Learning in Bayesian Networks

Florian Markowetz
Max-Planck-Institute for Molecular Genetics
– Computational Molecular Biology –
Berlin

Berlin: 20.06.2002

Overview

1. **Bayesian Networks**
   - Stochastic Networks
   - Bayesian statistics
   - Use of Bayesian Networks
     - Inference given a network,
     - Learn the network from data,
     - Causal Inference

2. **Analysis of Gene Expression Data**
   - Application: Cell Cycle Expression Patterns

References

1. **Bayesian Networks**

   HECKERMAN, David: *A Tutorial on Learning with Bayesian Networks*, MSR-TR-95-06

2. **Analysis of Gene Expression Data**

   FRIEDMAN *et. al.*: *Using Bayesian Networks to Analyze Expression Data*, RECOMB 1999.

   SPELLMAN *et. al.*: *Comprehensive identification of cell cycle-regulated genes of the yeast sac. cer. by microarray hybridization*, Molecular Biology of the Cell, 9:3273–3297, 1998.

Graphical Models

1. Markov Chains
2. Hidden Markov Models
3. Markov Random Fields
4. Bayesian Networks

...

Markov Chains

Example: Evolution of a protein sequence modelled by a first-order Markov Chain.

$$
\begin{array}{ccccccccccc}
\cdots & \longrightarrow & \mathtt{A} & \longrightarrow & \mathtt{A} & \longrightarrow & \mathtt{R} & \longrightarrow & \mathtt{N} & \longrightarrow & \cdots \\
\cdots & \longrightarrow & X_{i-2} & \longrightarrow & X_{i-1} & \longrightarrow & X_i & \longrightarrow & X_{i+1} & \longrightarrow & \cdots
\end{array}
$$

Markov Chains

Example: Evolution of a protein sequence modelled by a first-order Markov Chain.

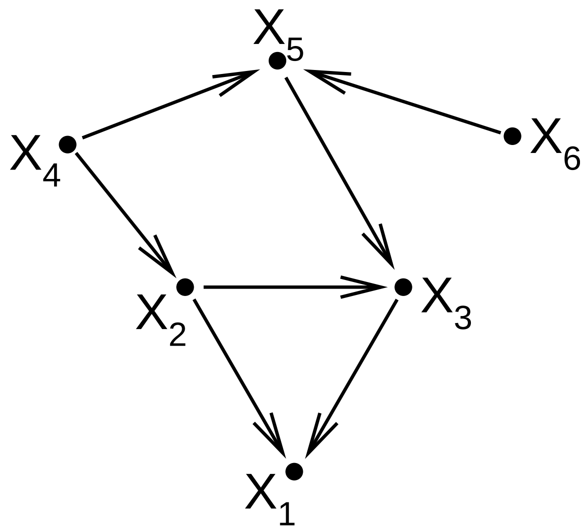
$$
\begin{array}{ccccccccccc}
\cdots & \longrightarrow & \texttt{A} & \longrightarrow & \texttt{A} & \longrightarrow & \texttt{R} & \longrightarrow & \texttt{N} & \longrightarrow & \cdots \\
\cdots & \longrightarrow & X_{i-2} & \longrightarrow & X_{i-1} & \longrightarrow & X_i & \longrightarrow & X_{i+1} & \longrightarrow & \cdots
\end{array}
$$

- **Markov Condition:**

$$P(\ X_i \mid X_{i-1}, X_{i-2}, \ldots) \ = \ P(\ X_i \mid X_{i-1})$$

- **Local probability** in X_i depends only on direct predecessor X_{i-1}.
- **Conditional Independence:** Given its predecessor, X_i is independent from the other nodes in the graph.

A more complex example



- **Markov Condition:** Each variable X_i is independent of its non-descendants given its parents.
- **Local probability** in X_i depends only on the parents.
- **Conditional Independence:** Given its parents, X_i is independent from the other nodes in the graph.

