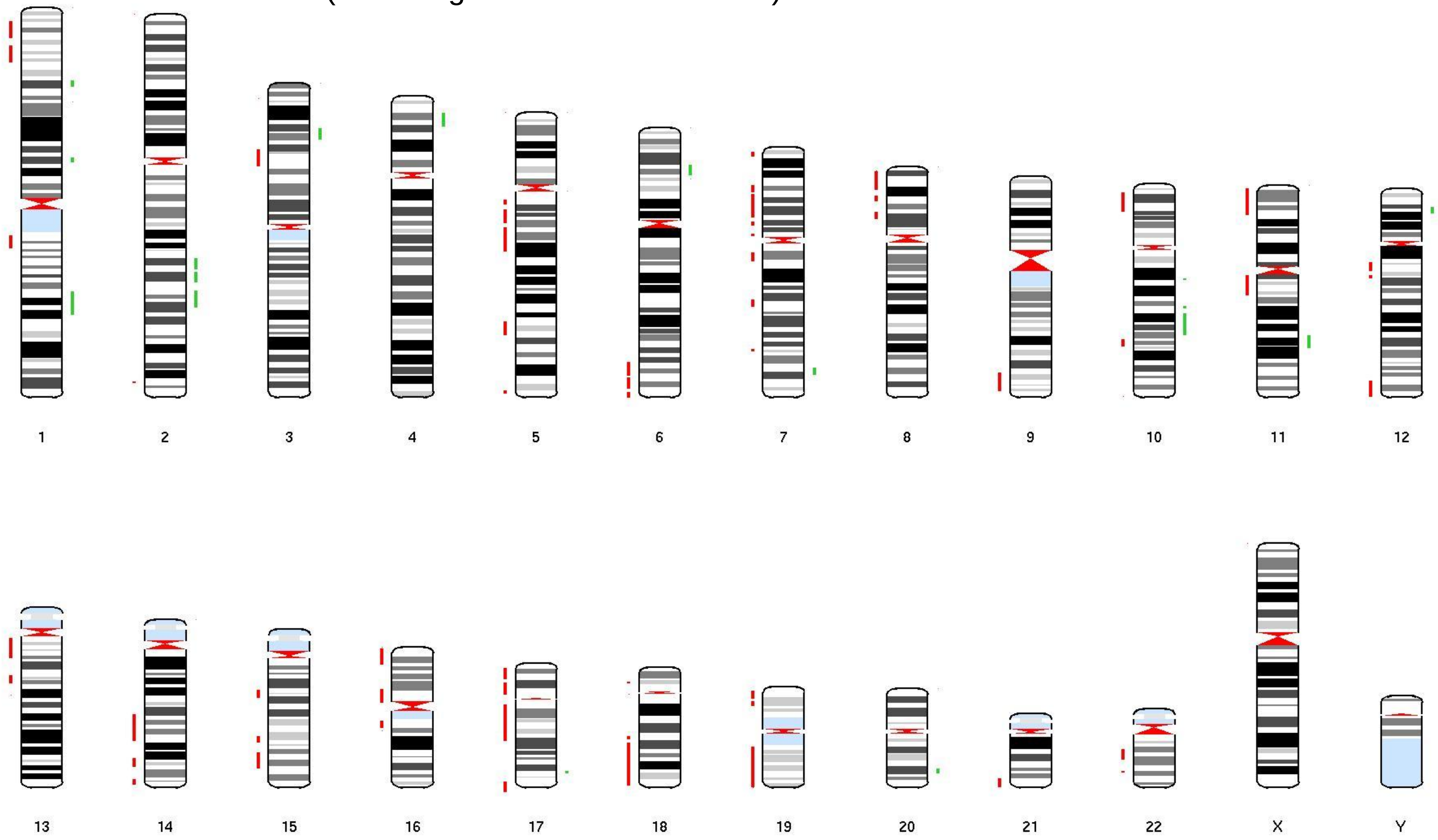
Deletion on chromosome 16

Results: BRCA1 genome

~ 730 Mb aberrated or 22%

(257 Mb gained and 473 Mb lost)

