COMPUTATIONAL BIOLOGY and MEDICINE
Causality

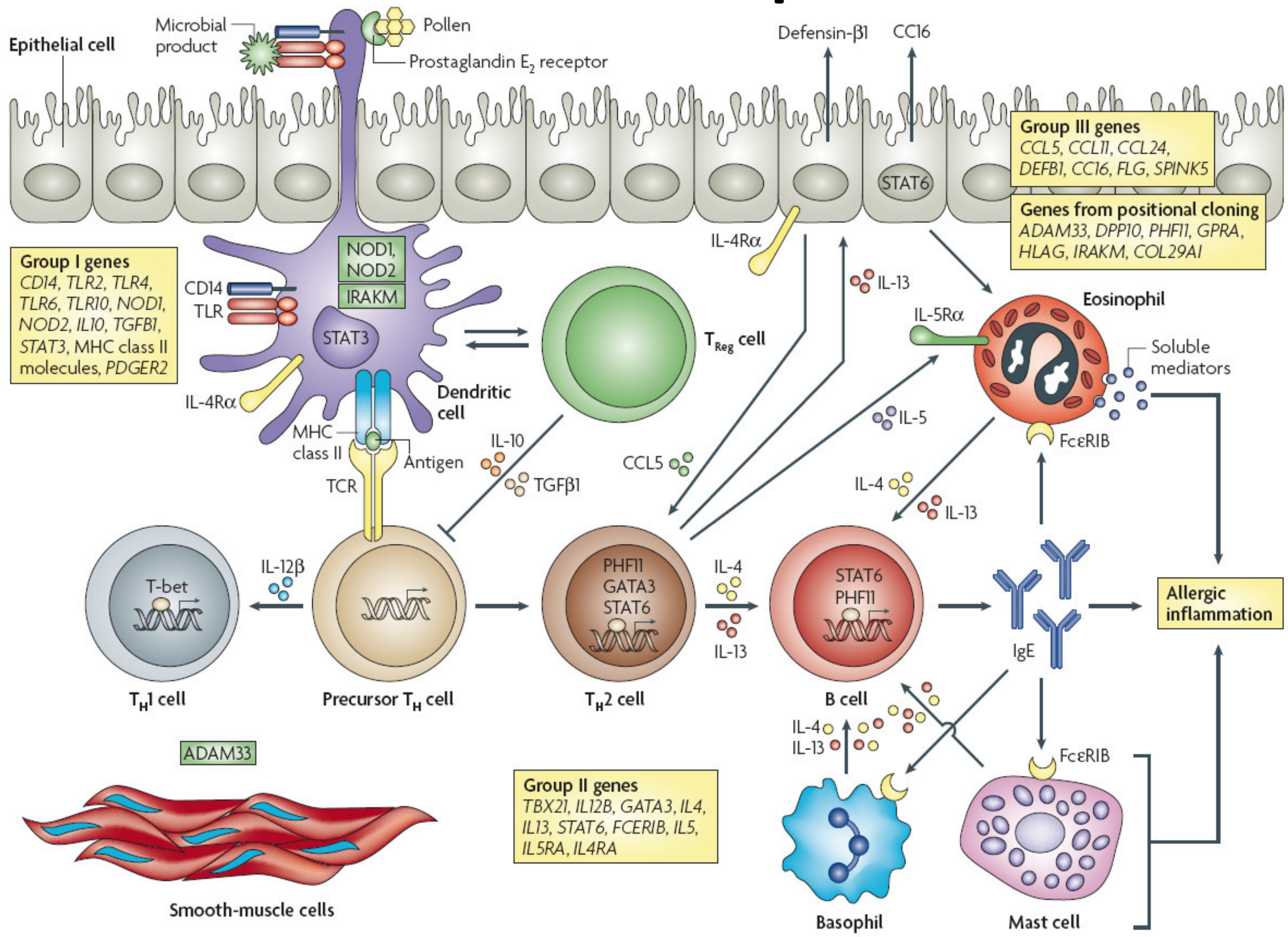
Andras Falus afalus@gmail.com
Peter Antal antal@mit.bme.hu
Gábor Csonka csonkagi@gmail.com

AIT, Budapest 2011. spring

Overview

- Statistical and causal models
  - Interpretations of probabilistic graphical models
- Observational equivalence
- Observational and interventional inference
- The Causal Markov Condition and faithfulness
- Learning causal relations

Formal models for complex diseases



„From allergy through asthma to COPD”:

Allergic pollen➔IgE➔Rhinitis➔Eosinophil➔Asthma ➔COPD

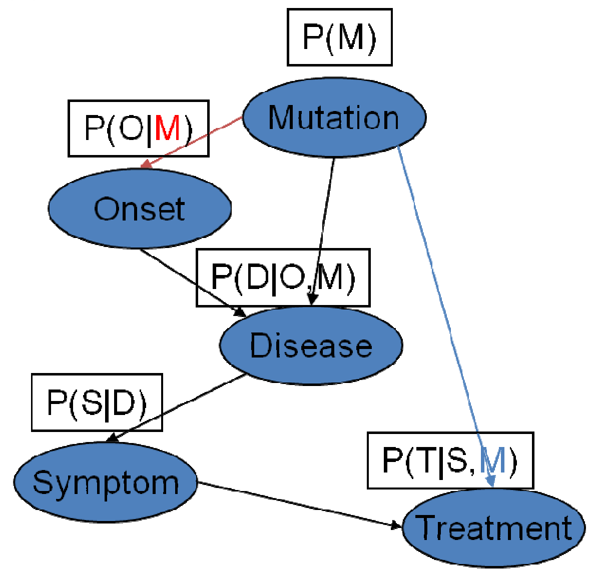
Bayesian networks

Directed acyclic graph (DAG)

- nodes – random variables/domain entities
- edges – direct probabilistic dependencies
  (edges- causal relations

Local models - $P(X_i | Pa(X_i))$

Three interpretations:



1. Causal model

3. Concise representation of joint distributions

$$P(M, O, D, S, T) =$$
$$P(M)P(O \mid M)P(D \mid O, M)P(S \mid D)P(T \mid S, M)$$

$$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\ldots\}$$

2. Graphical representation of (in)dependencies

The DAG space

The cardinality of the space of DAGs is given by the following recursion

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} 2^{i(n-1)} f(n-i) \; with \; f(0) = 1. \tag{42}$$