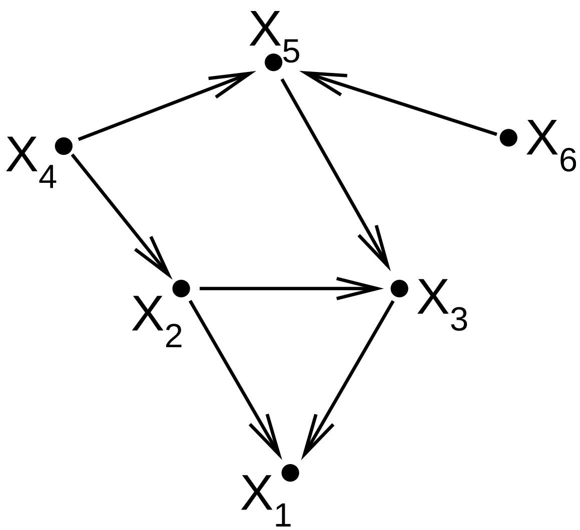
Stochastic Networks I

A **Bayesian Network** for $\mathbf{X} = \{X_1, \ldots, X_n\}$ consists of

- a **network structure** $\mathcal{S}$
  - directed acyclic graph (DAG),
  - nodes $\leftrightarrow$ variables,
  - lack of arc $\leftrightarrow$ conditional independence

- a set of **probability distributions** $\mathcal{P}$
  - locally: conditional distribution of a variable given its parents in $\mathcal{S}$:
    $\mathcal{P} = \{ P(X_i \mid pa_i) \}$

Stochastic Networks II



$(\mathcal{S}, \mathcal{P})$ encode the joint distribution

$$P(\mathbf{X}) = \prod_{i=1}^{n} P\left(X_i \mid pa_i\right)$$

Equivalence of Networks

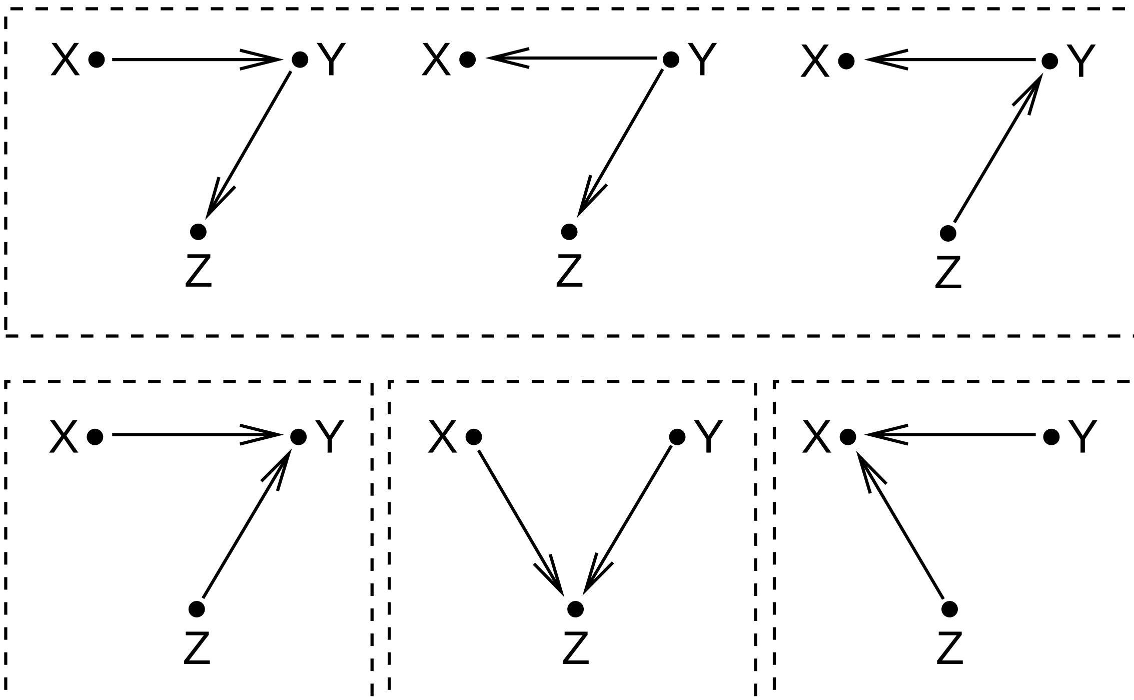
Markov equivalence

$\mathcal{S}_1 \stackrel{M}{\sim} \mathcal{S}_2$ if both structures represent the same set of independence assertions.