Leunen, Gevaert et al. Submitted

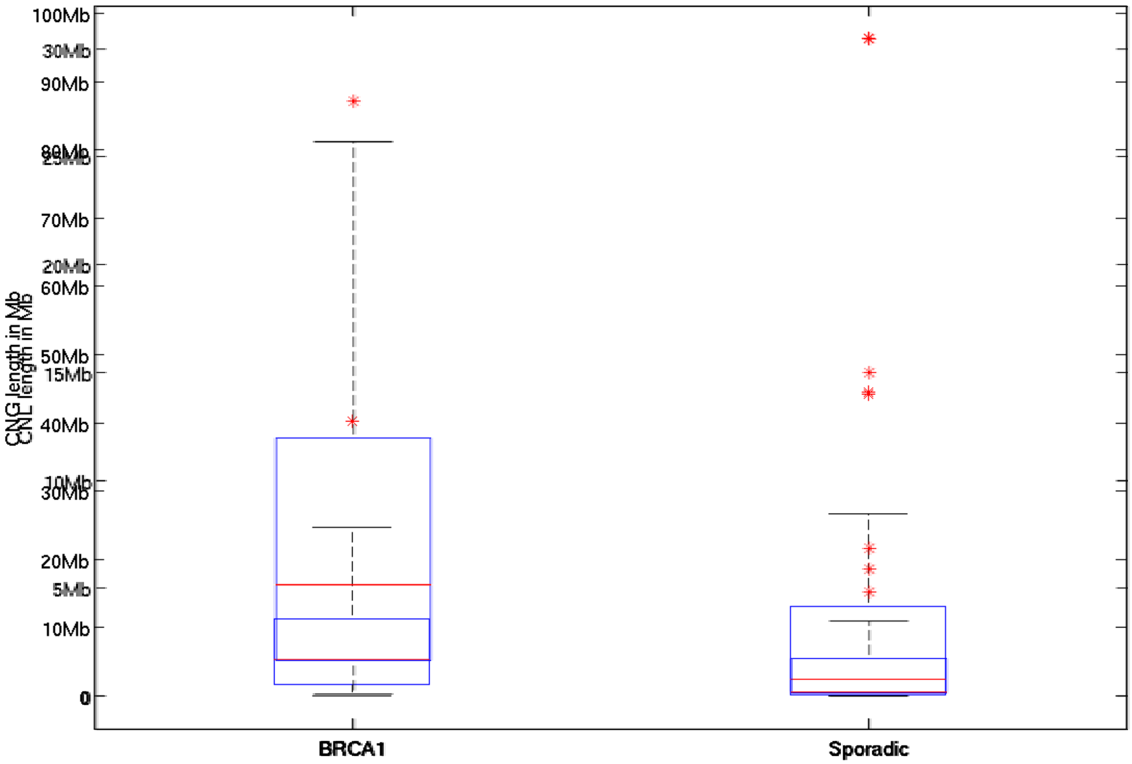


Results

Leunen, Gevaert et al. Submitted

Length of copy number gains

Length of copy number losses

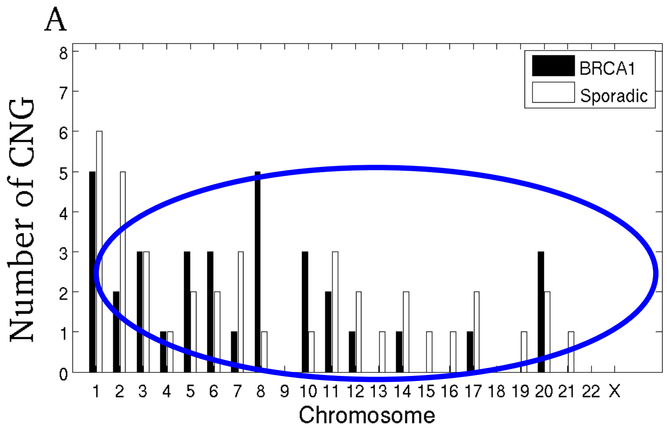


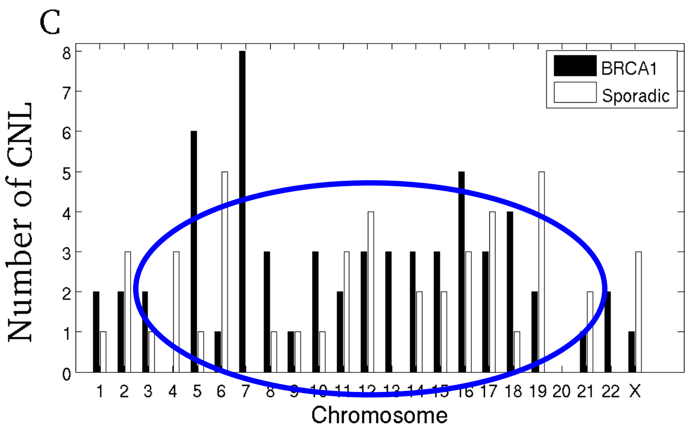
Results per chromosome

GAINS

LOSSES

Number





Length

Results

Pathways enriched in the BRCA1 group

Leunen, Gevaert et al. Submitted

Signature	Gene set name from MSigDB	P-value	Q-value	Overlapping Genes
GAINED	HOX GENES	0.00020	0.08684	HOXD10 HHEX HOXD11 HOXD9 HOXD13 HOXD1 HOXD12 HOXD4 HOXD3
GAINED	MATRIX METALLOPROTEINASES	0.00020	0.08684	MMP3 MMP10 MMP13 MMP27 MMP1 MMP20 MMP7 MMP8 MMP12
LOST	BREAST CANCER ESTROGEN SIGNALING	0.00180	0.09824	SPRR1B CLDN7 TP53 GATA3 ERBB2 CCND1 SCGB1D2 THBS2 C3 KLK5 FOSL1 KRT18 DLC1 KRT19 CTSB IL6ST RPL27 FLRT1 NGFR SERPINE1 IL2RA SCGB2A2 BCL2 HMGB1 SCGB2A1 TNFAIP2 AZGP1 ESR1 EGFR ESR2 RPL13A S100A2 SERPINB5 THBS4 BAD COL6A1 ACTB
LOST	TUMOR SUPRESSOR	0.00200	0.09824	BRCA2 CDKN2D BRCA1 LCMT2 EP300 TSC2 CDKN1C CFL1 TP53 RB1 NF2 CREBBP ACTB
LOST	HOX GENES	0.00223	0.09852	HOXA6 CBX8 LHX2 HOXB5 HOXB13 HOXA5 EZH1 HOXA2 HOXA4 PHC2 HOXA11 HOXA1 CBX4 HOXB3 HOXA3 DLX4 HOXA10 HOXB2 HOXB7 HOXA7 HOXB1 HOXB9 HOXA9 HOXB6

Conclusion

- Complex but powerful modeling strategies allows to identify recurrent CNAs
- CNAs from the two groups of patients are different
- Different pathways enriched
- We hypothesize that BRCA1-mutated tumors are driven by different biological processes and may benefit from different therapy strategies.

Aim 2: Integration of primary data sources

Data

- Case 1: Integration of clinical and microarray data
  - van ‘t Veer data set
- Case 2: Integration of microarray and proteomics data
  - Rectal cancer data set (University Hospitals Leuven)

Case 1: van't Veer

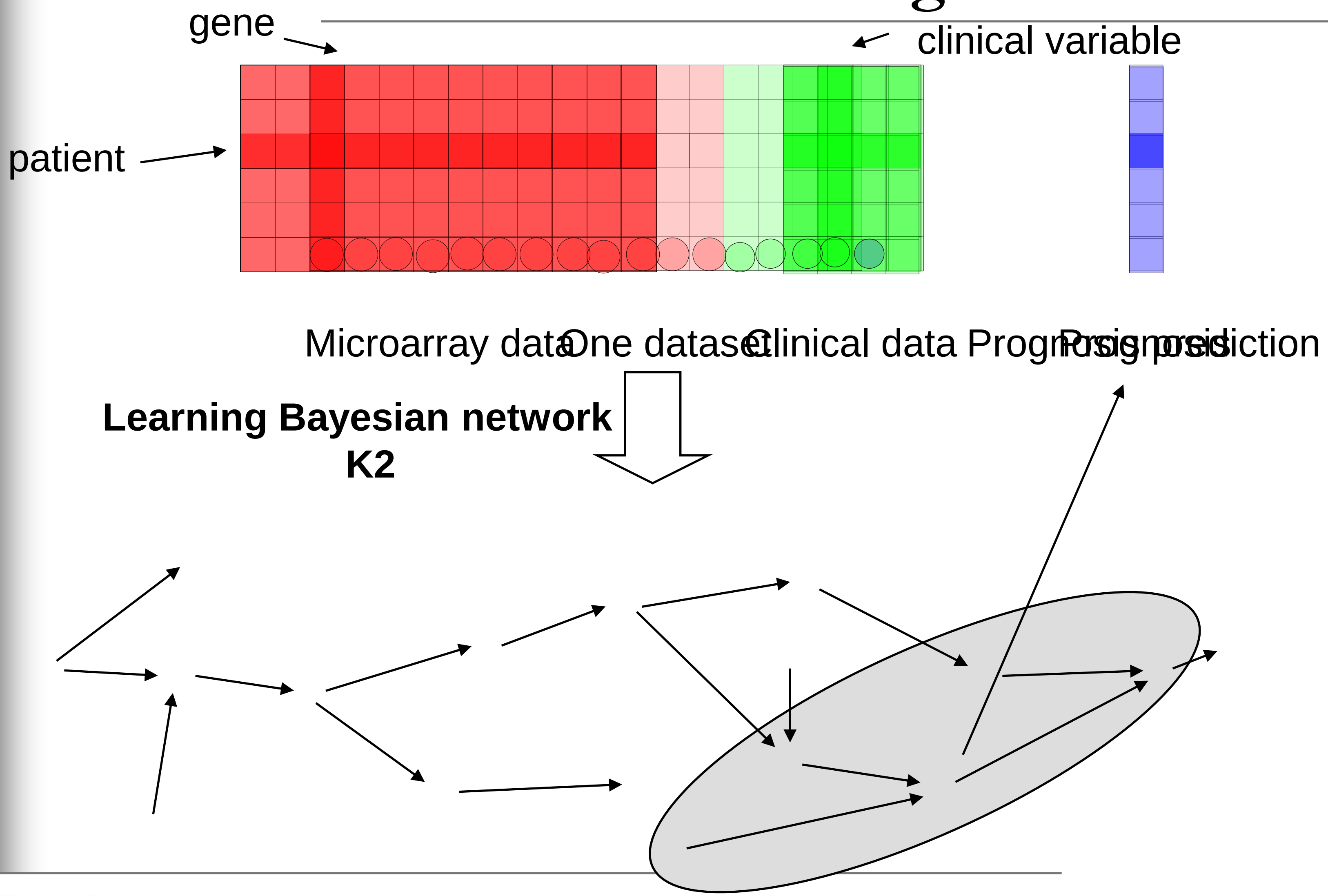
- Breast cancer microarray data *van 't Veer et al. Nature 2002*
- Microarray data consisted of ~20000 genes
- Clinical data consists of 7 variables:
  - age, diameter, grade, angioinvasion, ERP, PRP, lymphocytic infiltration
- Binary outcome variable had two states:
  - good prognosis (disease free interval of at least 5 years)
  - poor prognosis (recurrence within 5 years)

Data integration

- We have defined different methods for integrating both data sources with Bayesian networks
  - Full integration
  - Decision integration
  - Partial integration

- The difference between these methods lies “when” the data integration takes place