The number of orderings, DAGs and order-compatible DAGs with parental constraints. The columns shows respectively the number variables (nodes) (n), DAGs ($|DAG(n)|$), DAGs compatible with a given ordering ($|G_{\prec}|$), DAGs compatible with a given ordering and with maximum parental set size <=4 ($|G_{\prec}^{|\pi|\leq 4}|$) and <=2 ($|G_{\prec}^{|\pi|\leq 2}|$), the number of orderings (permutations) ($| \prec |$) and the total number of parental sets in an order-compatible DAG $|\pi^{\prec}|$ and in an order-compatible DAG with maximum parental set size <=4 ($||\pi^{\prec}| \leq 4|$) and <=2 ($||\pi^{\prec}| \leq 2|$).

n	$\lvert DAG(n)\rvert$	$\lvert G_{\prec}\rvert$	$\lvert G_{\prec}^{\lvert\pi\rvert\leq 4}\rvert$	$\lvert G_{\prec}^{\lvert\pi\rvert\leq 2}\rvert$	$\lvert \prec \rvert$	$\lvert\pi^{\prec}\rvert$	$\lvert\lvert\pi^{\prec}\rvert \leq 4\rvert$
5	2.9e+004	1e+003	1e+003	6.2e+002	1.2e+002	30	30
6	3.8e+006	3.3e+004	3.2e+004	9.9e+003	7.2e+002	62	61
7	1.1e+009	2.1e+006	1.8e+006	2.2e+005	5e+003	1.3e+002	1.2e+002
8	7.8e+011	2.7e+008	1.8e+008	6.3e+006	4e+004	2.5e+002	2.2e+002
9	1.2e+015	6.9e+010	2.9e+010	2.3e+008	3.6e+005	5.1e+002	3.8e+002
10	4.2e+018	3.5e+013	7.5e+012	1.1e+010	3.6e+006	1e+003	6.4e+002
15	2.4e+041	4.1e+031	2.1e+027	3.1e+019	1.3e+012	3.3e+004	4.9e+003
35	2.1e+213	1.3e+179	1.8e+109	8.5e+068	1e+040	3.4e+010	3.8e+005

Challenges in a complex domain

The domain is defined by the joint distribution

$$P(X_1,\ldots, X_n \mid \text{Structure},\text{parameters})$$

?

1. efficient description
   „small number of parameters”
2. representation of independencies
   „what is relevant for diagnosis”
3. representation of causal relations
   „what is the effect of a treatment”

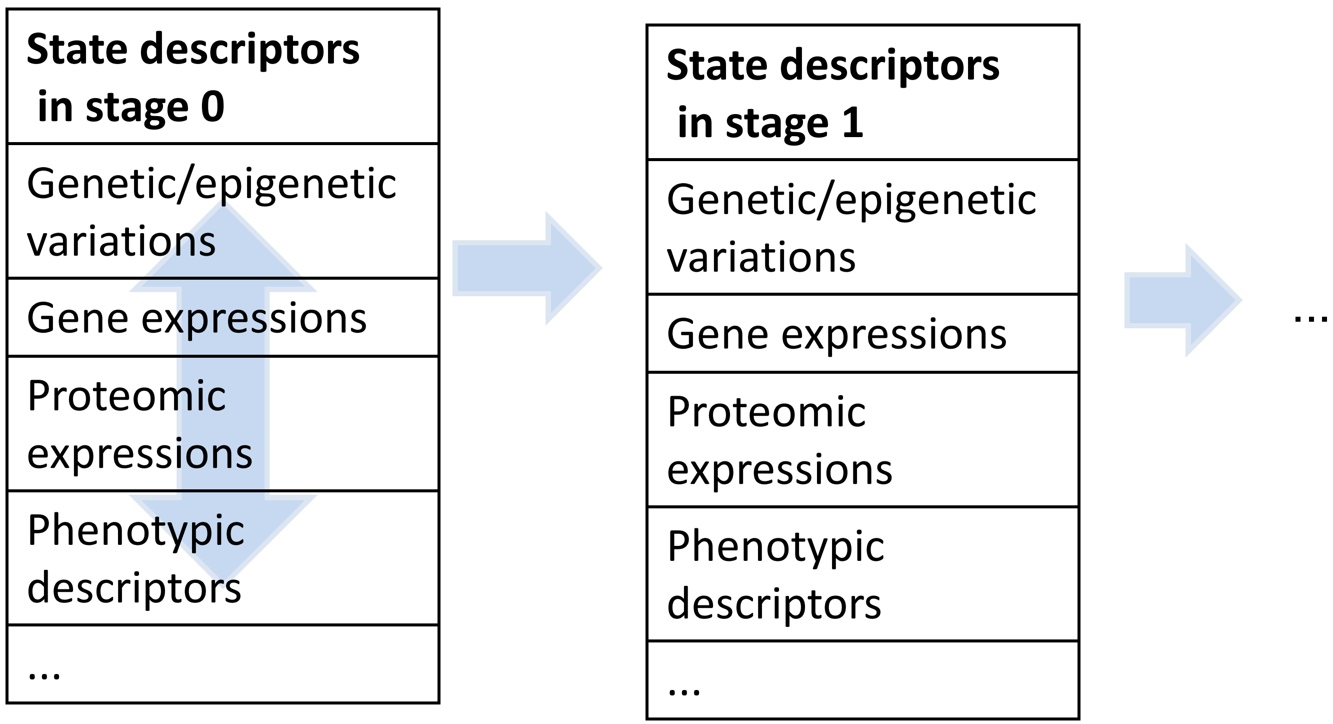
quantitave

qualitative

passive
(observational)

Active
(interventional)

A stochastic dynamical system view in biomedicine: systems biology



Stochastic internal dependencies

+ stochastic transitions

Models/knowledge representations for systems biology

- Declarative vs procedural
- Discrete vs continuous
- Deterministic vs stochastic
- Dynamic vs static
- Feedforward vs feedback
- Predictive vs domain models
- Associational vs independence vs causal

- E.g.
  - Logic
  - Boolean networks
  - Cellular automaton
  - Ordinary differential equations
  - Probabilistic models
    - Hidden Markov Models
    - (Dynamic) Bayesian networks

- Language
  - Systems Biology Markup Language

A Hidden Markov Model approach

State descriptors in stage 0
Genetic/epigenetic variations
Gene expressions
Proteomic expressions
Phenotypic descriptors

State descriptors in stage 1
Genetic/epigenetic variations
Gene expressions
Proteomic expressions
Phenotypic descriptors

……

Measurements in stage 0
Genetic/epigenetic variations
Gene expressions
Proteomic expressions
Phenotypic descriptors