Equivalence of Networks

Markov equivalence

$\mathcal{S}_1 \overset{M}{\sim} \mathcal{S}_2$ if both structures represent the same set of independence assertions.



Representation of an equivalence class

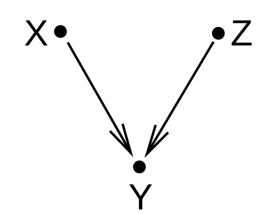
Theorem 1: [Verma and Pearl, 1990] Two DAGs are equivalent iff they have *the same skeletons* and *the same v-structures*.

Representation of an equivalence class

Theorem 1: [Verma and Pearl, 1990] Two DAGs are equivalent iff they have *the same skeletons* and *the same v-structures*.

- **skeleton:**
  the undirected graph resulting from ignoring the directionality of every edge.

- **v-structure:**
  an ordered triple of nodes (X, Y, Z) in a graph such that
  (1) $X \longrightarrow Y$ and $Z \longrightarrow Y$
  (2) X and Z are not adjacent.

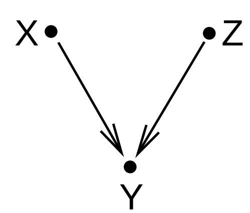


Representation of an equivalence class

Theorem 1: [Verma and Pearl, 1990] Two DAGs are equivalent iff they have *the same skeletons* and *the same v-structures*.

- **skeleton:**
  the undirected graph resulting from ignoring the directionality of every edge.

- **v-structure:**
  an ordered triple of nodes (X, Y, Z) in a graph such that
  (1) $X \longrightarrow Y$ and $Z \longrightarrow Y$
  (2) X and Z are not adjacent.



Using Theorem 1 we can uniquely represent a equivalence class of DAGs by a **partially directed graph** (PDAG).

DAG-2-PDAG

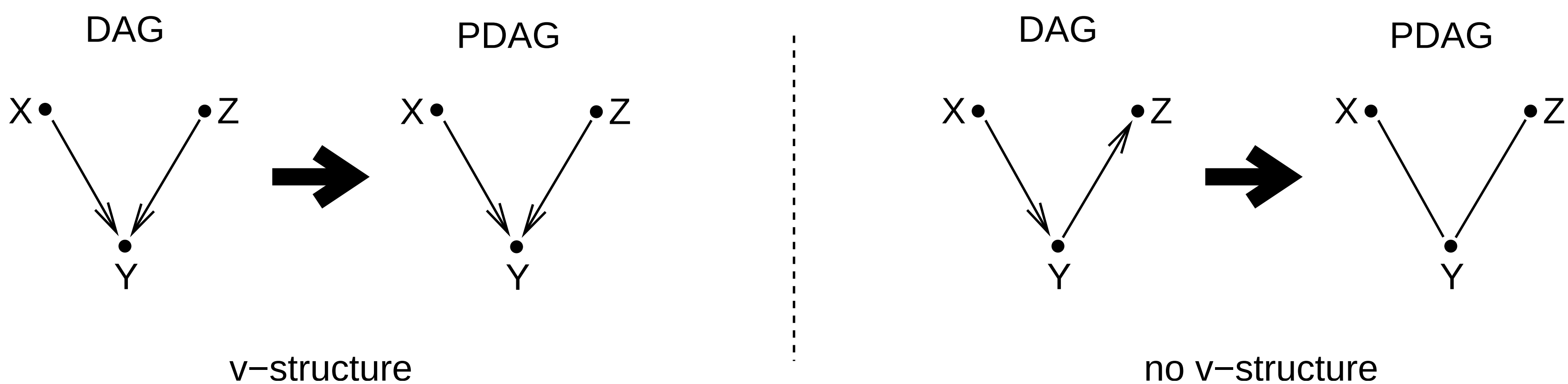
Construction:
The PDAG identifying the equivalence class of a given DAG contains

- a **directed edge** for every edge participating in a v-structure, and
- an **undirected edge** for all other edges.