Full integration



gene

clinical variable

patient

Microarray data

One dataset

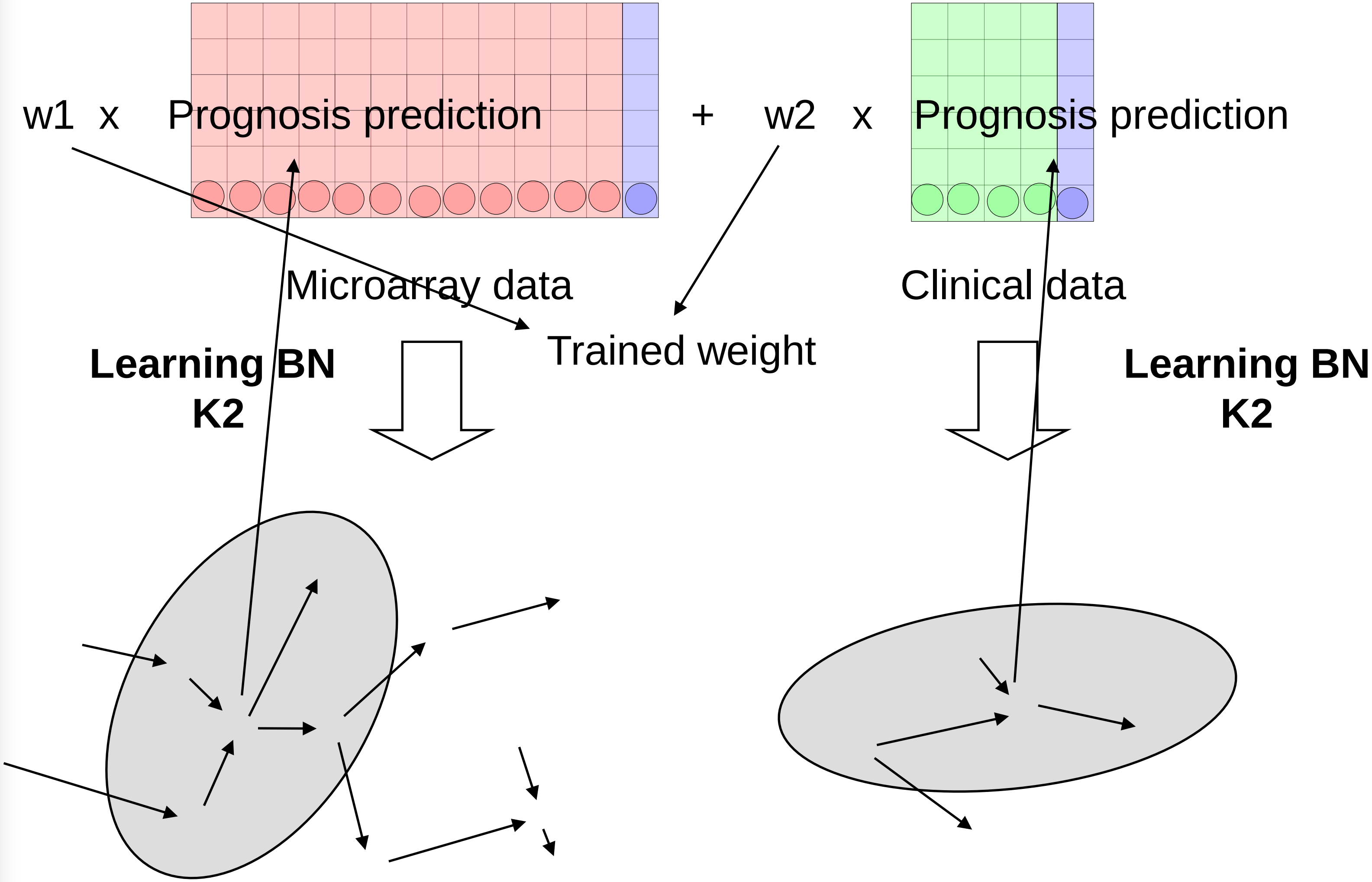
Clinical data

Prognosis

Prognosis prediction

Learning Bayesian network K2

Decision integration



w1 x Prognosis prediction + w2 x Prognosis prediction

Microarray data

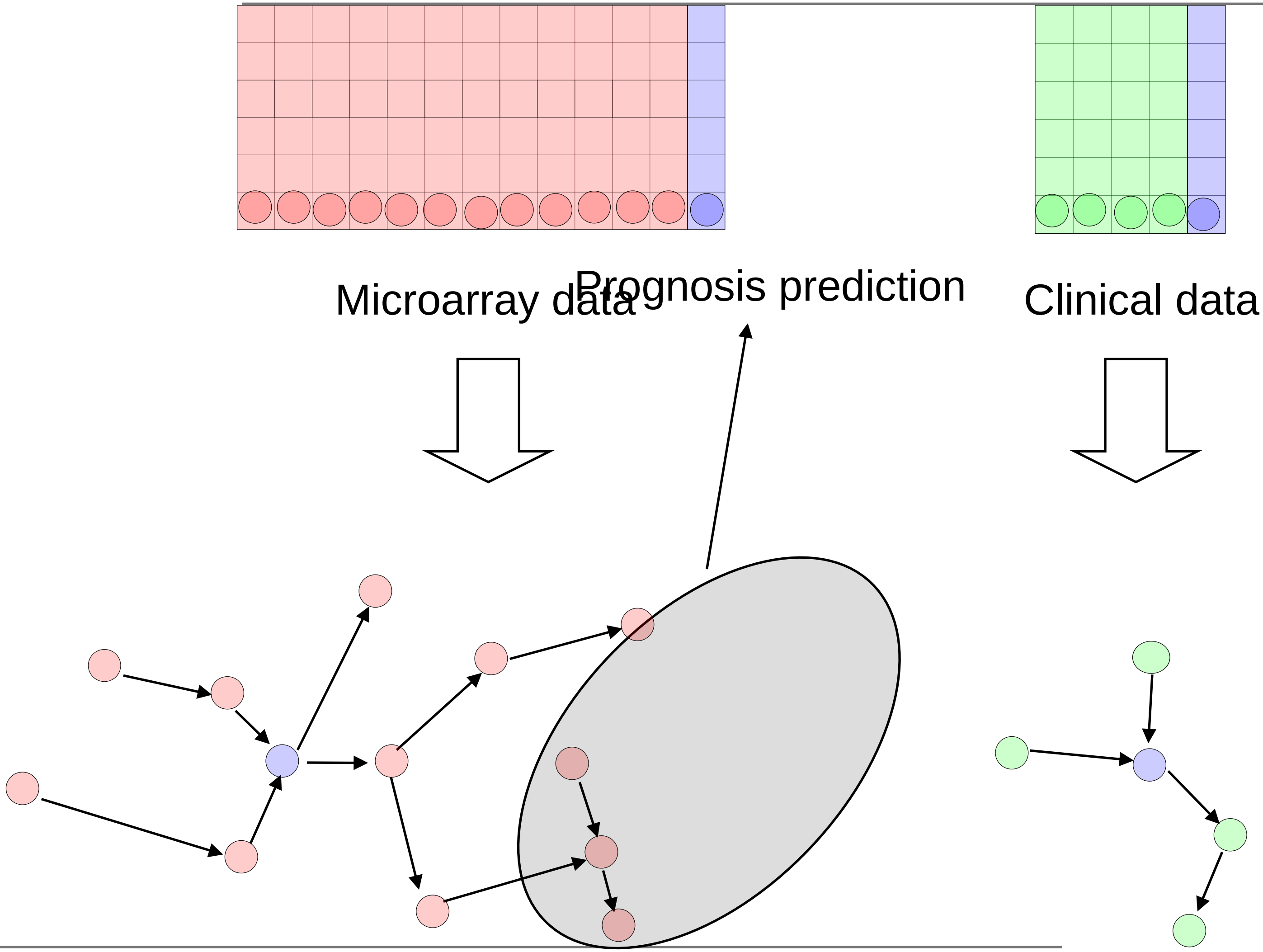
Clinical data

Trained weight

Learning BN K2

Learning BN K2

Partial integration



Results

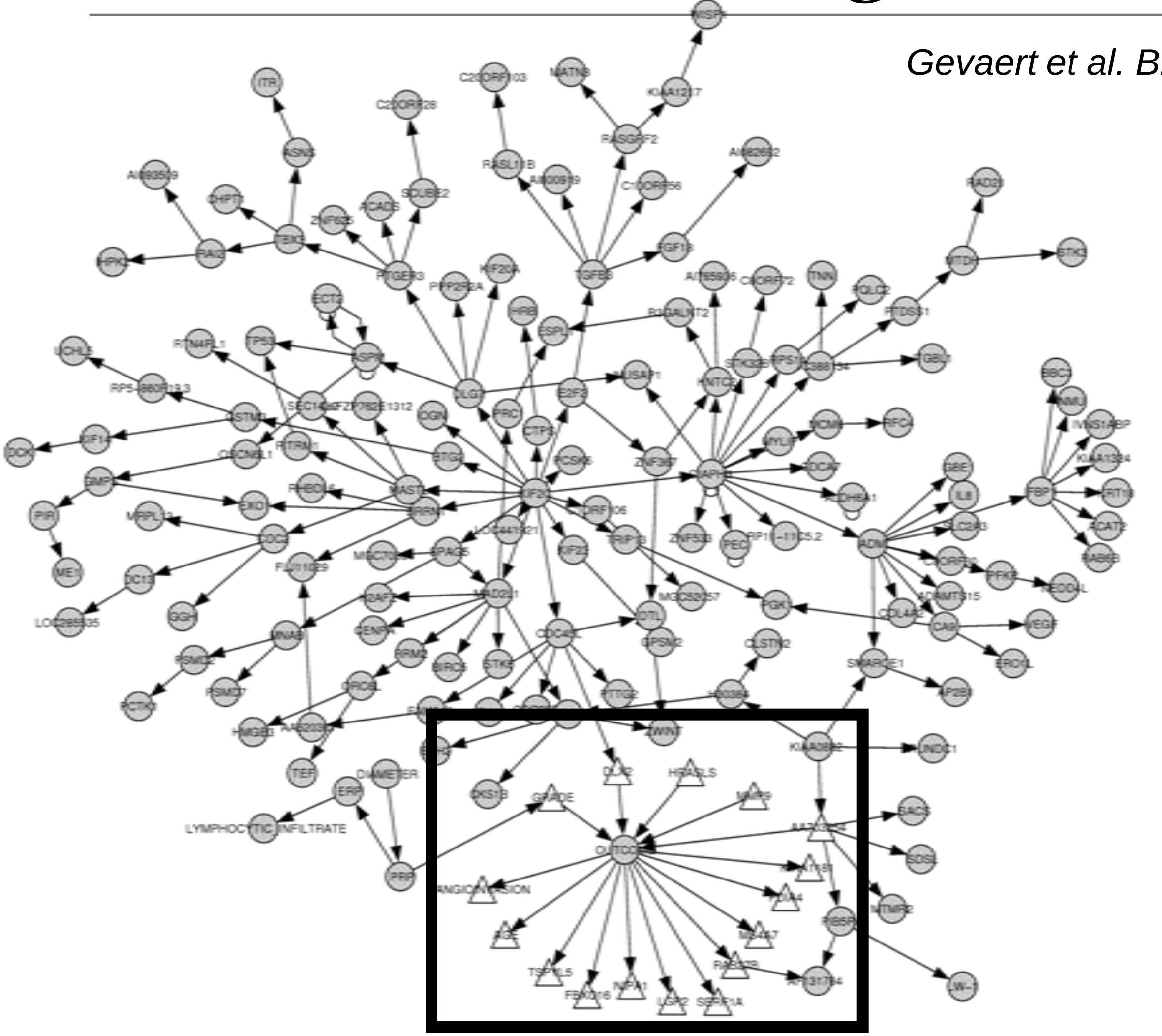
- Partial integration performs best
- Full integration is not better than either data source separately