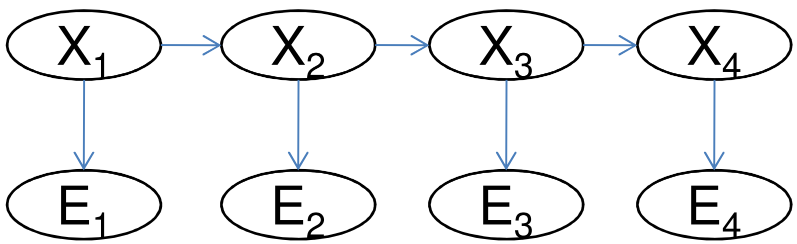
Measurements in stage 1
Genetic/epigenetic variations
Gene expressions
Proteomic expressions
Phenotypic descriptors

Stochastic internal dependencies

Stochastic measurements

Stochastic transitions

Hidden Markov Models



- First-order, homogenous Markov chain
  - Transition model: $P(X_i|X_{i-1})$
  - Sensor (or emission) model: $P(E_i|X_i)$
- Inference: P(Query|Observations)
  - Linear time complexity (w.r.t number of variables)
  - ....
- BUT WHAT TO DO in a complex state space???

Conditional independence

„Probability theory=measure theory+independence”

$I_P(X;Y|Z)$ or $(X \perp\!\!\!\perp Y|Z)_P$ denotes that X is independent of Y given Z: $P(X;Y|z)=P(Y|z)\,P(X|z)$ for all z with $P(z)>0$.

(Almost) alternatively, $I_P(X;Y|Z)$ iff

$P(X|Z,Y)= P(X|Z)$ for all z,y with $P(z,y)>0$.

Other notations: $D_P(X;Y|Z)$ =def= $\neg\, I_P(X;Y|Z)$

Contextual independence: for not all z.

The graphoid axioms

1. Symmetry: The observational probabilistic conditional independence is symmetric.

$$I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})\ iff\ I_p(\boldsymbol{Y};\boldsymbol{X}|\boldsymbol{Z})$$

2. Decomposition: Any part of an irrelevant information is irrelevant.

$$I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z})\Rightarrow\ I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})\ and\ I_p(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{Z})$$

3. Weak union: Irrelevant information remains irrelevant after learning (other) irrelevant information.

$$I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z})\Rightarrow\ I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}\cup\boldsymbol{W})$$

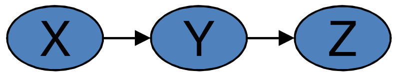
4. Contraction: Irrelevant information remains irrelevant after forgetting (other) irrelevant information.

$$I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})\ and\ I_p(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{Z}\cup\boldsymbol{Y})\Rightarrow\ I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z})$$

The independence model of a distribution

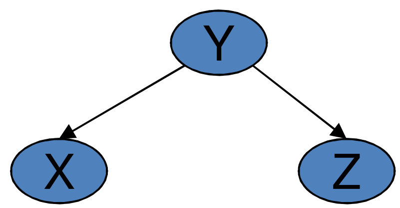
The independence map (model) M of a distribution P is the set of the valid independence triplets:

$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\ldots, I_{P,K}(X_K;Y_K|Z_K)\}$



If P(X,Y,Z) is a Markov chain, then
$M_P=\{D(X;Y), D(Y;Z), I(X;Z|Y)\}$
Normally/almost always: $D(X;Z)$
Exceptionally: $I(X;Z)$

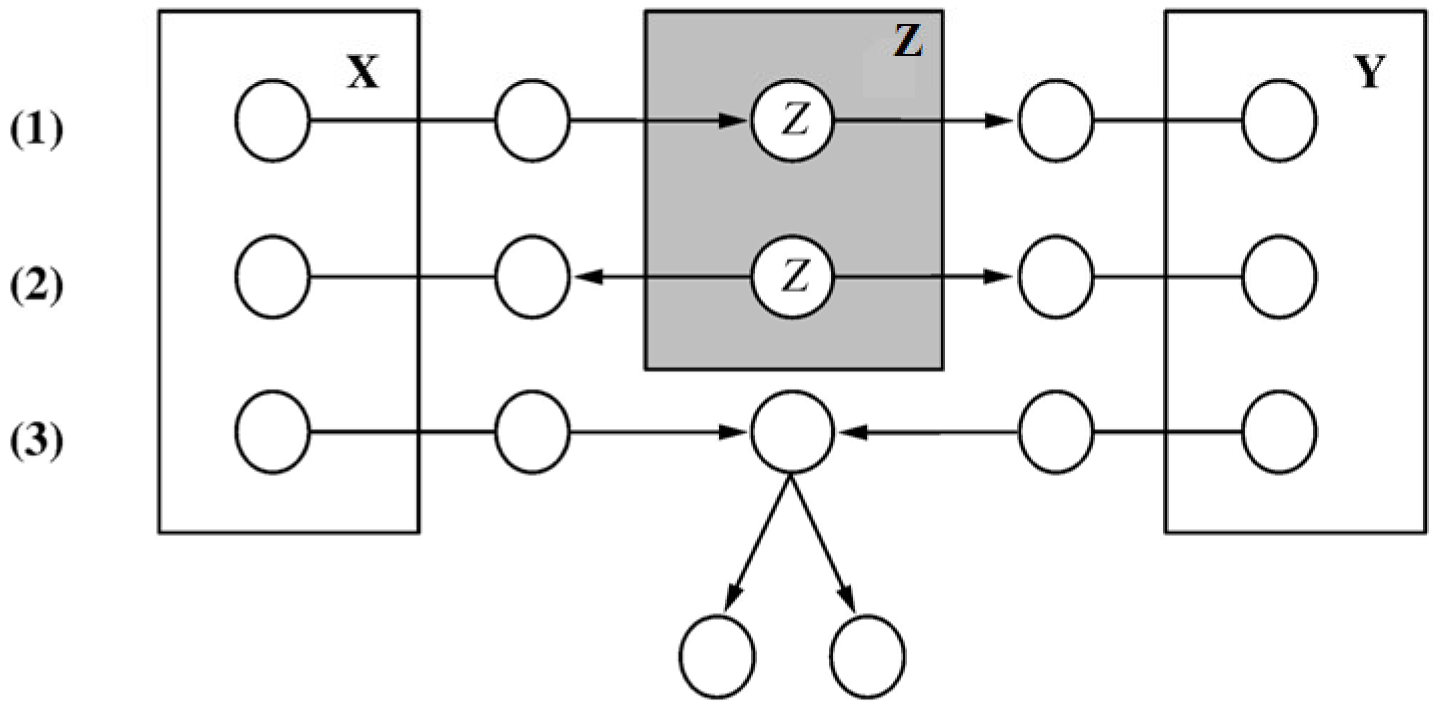
The independence map of a N-BN



If P(Y,X,Z) is a naive Bayesian network, then
M_P={D(X;Y), D(Y;Z), I(X;Z|Y)}
Normally/almost always: D(X;Z)
Exceptionally: I(X;Z)

D-separation

$I_G(X;Y|Z)$ denotes that X is d-separated (directed separated) from Y by Z in directed graph G.