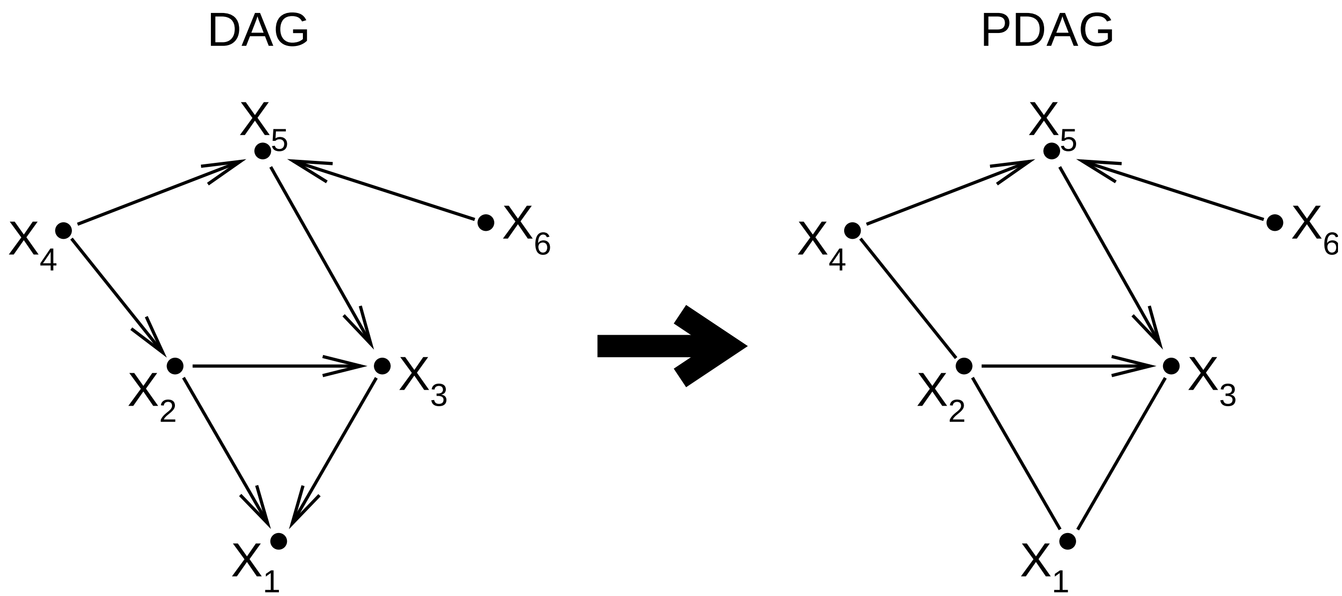
DAG-2-PDAG

Construction:
The PDAG identifying the equivalence class of a given DAG contains

- a **directed edge** for every edge participating in a v-structure, and
- an **undirected edge** for all other edges.



again: a more complex example



- $X \rightarrow Y$: all members of the class agree in this arc;
- $X - Y$: some $X \rightarrow Y$ and some $Y \rightarrow X$.

What is Probability?

Frequentist Answer:
Probability is . . .

- a physical property of the world
- measured by repeated trials

Bayesian Answer:
Probability is . . .