Method	AUC	Std
Clinical data	0.751	0.086
Microarray data	0.750	0.073
Decision integration	0.773	0.071
Partial integration	0.793	0.068
Full integration	0.747	0.099

Gevaert et al. Bioinformatics 2006

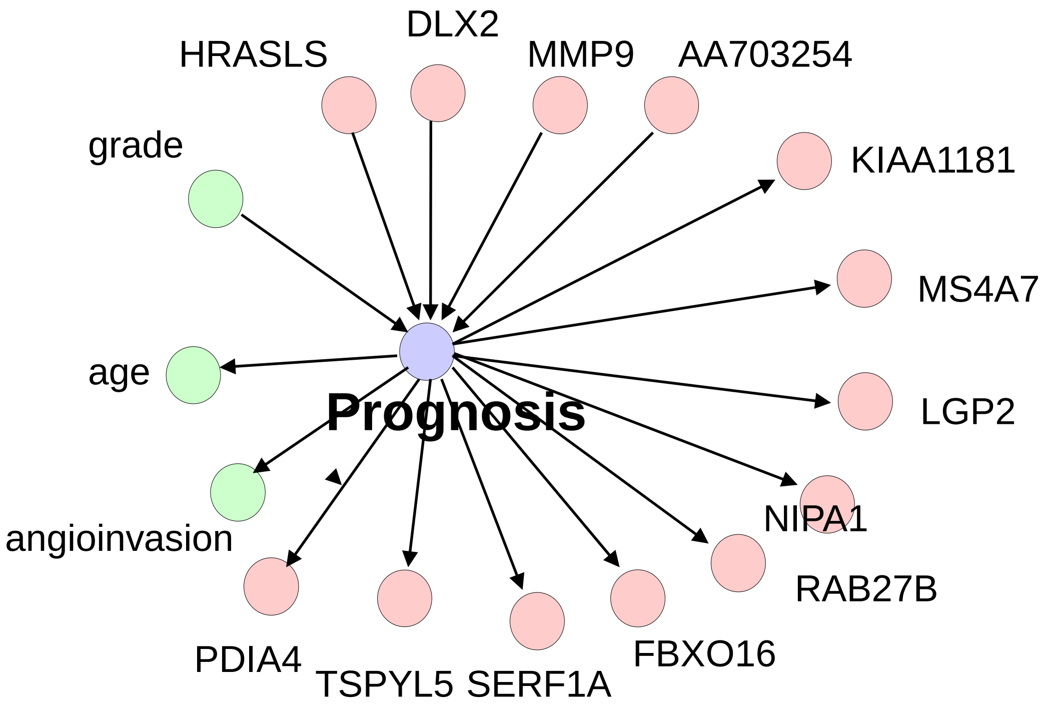
Partial integration

Gevaert et al. Bioinformatics 2006



Results

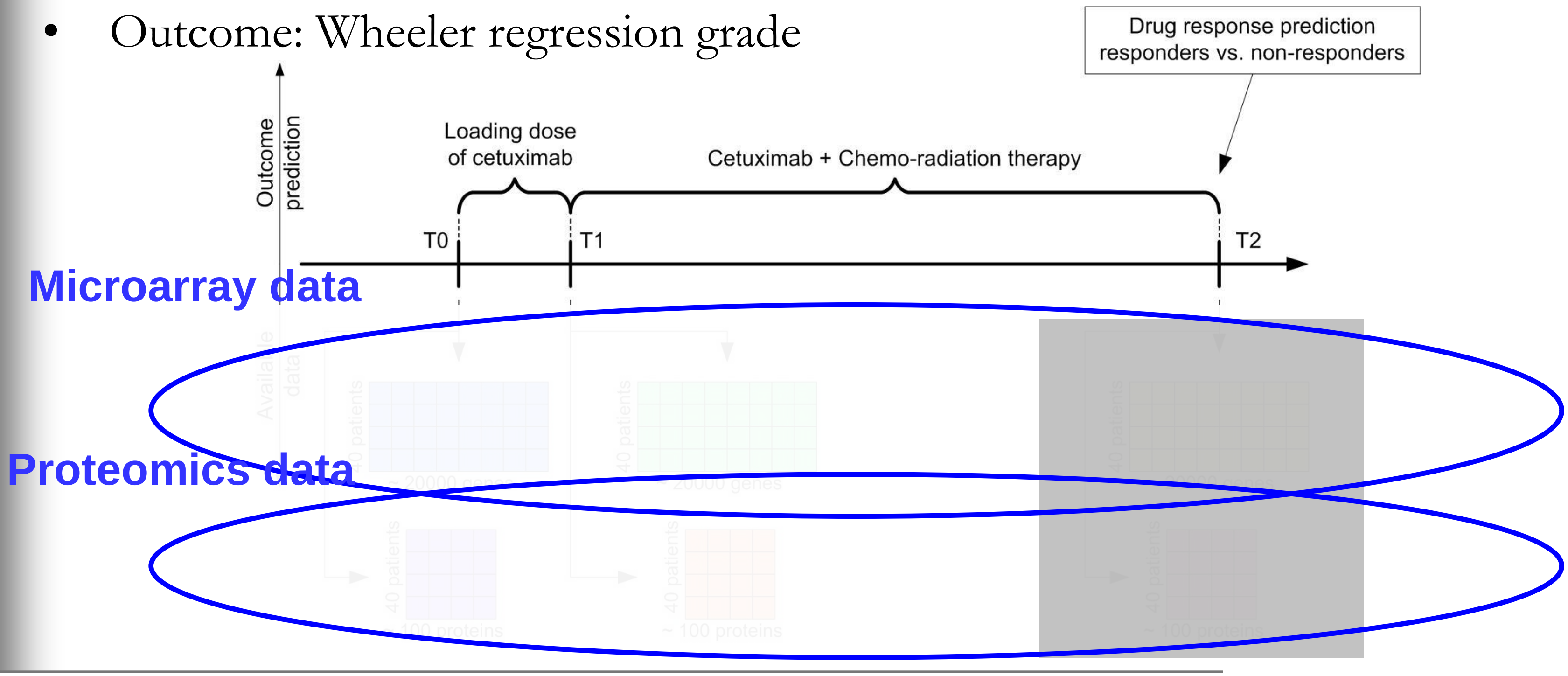
- 3 clinical variables
- 13 genes



Gevaert et al. Bioinformatics 2006

Case 2: rectal cancer

- Rectal cancer therapy timeline:
  - T0: start of therapy
  - T1: after 1 loading dose of cetuximab
  - T2: before surgery
- Outcome: Wheeler regression grade



