AIT-BUDAPEST

AQUINCUM INSTITUTE OF TECHNOLOGY

Creativity in
Computer Science &
Engineering

COMPUTATIONAL BIOLOGY and MEDICINE
Biomarker discovery I.

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

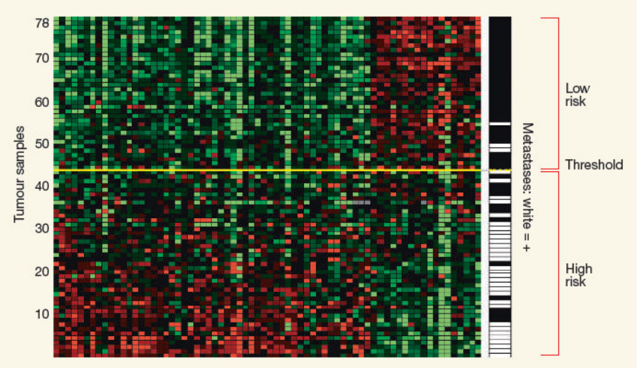
AIT, Budapest 2011. fall

Overview

- Biomarkers
- Multivariate approach to GAS
- Models: Naïve Bayesian network, logistic regression
- Feature relevance, the feature subset selection problem
- Sufficient and necessary set for diagnosis:
  - Markov blankets
  - Strong relevance
- Identification methods of biomarkers
- The Bayesian statistical approach
- Partial multivariate analysis
- Genagrid
- BayesEye

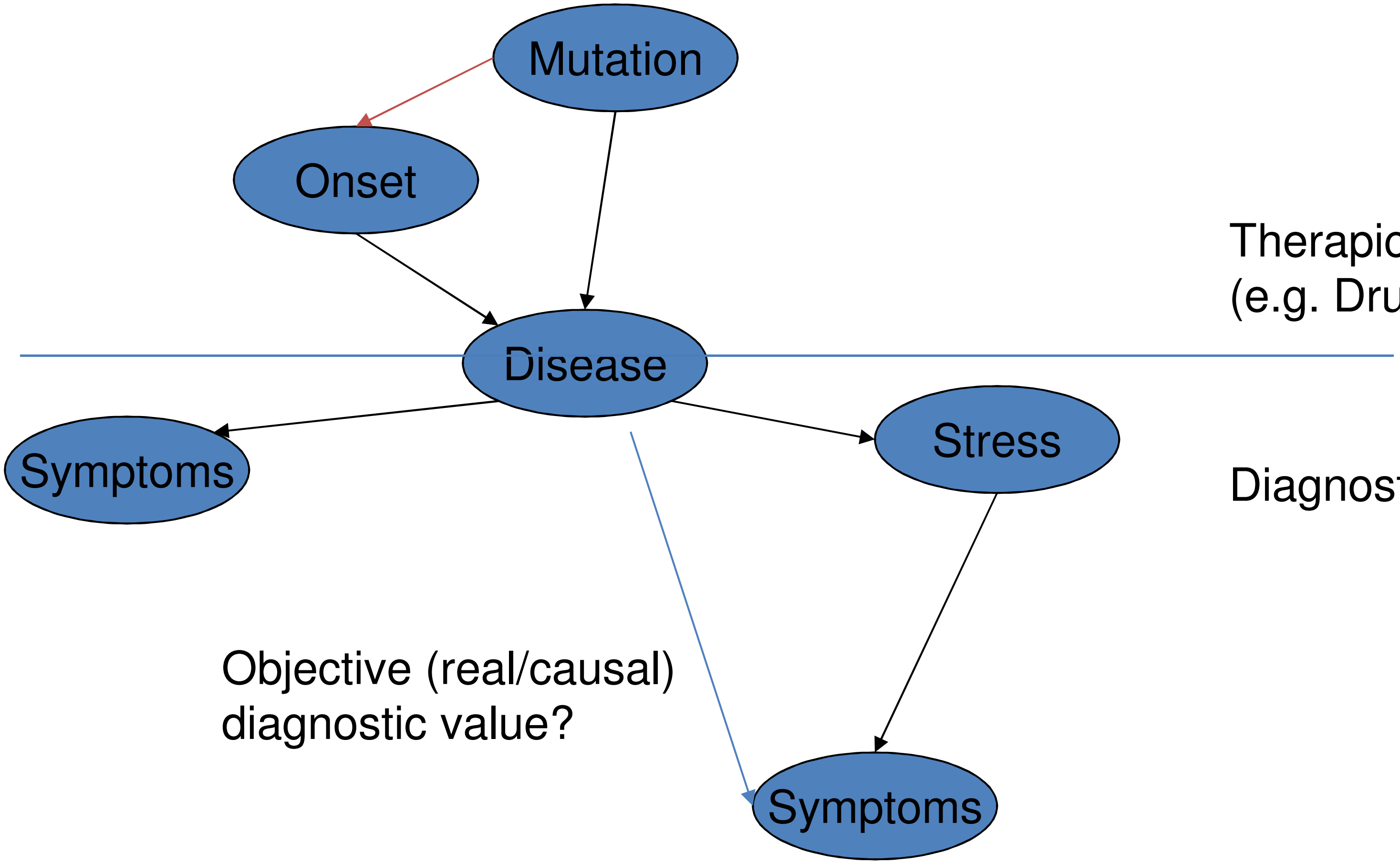
CA125 and its pretenders

- CA125 is very indicative tumor marker in cancer.
- Missing the mark, 2007, Nature