D-separation and the global Markov condition

Definition 7 *A distribution $P(X_1, \ldots, X_n)$ obeys the global Markov condition w.r.t. DAG G, if*

$$\forall\, X, Y, Z \subseteq U \; (X \perp\!\!\!\perp Y|Z)_G \Rightarrow (X \perp\!\!\!\perp Y|Z)_P, \tag{9}$$

where $(X \perp\!\!\!\perp Y|Z)_G$ denotes that X and Y are d-separated by Z, that is if every path p between a node in X and a node in Y is blocked by Z as follows

1. *either path p contains a node n in Z with non-converging arrows (i.e. $\rightarrow n \rightarrow$ or $\leftarrow n \rightarrow$),*
2. *or path p contains a node n not in Z with converging arrows (i.e. $\rightarrow n \leftarrow$) and none of its descendants of n is in Z.*

Representation of independencies

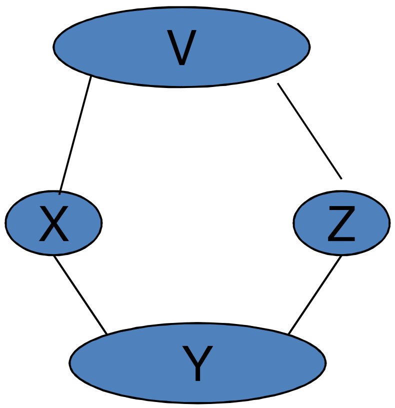
D-separation provides a sound and complete, computationally efficient algorithm to read off an (in)dependency model consisting the independencies that are valid in all distributions Markov relative to G, that is $\forall\ X, Y, Z \subseteq V$

$$(X \perp\!\!\!\perp Y|Z)_G \Leftrightarrow ((X \perp\!\!\!\perp Y|Z)_P \text{ in all P Markov relative to G}). \qquad (10)$$

For certain distributions exact representation is not possible by Bayesian networks, e.g.:

1. Intransitive Markov chain: X➔Y➔Z
2. Pure multivariate cause: {X,Z}➔Y
3. Diamond structure:

P(X,Y,Z,V) with M_P={D(X;Z), D(X;Y), D(V;X), D(V;Z), I(V;Y|{X,Z}), I(X;Z|{V,Y}).. }.



Markov conditions

Definition 4 *A distribution $P(X_1, \ldots, X_n)$ is* Markov relative *to DAG G or factorizes w.r.t G, if*

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i)), \tag{6}$$

where $Pa(X_i)$ denotes the parents of X_i in G.

Definition 5 *A distribution $P(X_1, \ldots, X_n)$ obeys the* ordered Markov *condition w.r.t. DAG G, if*

$$\forall\, i = 1, \ldots, n : (X_{\pi(i)} \perp\!\!\!\perp \{X_{\pi(1)}, \ldots X_{\pi(i-1)}\} / Pa(X_{\pi(i)}) | Pa(X_{\pi(i)}))_P, \tag{7}$$

where $\pi()$ is some ancestral ordering w.r.t. G (i.e. compatible with arrows in G).

Definition 6 *A distribution $P(X_1, \ldots, X_n)$ obeys the* local (or parental) Markov *condition w.r.t. DAG G, if*

$$\forall\, i = 1, \ldots, n : (X_i \perp\!\!\!\perp \mathrm{Nondescendants(X_i)} | \mathrm{Pa(X_i)})_{\mathrm{P}}, \tag{8}$$

where $\mathrm{Nondescendants(X_i)}$ denotes the nondescendants of X_i in G.

Bayesian network definitions

Theorem 1 *Let $P(U)$ a probability distribution and* **G** *a DAG, then the conditions above (repeated below) are equivalent:*

F P *is Markov relative* G *or* P *factorizes w.r.t* G,

O P *obeys the ordered Markov condition w.r.t.* G,

L P *obeys the local Markov condition w.r.t.* G,

G P *obeys the global Markov condition w.r.t.* G.

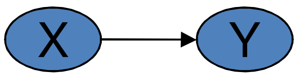
Definition 8 *A directed acyclic graph* **(DAG)** G *is a* Bayesian network *of distribution* $P(U)$ *iff the variables are represented with nodes in* G *and* (G, P) *satisfies any of the conditions* F, O, L, G *such that* G *is minimal (i.e. no edge(s) can be omitted without violating a condition* F, O, L, G*).*

A practical definition

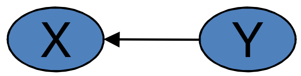
Definition 9 *A Bayesian network model M of domain with variables U consists of a structure G and parameters $\boldsymbol{\theta}$. The structure G is a* DAG *such that each node represents a variable and local probabilistic models $p(X_i|pa(X_i))$ are attached to each node w.r.t. the structure G, that is they describe the stochastic dependency of variable X_i on its parents $pa(X_i)$. As the conditionals are frequently from a certain parametric family, the conditional for X_i is parameterized by $\boldsymbol{\theta}_i$, and $\boldsymbol{\theta}$ denotes the overall parameterezation of the model.*

Association vs. Causation

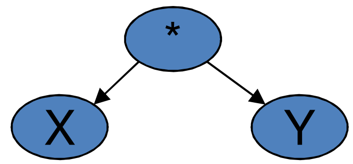
Causal models:



X causes Y

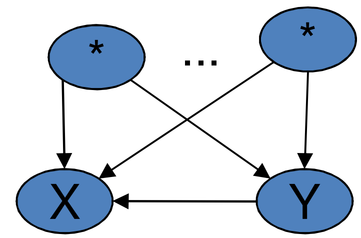


Y causes X



There is a common cause (pure confounding)

...

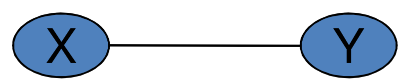


Causal effect of Y on X is confounded by many factors

From passive observations:

P(X,Y)

$M_P=\{D(X;Y)\}$



„X and Y are associated"

Reichenbach's Common Cause Principle:

a correlation between events *X* and *Y* indicates either that *X* causes *Y*, or that *Y* causes *X*, or that *X* and *Y* have a common cause.