- a personal degree of belief
- arbitrary (techniques for sensible choice)

Bayes Formula

Bayes Formula:

$$P(\ \vartheta \mid D\ ) \ = \ \frac{P(\vartheta)\ P(D \mid \vartheta)}{P(D)}$$

Joint distribution:

$$P(\vartheta, D) \ = \ P(\vartheta)\ P(D \mid \vartheta)$$

Bayes Formula

Bayes Formula:

$$\underbrace{P(\ \vartheta \mid D\ )}_{\text{Posterior}} = \frac{\overbrace{P(\vartheta)}^{\text{Prior}}\ \overbrace{P(D \mid \vartheta)}^{\text{likelihood}}}{P(D)}$$

Joint distribution:

$$P(\vartheta, D) \ = \ P(\vartheta)\ P(D \mid \vartheta)$$

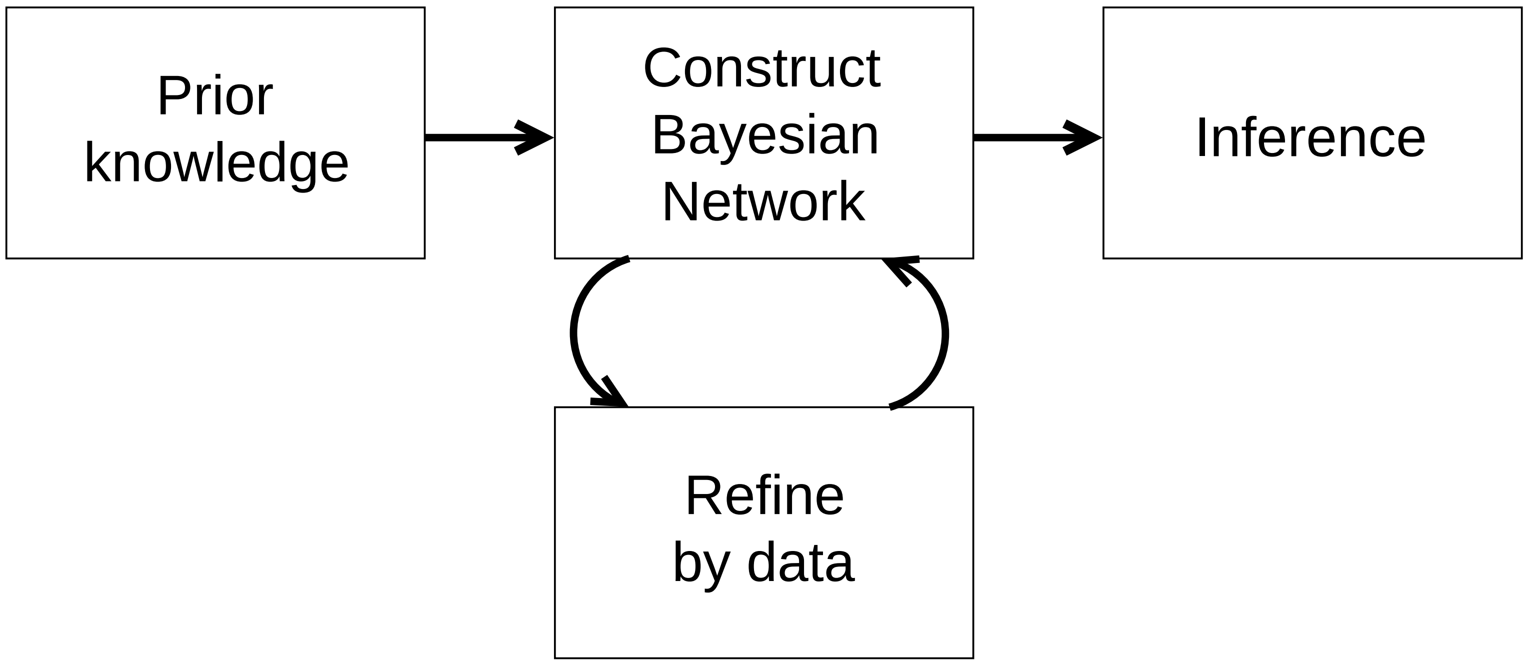
How to use a Bayesian Network?

1. **Probabilistic inference:**
   Determine probability of interest from model

2. **Learn Network:**
   Infer Structure and Probabilities from Data

3. **Causal Inference:**
   Detect causal patterns.