- MISSING THE MARK: *Why is it so hard to find a test to predict cancer?, 2011, March*
  - Series of proteomic kits... with poor performance.
  - The „prolactin” case

Causal vs. diagnostic markers
Direct =/= Causal



Mutation

Onset

Disease

Symptoms

Stress

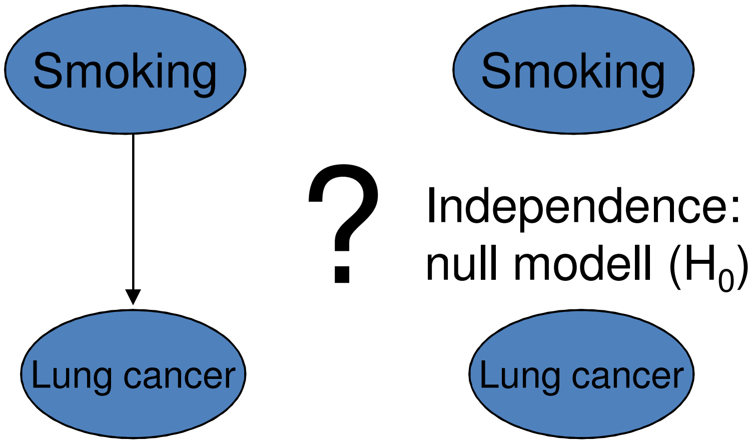
Symptoms

Therapic value
(e.g. Drug target)

Diagnostic value

Objective (real/causal)
diagnostic value?

Conditional probabilities, odds, odds ratios



Independence: null modell (H_0)

	¬S	S	
¬LC	8	7	15
LC	1	4	5
	9	11	20

Contingency table with marginals

	¬S	S	
¬LC	.4	.35	.75
LC	.05	.2	.25
	.45	.55	

Conditional probabilities:

P(LC| ¬S)=.11 ??? P(LC| S)=.36 ??? P(LC)=.25

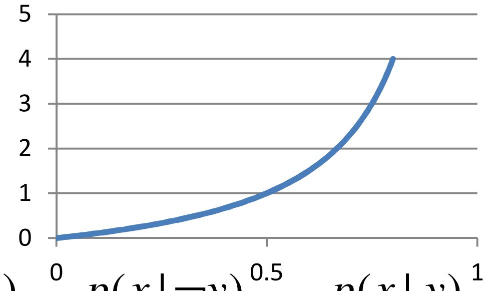
Odds:

[0,1] →[0,∞]: Odds(p)=p/(1-p)

O(LC| ¬S)=.12 ??? O(LC| S)=.56

Odds Ratio (OR):

OR(LC,S)=O(LC| S)/O(LC| ¬S)=4.6



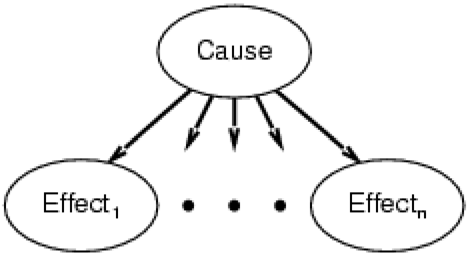
$$\frac{p(x \mid y)}{p(\neg x \mid y)} : \frac{p(x \mid \neg y)}{p(\neg x \mid \neg y)} = \frac{p(x \mid y)}{p(\neg x \mid y)} \frac{p(\neg x \mid \neg y)}{p(x \mid \neg y)}$$

➔prevalences + odds: joint distribution, e.g. P(LC, S)=P(LC| S) P(S)

Naive Bayesian network

Assumptions:

1, Two types of nodes: a cause and effects.