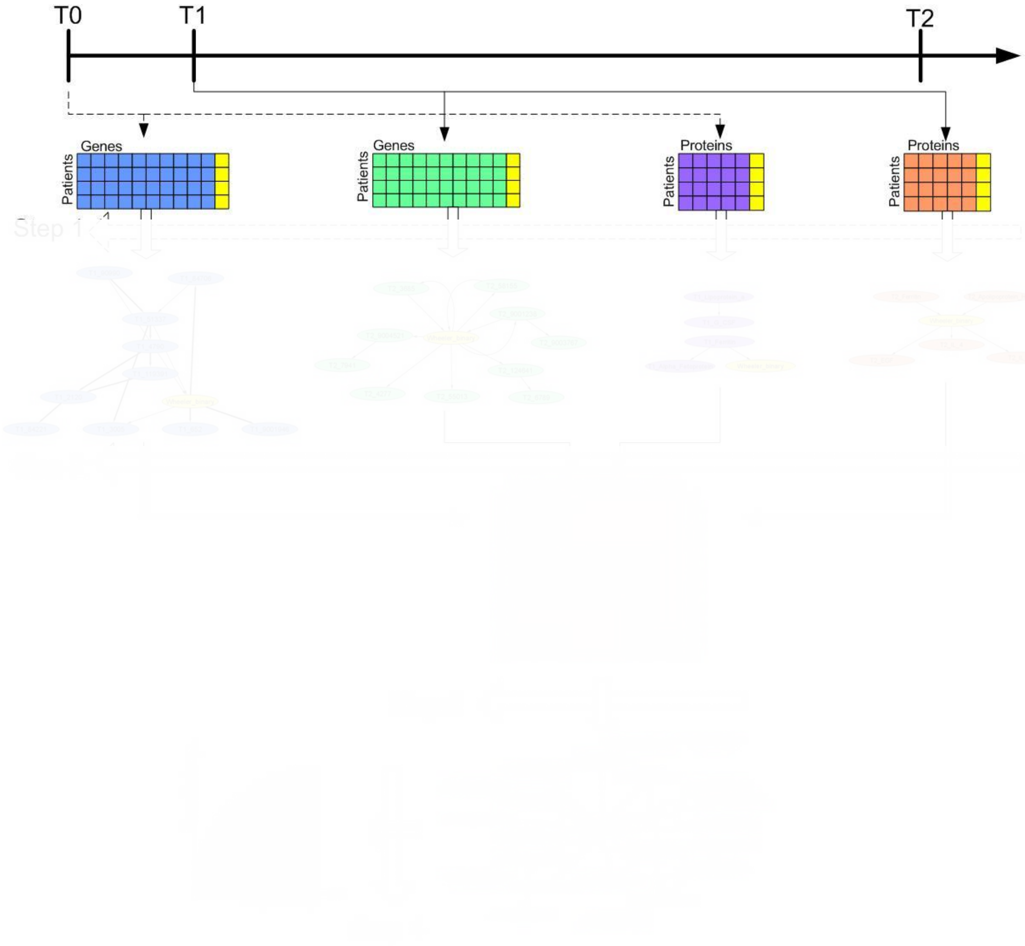
Drug response prediction responders vs. non-responders

Microarray data

Proteomics data

Overview

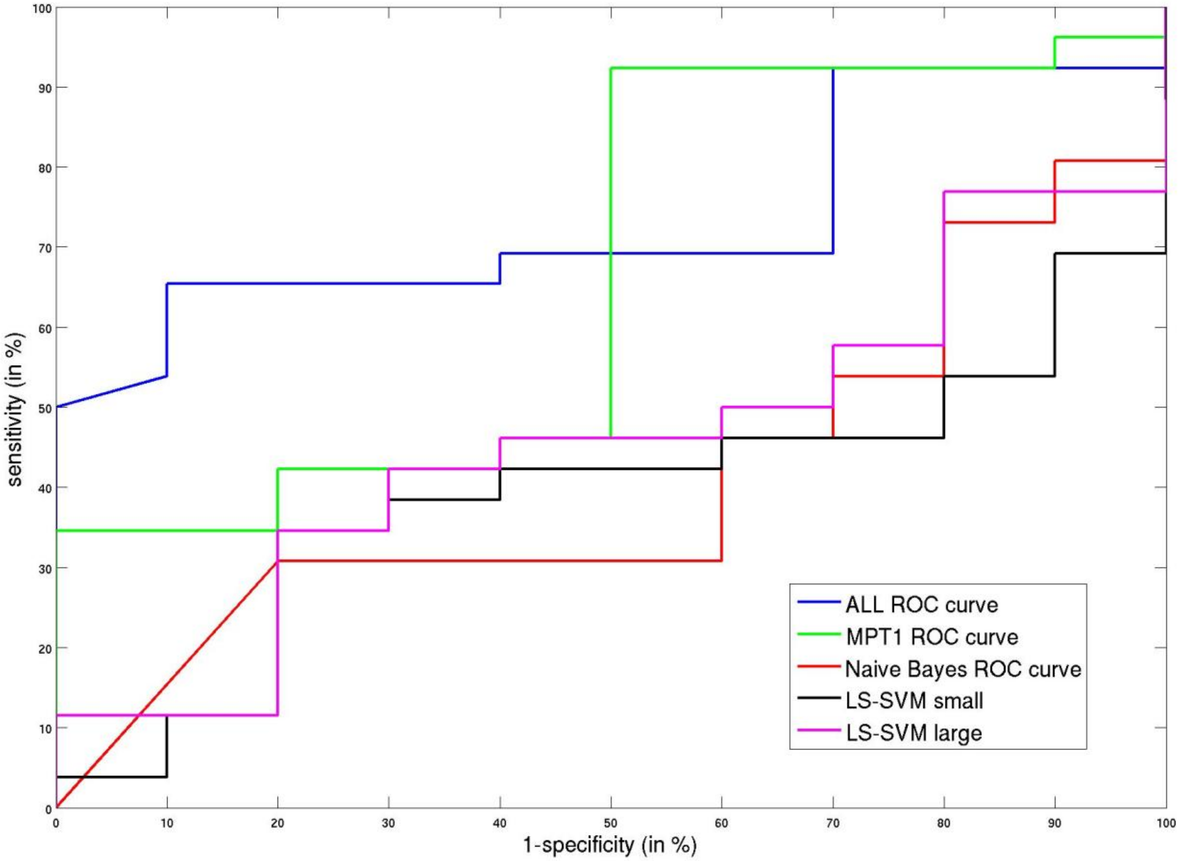
- Partial integration -> Bayesian integration
  - Step 1: represent each data source with its posterior distribution
  - Step 2: integrate posterior in a structure prior
  - Step 3: learn integrated network
  - Step 4: estimate predictive performance



Results

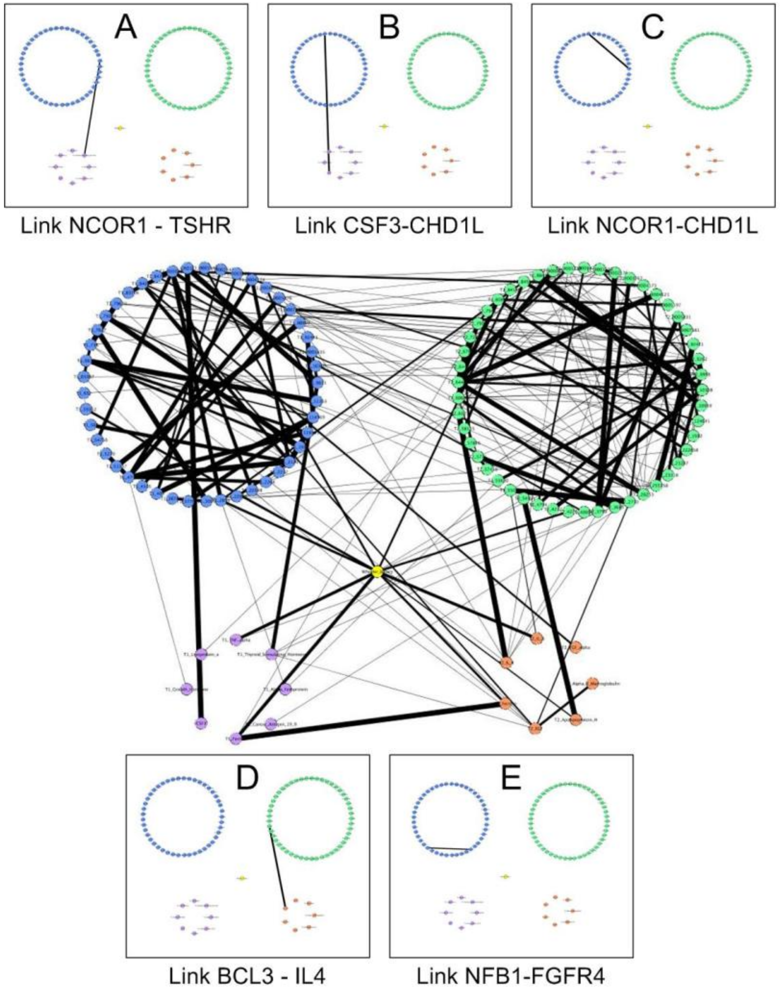
Model abbreviation	AUC	SE
ALL	0.73	0.08
MPT0	0.23	0.09
MPT1	0.67	0.1
MT0T1	0.54	0.11
PT0T1	0.55	0.12
MT0	0.41	0.1
MT1	0.55	0.11
PT0	0.49	0.11
PT1	0.57	0.1
Partial integration	0.61	0.11
Full integration	0.51	0.1
Naïve Bayes	0.41	0.1
LS-SVM small	0.39	0.1
LS-SVM large	0.45	0.1

Results ROC curve



Results

- Thickness of the edge reflects its confidence
- A, B, C, D and E are links with strong support in literature



Link NCOR1 - TSHR

Link CSF3-CHD1L

Link NCOR1-CHD1L

Link BCL3 - IL4

Link NFB1-FGFR4

Conclusions

- We have developed a Bayesian network integration framework
- The breast cancer and rectal cancer case show that integrating information improves predictive performance.
- Additionally, new biological hypothesis are generated

Aim 3: Integration of secondary data sources

Motivation

- Recently there has been a significant increase of publicly available databases containing secondary data:
  - E.g 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
- Therefore we investigated if literature abstracts in the structure prior of a Bayesian network improved prognosis prediction

Structure prior

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

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

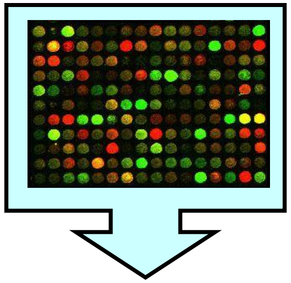
Parameter prior

Structure prior

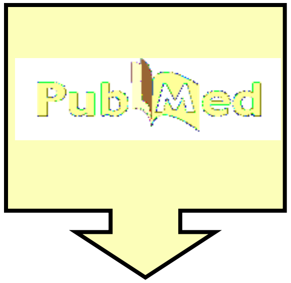
Heckerman, Machine Learning, Vol. 20 (1995), pp. 197-243.