1. Inference in a Bayesian Network



1.1 Construction

- Determine the **variables to model**,
- **build DAG** that encodes conditional independence

  edge: cause $\longrightarrow$ effect

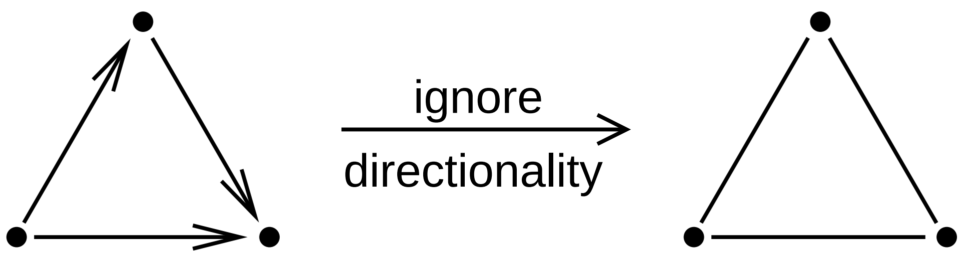
- assess **local probability distributions** $P(X_i \mid pa_i)$

1.2 Probabilistic Inference

Probabilistic Inference: compute a probability of interest given a model.

$\Longrightarrow$ Use Bayes theorem and simplify by conditional independence

Source of difficulty: undirected cycles



1.3 Refinement

Using data to update the probabilities of a given Bayesian Network structure:

$$P\left(\mathbf{X} \mid \vartheta, \mathcal{S}\right) = \prod_{i=1}^{n} P\left(X_i \mid pa_i, \vartheta_i, \mathcal{S}\right)$$

Uncertainty about local distributions is encoded in Prior $P\left(\vartheta \mid \mathcal{S}\right)$

Refinement: Compute Posterior from Prior.

$$\text{Prior} \quad P\left(\vartheta \mid \mathcal{S}\right) \quad \Longrightarrow \quad P\left(\vartheta \mid D, \mathcal{S}\right) \quad \text{Posterior}$$

2. Learning a network structure

Learning: find network structure which fits the prior knowledge and data.

- **Measure** this by a score function, e. g.

$$\text{Score}_D(\mathcal{S}) \quad = \quad \log \underbrace{p(\mathcal{S})}_{\text{Prior}} + \log \underbrace{p(D|\mathcal{S})}_{\text{likelihood}}$$

- **Search** for highscoring structure by *greedy search*, *simulated annealing*, ...

3. Learning Causal Relationships

Objective: Determine the "flow of causality" in a system.

Causal Graph: DAG $\mathcal{C}$ is a *causal graph* for $\mathbf{X} = \{X_1, \ldots, X_n\}$ iff

nodes $n_i \quad \Leftrightarrow \quad$ variables X_i,

$n_i \rightarrow n_j \quad \Leftrightarrow \quad X_i$ is **direct cause** of X_j.

Causal Networks vs. Bayesian Networks

1. Connection

Causal Markov Condition:

Given the values of a variables immediate causes,
it is independent of its earlier causes.

Thus:

$\mathcal{C}$ is a causal graph for $\mathbf{X}$ $\Longrightarrow$

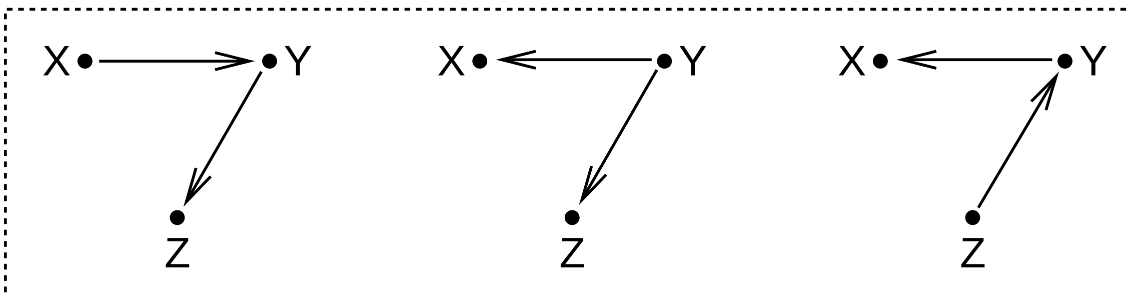
$\mathcal{C}$ is a Bayesian Network structure for the joint probability distribution of $\mathbf{X}$.