Association vs. Causation: Markov chain

Causal models:

$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$ $\qquad$ $X_1 \leftarrow X_2 \leftarrow X_3 \leftarrow X_4$

Markov chain

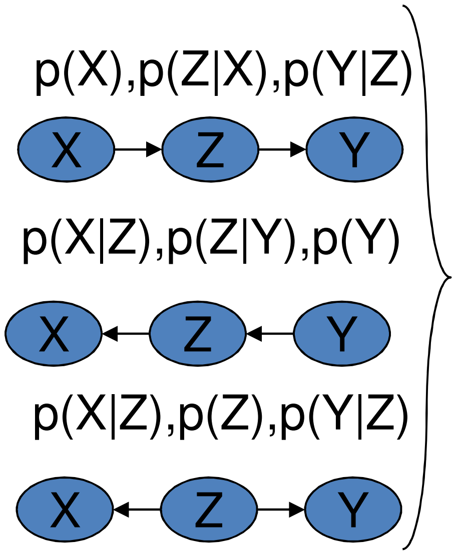
$P(X_1,\ldots)$

$M_P=\{I(X_{i+1};X_{i-1}|X_i)\}$

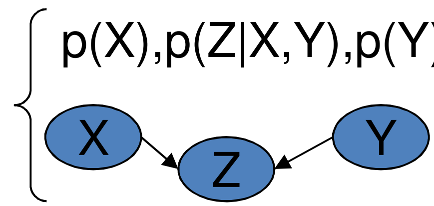
„first order Markov property”

Flow of time?

The building block of causality: v-structure



“transitive” M ≠ „intransitive” M



„v-structure”

M_P={D(X;Z), D(Z;Y), D(X,Y), I(X;Y|Z)}

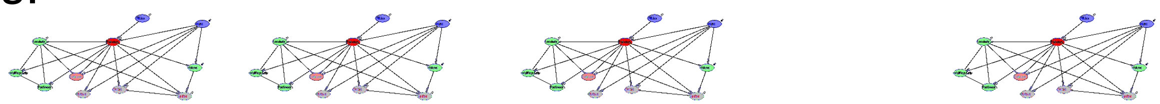
M_P={D(X;Z), D(Y;Z), I(X;Y), D(X;Y|Z) }

Often: present knowledge renders future states conditionally independent.
(confounding)

Ever(?): present knowledge renders past states conditionally independent.
(backward/atemporal confounding)

Observational equivalence of causal models

Causal models:



J.Pearl:
~„3D objects”

From passive observations:

$P(X_1,\dots, X_n)$

$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\dots, I_{P,K}(X_K;Y_K|Z_K)\}$

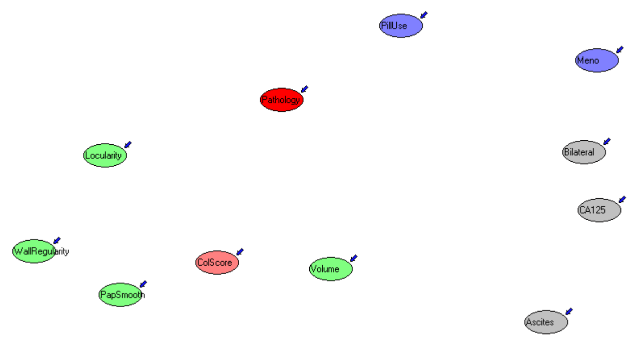
„2D projection”

Different causal models can have the same independence map!

Typically causal models cannot be identified from passive observations, they are observationally equivalent.

Observational equivalence: total independence

„Causal” model:



One-to-one relation

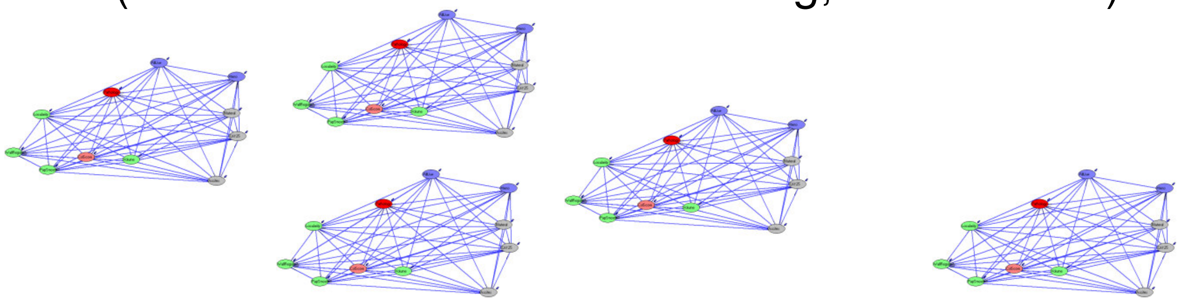
Dependency map:

$P(X_1, \dots, X_n)$

$M_P = \{I_{P,1}(X_1; X_2), \dots\}$

Observational equivalence: full dependence

„Causal" models (there is a DAG for each ordering, i.e. n! DAGs):



One-to-many relation

Dependency map:

$P(X_1,\ldots, X_n)$

$M_P=\{D_{P,1}(X_1;X_2),\ldots\}$

Observational equivalence of causal models

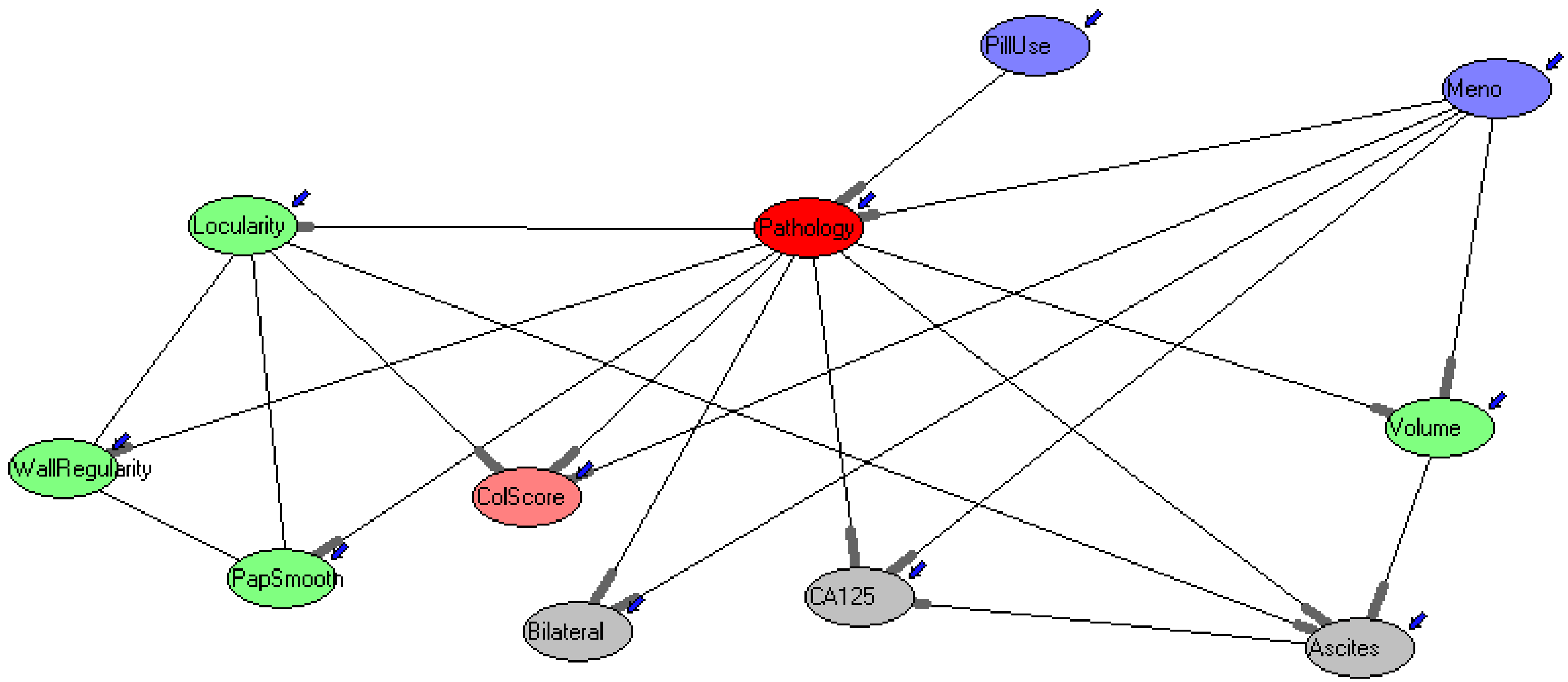
Definition 11 *Two DAGs G_1, G_2 are* observationally equivalent *, if they imply the same set of independence relations (i.e. $(X \perp\!\!\!\perp Y|Z)_{G_1}) \Leftrightarrow (X \perp\!\!\!\perp Y|Z)_{G_2}$).*

The implied equivalence classes may contain $n!$ number of DAGs (e.g. all the full networks representing no independencies) or just 1.

Theorem 2 *Two DAGs G_1, G_2 are* observationally equivalent *, iff they have the same skeleton (i.e. the same edges without directions) and the same set of* v-structures *(i.e. two converging arrows without an arrow between their tails).*

Definition 12 *The* essential graph *representing observationally equivalent DAGs is a partially oriented DAG (PDAG), that represents the identically oriented edges called* compelled edges *of the observationally equivalent DAGs (i.e. in the equivalence class), such a way that in the common skeleton only the compelled edges are directed (the others are undirected representing inconclusiveness).*

Compelled edges and PDAG



The Causal Markov Condition

- A DAG is called a *causal structure* over a set of variables, if each node represents a variable and edges direct influences. A *causal model* is a causal structure extended with local probabilistic models.
- A causal structure *G* and distribution *P* satisfies the Causal Markov Condition, if P obeys the local Markov condition w.r.t. G.
- The distribution P is said to stable (or faithful), if there exists a DAG called *perfect map* exactly representing its (in)dependencies (i.e. $I_G(X;Y|Z) \Leftrightarrow I_P(X;Y|Z)\ \forall\ X,Y,Z \subseteq V$).

- CMC: sufficiency of G (there is no extra, acausal edge)
- Faithfulness/stability: necessity of G (there are no extra, parametric independency)

Inference in Bayesian networks

- **(Passive, observational) inference**
  - **P(Query|Observations)**
- Interventionist inference
  - P(Query|Observations, Interventions)
- Counterfactual inference
  - P(Query| Observations, Counterfactual conditionals)

Inference by enumeration

If the joint distribution is efficiently represented by a Bayesian network, then any conditional is exactly defined:

$$P(Q=q|E=e),$$

where Q is the query variable, E are the evidence variables.

By definition

$\mathbf{P}(\mathbf{Q=q} \mid \mathbf{E} = \mathbf{e})$

$= \mathbf{P}(\mathbf{Q=q},\mathbf{E} = \mathbf{e}) / \mathbf{P}(\mathbf{E} = \mathbf{e})$

$= \Sigma_h \mathbf{P}(\mathbf{Q=q},\mathbf{E}= \mathbf{e}, \mathbf{H} = \mathbf{h}) / \Sigma_{h,q} \mathbf{P}(\mathbf{Q=q},\mathbf{E}= \mathbf{e}, \mathbf{H} = \mathbf{h})$

where $\mathbf{H} = \mathbf{X} - \mathbf{Q} - \mathbf{E}$ are the hidden variables, and $\mathbf{P}(\mathbf{Q=q},\mathbf{E}= \mathbf{e}, \mathbf{H} = \mathbf{h}) = \prod_i P(X_i,|Pa(X_i))$.

Problem:

Worst-case time complexity $O(d^n)$ where d is the largest arity

Complexity of exact inference

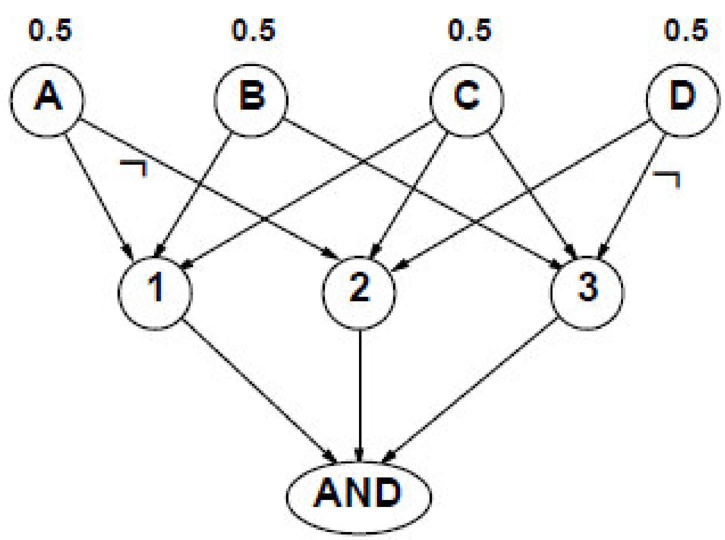
Singly connected networks (or polytrees):