Integration of secondary data

Primary data source
(e.g. Microarray data)

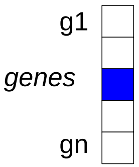


Secondary data source
(e.g. literature abstracts)



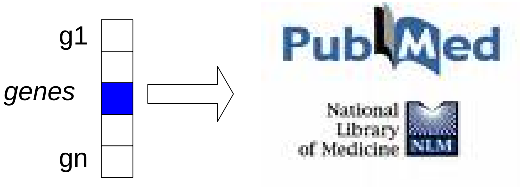
$$\underbrace{p\left(S\middle|D\right)}_{\text{Posterior}} \propto \underbrace{\prod_{i=1}^{n}\prod_{j=1}^{q_i}\left[\frac{\Gamma\left(N'_{ij}\right)}{\Gamma\left(N'_{ij}+N_{ij}\right)}\right]\prod_{k=1}^{r_i}\frac{\Gamma\left(N'_{ijk}+N_{ijk}\right)}{\Gamma\left(N'_{ijk}\right)}}_{\text{Likelihood}}\;\underbrace{P\left(S\right)}_{\text{Prior}}$$

Structure prior construction



Text mining

Structure prior construction



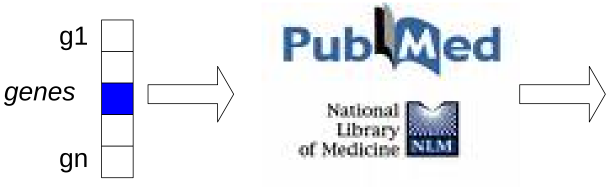
g1

genes

gn

Text mining

Structure prior construction



1: Nature. 2006 Jan 19;439(7074):353-7. Epub 2005 Nov 6.

Comment in:
Nature. 2006 Jan 19;439(7074):274-5.

Oncogenic pathway signatures in human cancers as a guide to targeted t

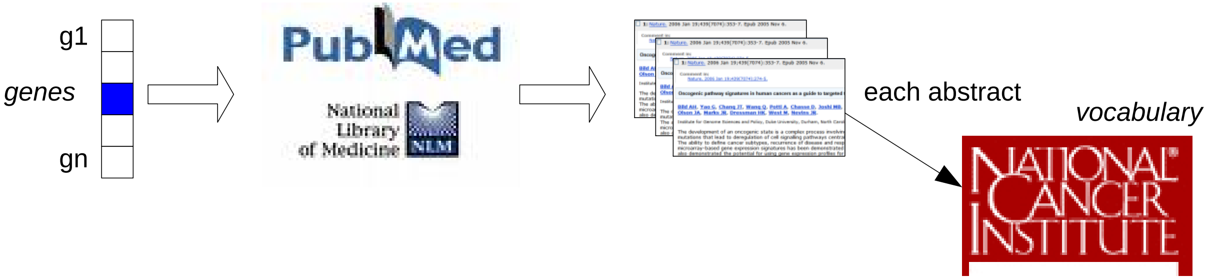
Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Olson JA, Marks JR, Dressman HK, West M, Nevins JR.

Institute for Genome Sciences and Policy, Duke University, Durham, North Caroli

The development of an oncogenic state is a complex process involvin
mutations that lead to deregulation of cell signalling pathways centra
The ability to define cancer subtypes, recurrence of disease and resp
microarray-based gene expression signatures has been demonstrated
also demonstrated the potential for using gene expression profiles for