Causal Networks vs. Bayesian Networks

2. Differences

A **Bayesian Network** models the distribution of observations.

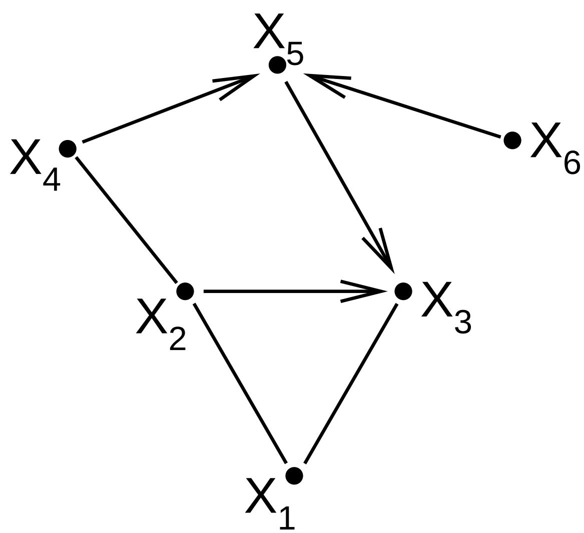
A **Causal Network** models the distribution of observations *and* effects of interventions.



Equivalent as Bayesian Networks, **but not** equivalent as Causal Networks.

Causal Inference from Data

From observations alone, we can only learn a PDAG: a whole equivalence class of structures.



$X \longrightarrow Y$ in the PDAG: all networks agree on this directed arc.
$\Longrightarrow$ Infer causal direction **X causes Y**.

Application: Cell Cycle in yeast

- Data from SPELLMAN *et. al.*: *Comprehensive identification of cell cycle-regulated genes of the yeast sac. cer. by microarray hybridization*, Molecular Biology of the Cell, 9:3273–3297, 1998.
- contains 76 samples of all the yeast genome (time series in few minutes intervals)
- Spellman *et. al.* identified 800 cell cycle regulated genes, and clustered them: 250 genes in 8 clusters
- Friedman *et. al.* analysed these 250 genes by a bayesian network.

Data Representation

Random Variables:

- Expression levels of each gene
- In addition:
  - experimental conditions
  - temporal indication (cell cycle phase)
  - background variables
  - exogenous cellular conditions

Discretization

Focus on **qualitative aspects** of the data.

Discretize gene expression data in three categories:

-1	lower than control
0	equal
1	greater

Pairwise Features

Small datasets with many variables: many different networks are reasonable explanations of the data
$\Longrightarrow$ focus on features that are common to most of these networks

1. **Markov relations:**
   Is X an direct relative of Y?
   (local property)

2. **Order relations:**
   Is X an ancestor of Y in all networks of a given equivalence class?
   (global property)

The Sparse Candidate algorithm

Learning a network structure = solving optimization problem in the space of DAGs
$\Longrightarrow$ Efficient search algorithms

Sparse Candidate algorithm (Friedman *et. al.*):

- Identify a small number of **candidate parents**,
  (simple local statistics like correlation)
- **restrict search** to networks in which only the candidate parents of a variable can be its parents,
- **iteratively adapt** candidate set during search.