- any two nodes are connected by at most one (undirected) path
- time and space cost of variable elimination are $O(d^k n)$

Multiply connected networks:

- can reduce 3SAT to exact inference $\Rightarrow$ NP-hard
- equivalent to counting 3SAT models $\Rightarrow$ #P-complete

1. A ∨ B ∨ C
2. C ∨ D ∨ ¬A
3. B ∨ C ∨ ¬D



Russel&Norvig: Artificial intelligence, ch.14

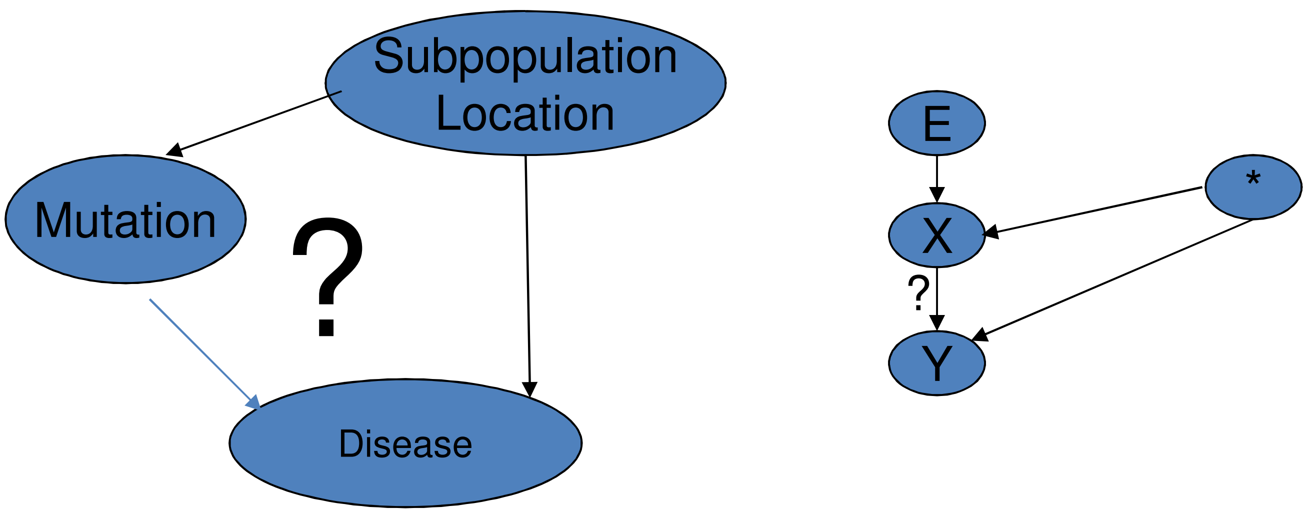
Inference in Bayesian networks

- (Passive, observational) inference
  - P(Query|Observations)
- **Interventionist inference**
  - **P(Query|Observations, Interventions)**
- Counterfactual inference
  - P(Query| Observations, Counterfactual conditionals)

Interventions and graph surgery

If G is a causal model, then compute p(Y|do(X=x)) by

1. deleting the incoming edges to X
2. setting X=x
3. performing standard Bayesian network inference.



Statistical vs causal inference

- Statistical concepts:
  - „correlation, regression, dependence, conditional independence, likelihood, collapsibility, propensity score, risk ratio, odds ratio, marginalization, conditionalization, “controlling for,”..
  - Any relation based on the joint distribution of observations.
- Causal concepts:
  - randomization, influence, effect, confounding, “holding constant,” disturbance, spurious correlation, faithfulness/stability, instrumental variables, intervention, explanation, attribution
  - Causal inference: statistical inference + causal assumptions

J.Pearl: Causality

A deterministic concept of causation

- H.Simon
  - $X_i = f_i(X_1,..,X_{i-1})$ for i=1..n
  - In the linear case the sytem of equations indicates a natural causal ordering

					X
				X	X
			X	X	X
		X	X	X	X
	....				

In fact the probabilistic conceptualization is its generalization:

$$P(X_i, | X_1,..,X_{i-1}) \sim X_i = f_i(X_1,..,X_{i-1})$$

The Inductive Causation algorithm

Assuming a stable distribution P (Pearl,2000):

1. *Skeleton:* Construct an undirected graph (skeleton), such that variables $X, Y \in \boldsymbol{V}$ are connected with an edge iff $\forall S (X \perp\!\!\!\perp Y | S)_P$, where $S \subseteq \boldsymbol{V} \setminus \{X, Y\}$.
2. *v-structures:* Orient $X \rightarrow Z \leftarrow Y$ iff X, Y are nonadjacent, Z is a common neighbour and $\neg \exists S$ that $(X \perp\!\!\!\perp Y | S)_P$, where $S \subseteq \boldsymbol{V} \setminus \{X, Y\}$ and $Z \in S$.
3. *propagation:* Orient undirected edges without creating new v-structures and directed cycle.

8. Theorem. *The following four rules are necessary and sufficient.*

R_1 *if* $(a \not- c) \wedge (a \rightarrow b) \wedge (b - c)$, *then* $b \rightarrow c$

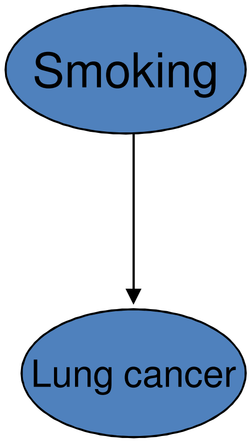
R_2 *if* $(a \rightarrow c \rightarrow b) \wedge (a - b)$, *then* $a \rightarrow b$