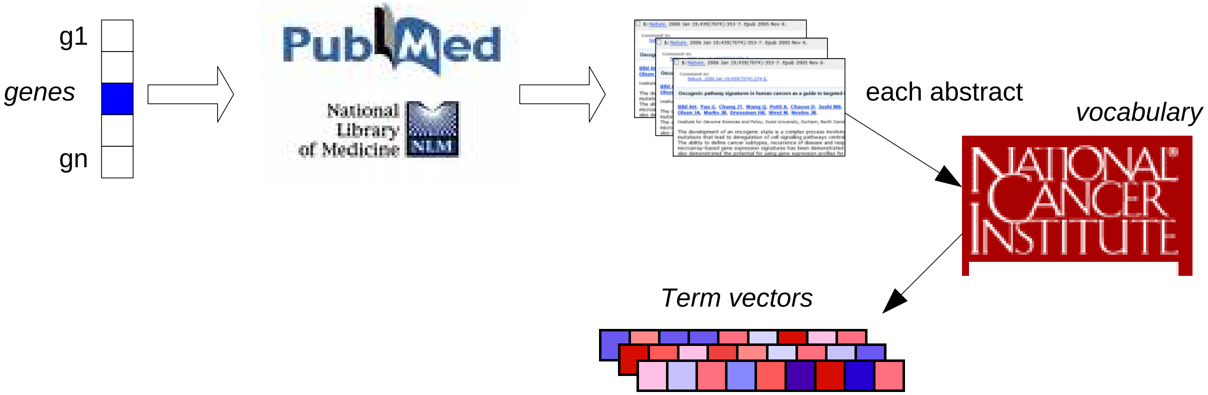
Text mining

Structure prior construction



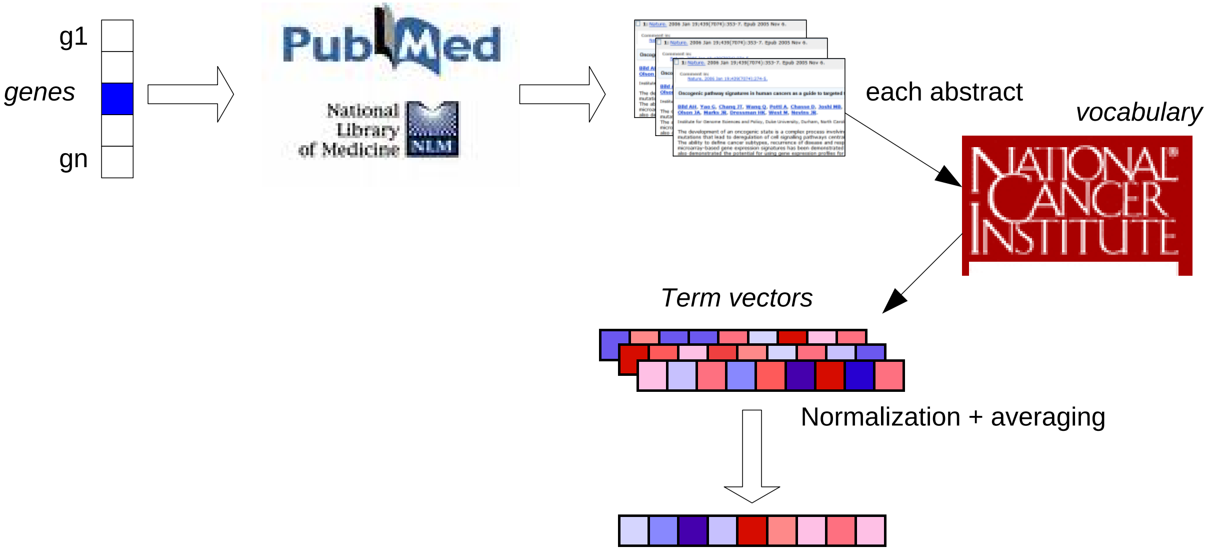
Text mining

Structure prior construction



Text mining

Structure prior construction



g1

genes

gn

each abstract

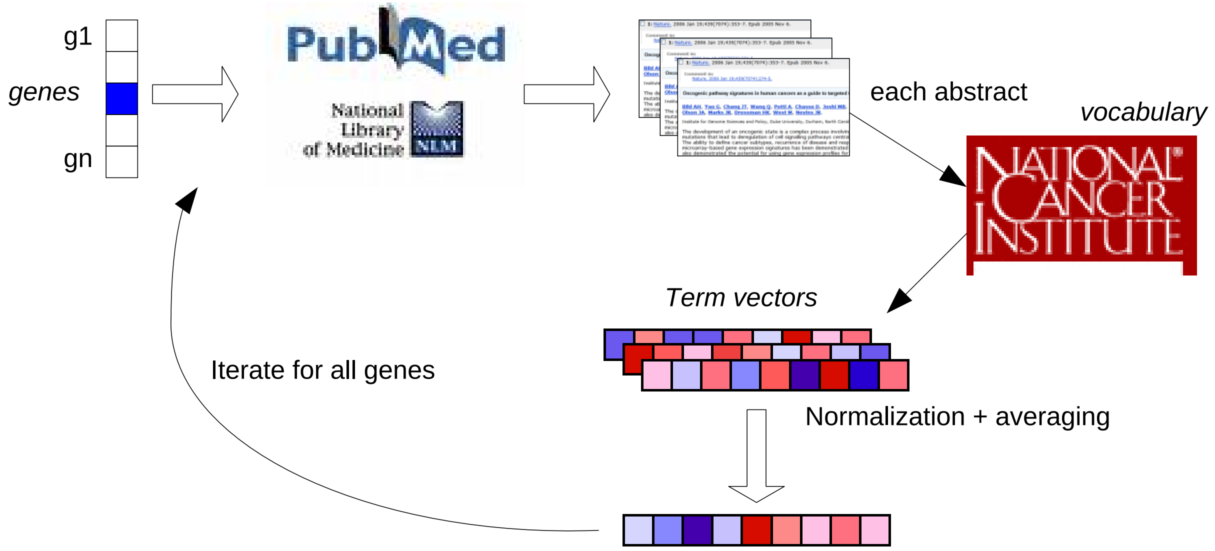
vocabulary

Term vectors

Normalization + averaging

Text mining

Structure prior construction



g1

genes

gn

each abstract

vocabulary

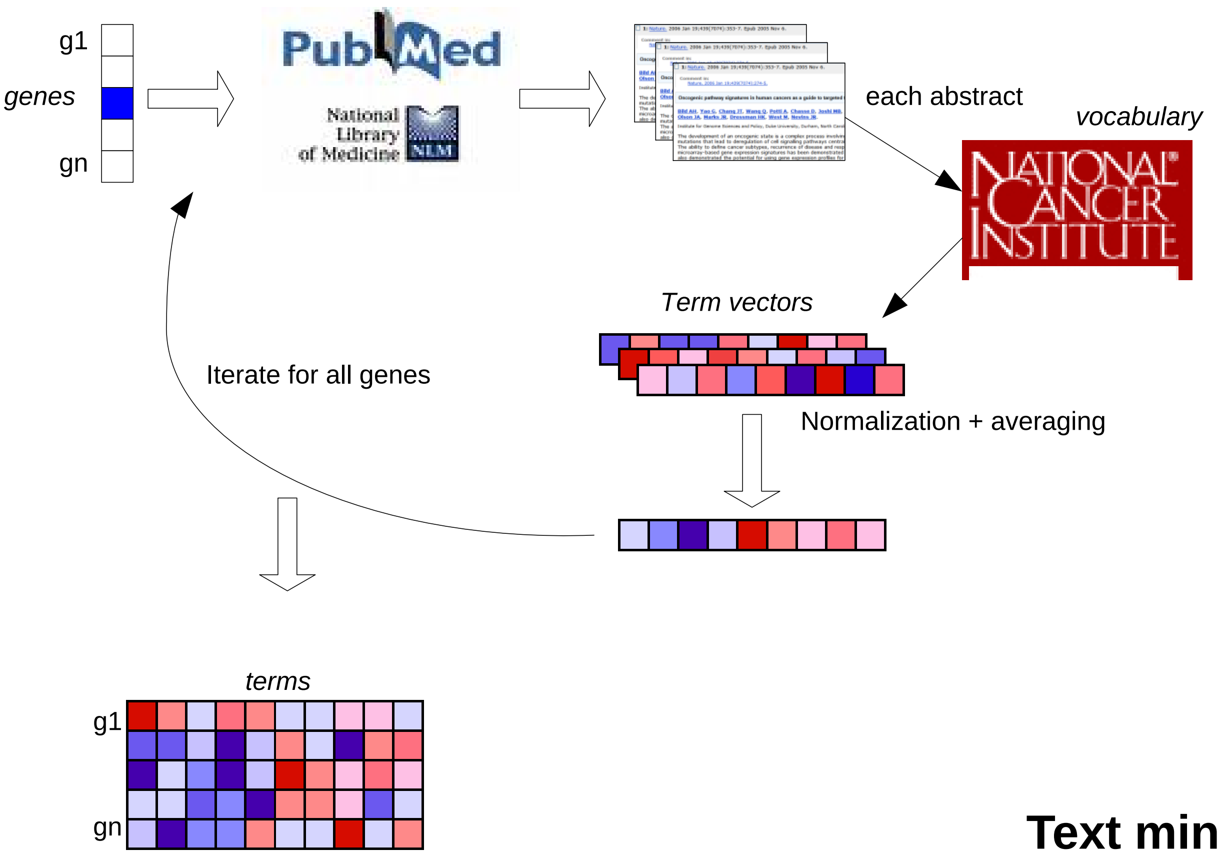
Term vectors

Normalization + averaging

Iterate for all genes

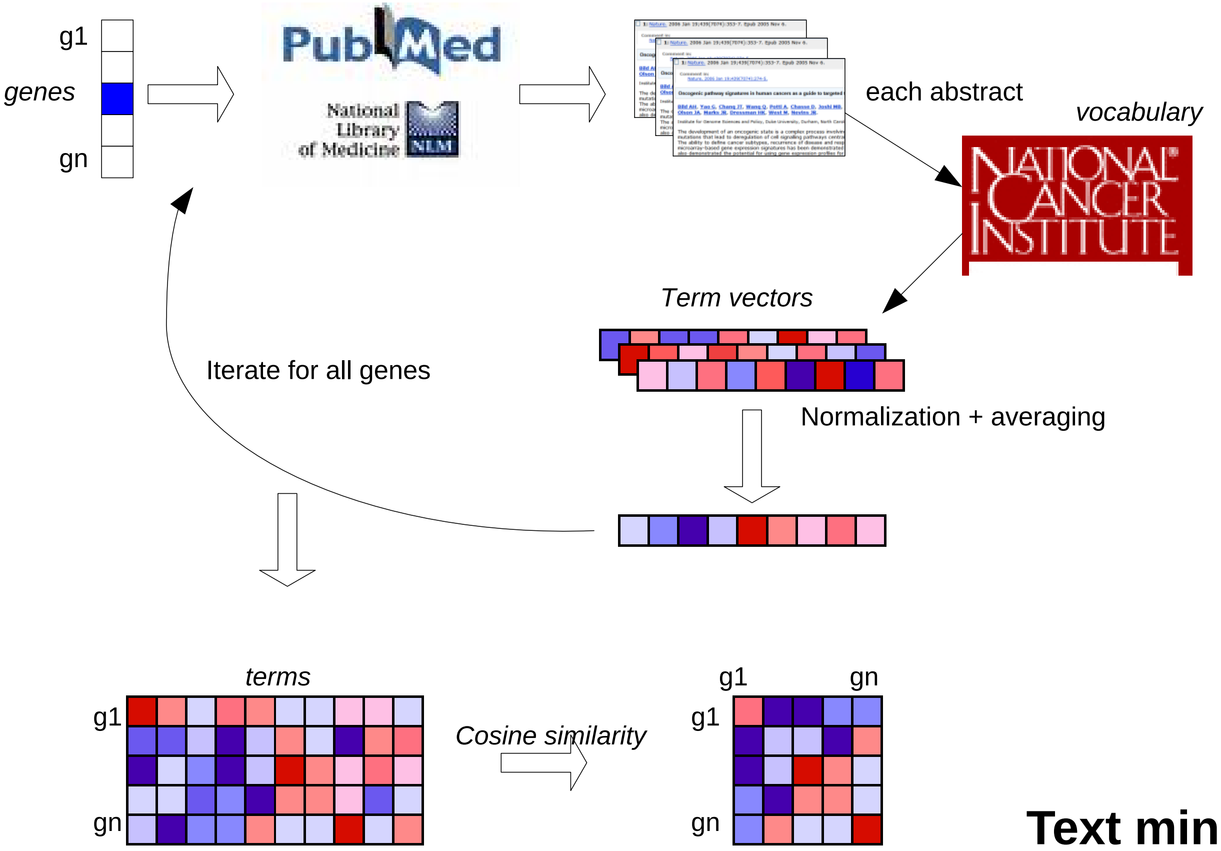
Text mining

Structure prior construction



Text mining

Structure prior construction



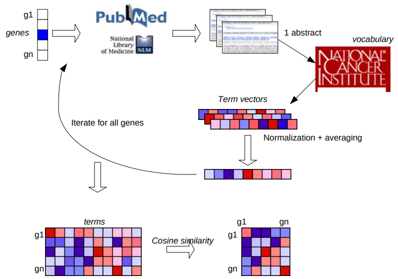
Text mining

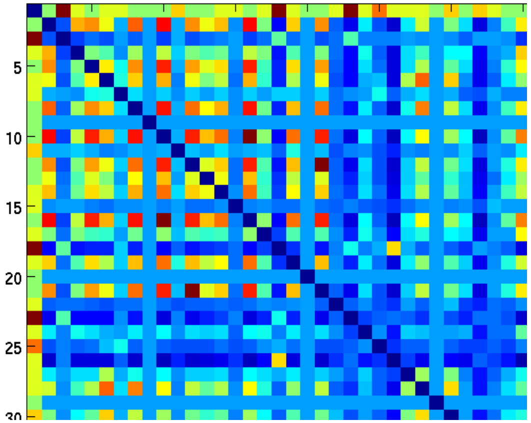
Structure prior: scaling

- 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
    - Structure prior will be scaled according to this mean density
- Low mean density ⇒ less edges ⇒ less complex networks