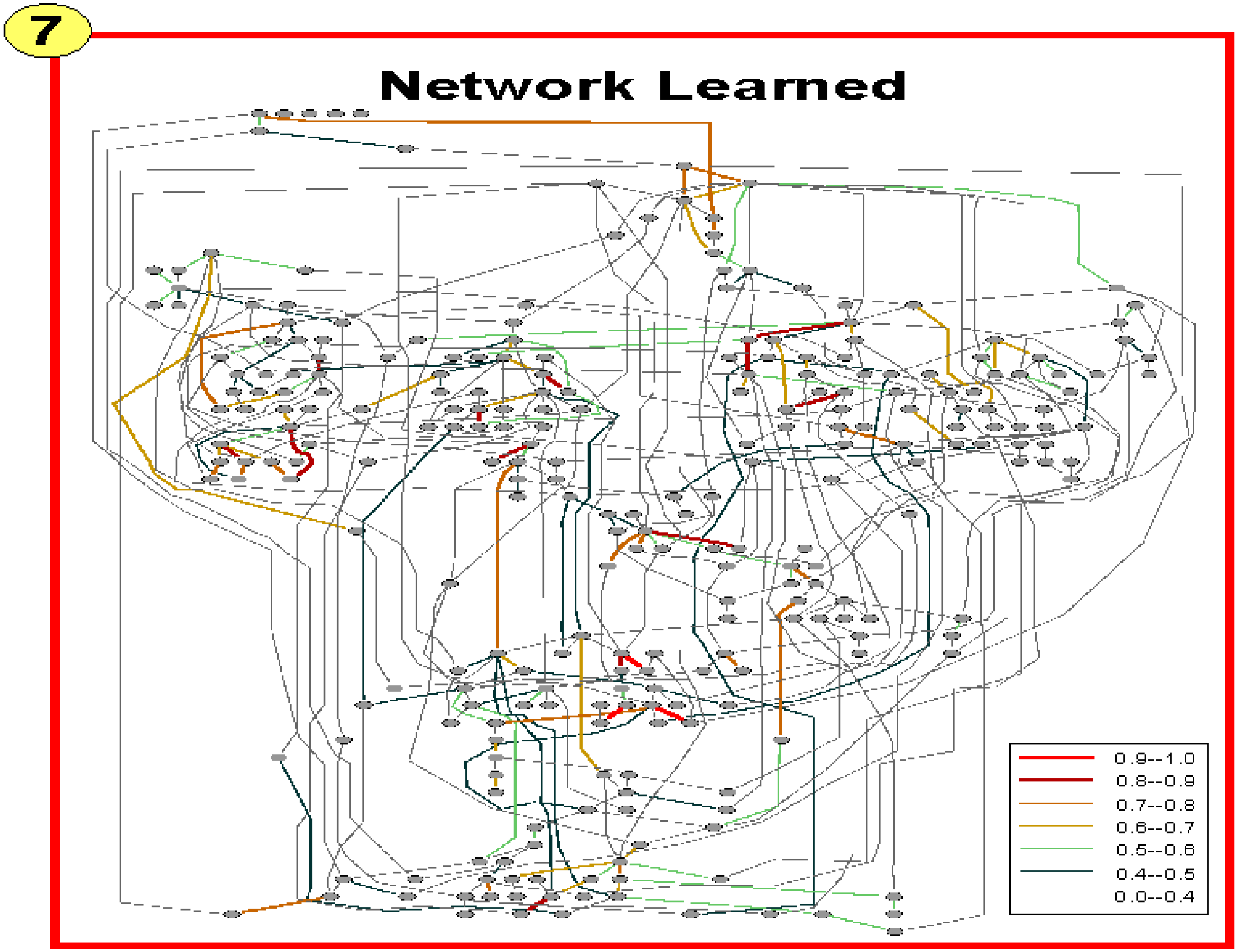
Results

- **Order relations:** only few genes dominate the order (i. e. appear before many genes). These were found to be key genes in the cell-cycle process.

- **Markov relations:** top scoring markov relations between genes were found to indicate a relation in biological function.

- **In general:** Friedman *et. al.* emphasize that BN provide us with a tool that allows biologically plausible conclusions from the data.

Don't get confused!



Ideas

- Incorporating biological knowledge
- *Interventional* instead of *observational* data
- How to learn causality from knockout experiments?
- Design of knockout experiments?