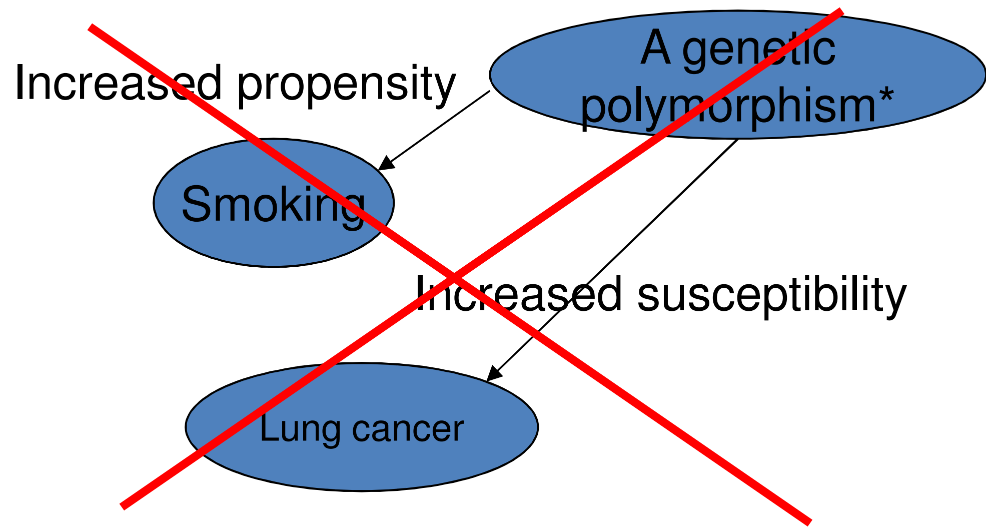
R_3 *if* $(a - b) \wedge (a - c \rightarrow b) \wedge (a - d \rightarrow b) \wedge (c \not- d)$, *then* $a \rightarrow b$

R_4 *if* $(a - b) \wedge (a - c \rightarrow d) \wedge (c \rightarrow d \rightarrow b) \wedge (c \not- b) \wedge (a - d)$, *then* $a \rightarrow b$

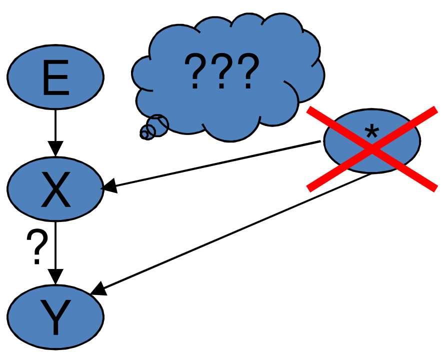
Local Causal Discovery

- Can we learn causal relations from observational data in presence of confounders???





- Automated, tabula rasa causal inference from (passive) observation is possible, i.e. hidden, confounding variables can be excluded



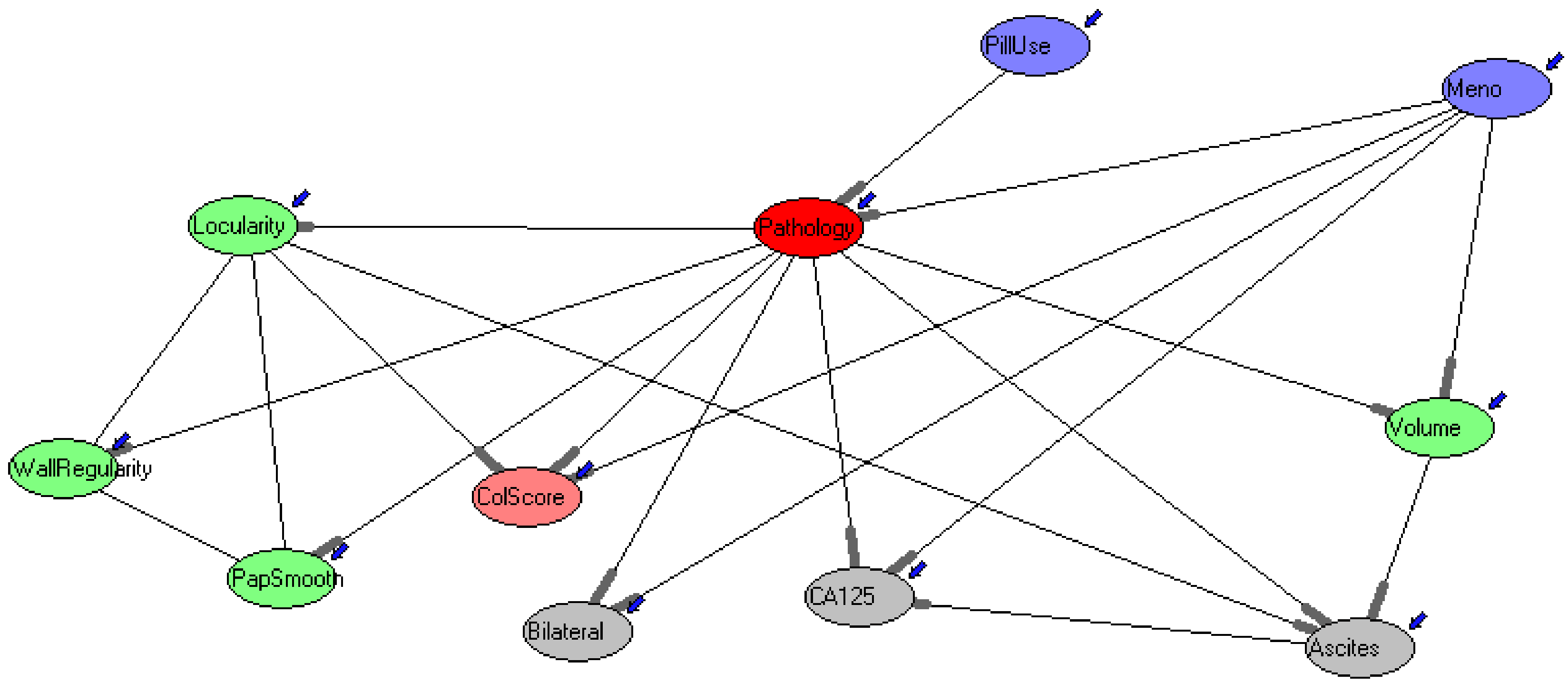
„Plato's two surprises:

1. Not all true theorems can be proved
2. Causal inference is possible from observations"

Statistical time

- Newtonian physics is symmetric.
- Macroscopic is not: entropy/thermodynamic.
- Quantum mechanics is not: state collapse
  - (R.Penrose: The emperor’s new mind)
- Subjective experience(?):
  - R.Penrose, S. Hawking
- J.Pearl: statistical time is compatible with a minimal causal model (with compelled edges).

Statistical time: example



A „causal”(?) chain: MenoPausalState➔Volume➔Ascites➔CA125

Summary

- Statistical and causal models
  - Interpretations of probabilistic graphical models
- Observational equivalence
- Observational and interventional inference
- The Causal Markov Condition and faithfulness
- Learning causal relations

- Homework: construct a model for a disease

  http://redmine.genagrid.eu/

  Login: bayeseyestudent Files: Wiki