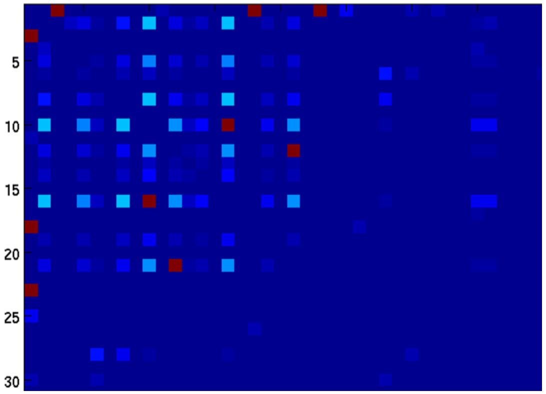
Summary

Text mining

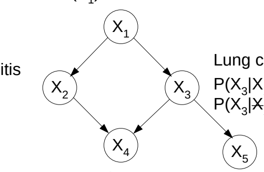




Scaling



Text prior



History of smoking
$P(X_1)= 20\%$

Chronic bronchitis
$P(X_2|X_1)= 25\%$
$P(X_2|\cancel{X}_1)= 5\%$

Lung cancer
$P(X_3|X_1)= 0.3\ \%$
$P(X_3|\cancel{X}_1)= 0.005\ \%$

Mass seen on X-ray
$P(X_5|X_3)= 60\ \%$
$P(X_5|\cancel{X}_3)= 2\ \%$

Fatigue
$P(X_4|X_2,X_3)= 75\%$
$P(X_4|X_2,\cancel{X}_3)= 10\%$
$P(X_4|\cancel{X}_2,X_3)= 50\%$
$P(X_4|\cancel{X}_2,\cancel{X}_3)= 5\%$

Results

- First case: Breast cancer (van’t Veer data)

Mean density	Text prior mean AUC	Uniform prior mean AUC	P-value
1	0.80 (0.08)	0.75(0.08)	$0.000396^{\S}$
2	0.80 (0.08)	0.75(0.07)	$<2e\text{-}06^{\S}$
3	0.79 (0.08)	0.75(0.08)	$0.00577^{\S}$
4	0.79 (0.07)	0.74(0.08)	$<6e\text{-}06^{\S}$

Average number of parents per variable

Gevaert et al. PSB 2008
Gevaert et al. Ann NY Acad Sci 2007

Results

- Second Case: Bild data (3 data sets)
  - Breast
  - Ovarian
  - Lung
- Mean density is set to 1 based on van’t Veer results

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

Gevaert et al. PSB 2008
Gevaert et al. Ann NY Acad Sci 2007

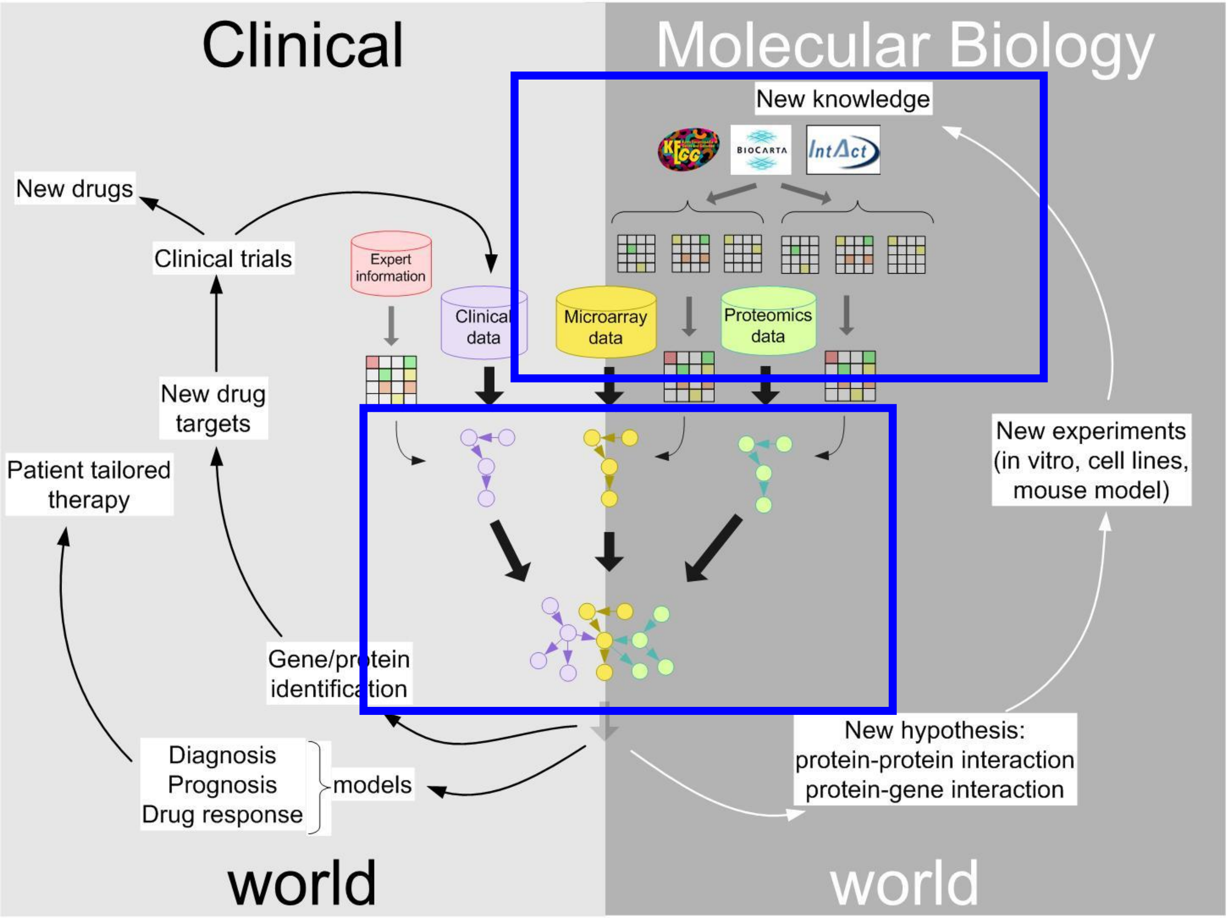
Conclusions

- 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
- Also allows to select a set of genes based on both gene expression data and knowledge available in the literature related to cancer outcome

Overall conclusions

- Our main goal was to develop a Bayesian network integration framework to model primary and secondary data
- First, we illustrated Bayesian network model on two primary data sources:
  - Clinical data
  - Genomic data
- Secondly, we illustrated the integration of primary data sources on two cases
  - Integrating clinical and microarray data of breast cancer patients
  - Integrating microarray and proteomics data of rectal cancer patients
- Thirdly, we integrated secondary data in the form of literature abstracts

Overall conclusions



Future work

- We see two important future directions
  - Integration of other secondary data sources:
    - Protein-DNA interactions (TRANSFAC), Pathway information (KEGG, Biocarta), …
    - Main issue is standardization of databases: being solved thanks to efforts such as BIOPAX
  - New technologies
    - Exon microarrays, SNP microarrays, second generation sequencing will probably unlock a whealth of information
    - Amount of data will increase super exponentially which may cause serious computational problems
    - Possible solution is parallellization: HPC cluster K.U.Leuven
      - Calculation time on VIC cluster used during PhD amounts to 1.4 years of CPU time

Acknowledgements

- ESAT-Sista
  - Prof. Bart De Moor
  - Bioinformatics group
  - Frank De Smet
  - Nathalie Pochet
  - Anneleen Daemen
  - Raf van de Plas
  - Steven van Vooren
- UZ Leuven, Gynecology
  - Prof. Dirk Timmerman
  - Prof. Ignace Vergote
  - Toon Van Gorp
  - Isabelle Cadron
  - Caroline van Holsbeke
  - Karin Leunen
- UZ Leuven, Radiation oncology
  - Prof. Karin Hautermans
- University Hospital St. Luc (Brussels)
  - Prof. Jean-Pascal Machiels