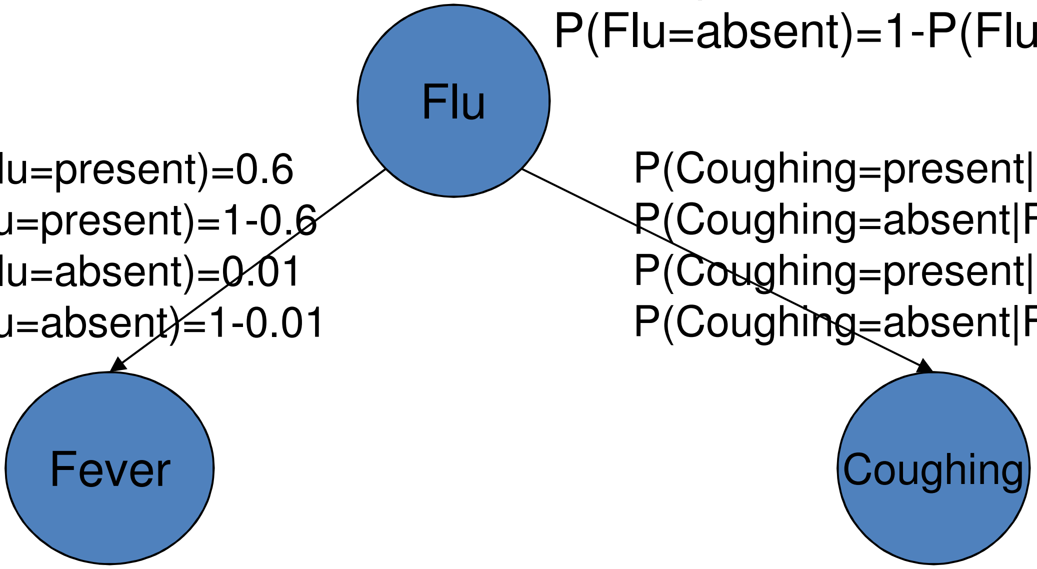


2, Effects are conditionally independent of each other given their cause.

Variables (nodes)

Flu: present/absent
FeverAbove38C: present/absent
Coughing: present/absent

Model



P(Flu=present)=0.001
P(Flu=absent)=1-P(Flu=present)

P(Fever=present|Flu=present)=0.6
P(Fever=absent|Flu=present)=1-0.6
P(Fever=present|Flu=absent)=0.01
P(Fever=absent|Flu=absent)=1-0.01

P(Coughing=present|Flu=present)=0.3
P(Coughing=absent|Flu=present)=1-0.7
P(Coughing=present|Flu=absent)=0.02
P(Coughing=absent|Flu=absent)=1-0.02

Naive Bayesian network (NBN)

Decomposition of the joint:

$P(Y,X_1,..,X_n) \quad = P(Y)\prod_i P(X_i,|Y, X_1,..,X_{i-1})$ //by the chain rule

$= P(Y)\prod_i P(X_i,|Y)$ // by the N-BN assumption

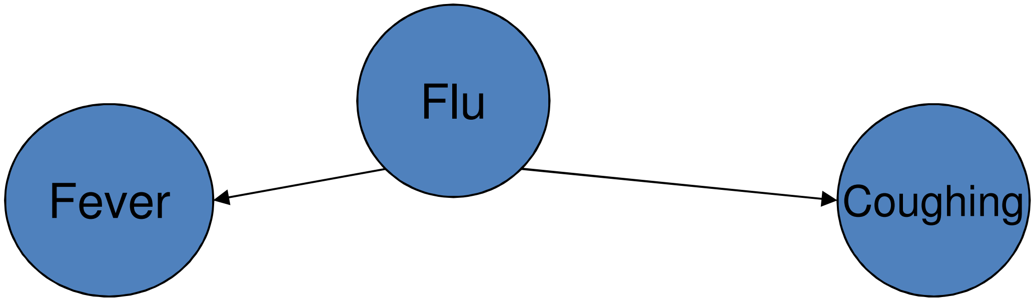
2n+1 parameteres!

Diagnostic inference:

$P(Y|x_{i1},..,x_{ik}) \quad = P(Y)\prod_j P(x_{ij},|Y) / P(x_{i1},..,x_{ik})$

If Y is binary, then the odds

$P(Y=1|x_{i1},..,x_{ik}) / P(Y=0|x_{i1},..,x_{ik}) \; = P(Y=1)/P(Y=0) \prod_j P(x_{ij},|Y=1) / P(x_{ij},|Y=0)$



$$p(Flu = present \mid Fever = absent, Coughing = present)$$

$$\propto p(Flu = present)\, p(Fever = absent \mid Flu = present)\, p(Coughing = present \mid Flu = present)$$

Logistic regression

Assume binary outcomes $y, \bar{y}$ and predictors $x_i, \bar{x}_i$. Logistic regression without interactions can be defined by the odds ratios for the predictors x_i, $i = 1, \ldots, n$ and the bias Ψ_0 ($x_0 \triangleq 1$):

$$\Psi_i = \frac{P(y|x_i)P(\bar{y}|\bar{x}_i)}{P(\bar{y}|x_i)P(y|\bar{x}_i)} \triangleq \exp^{\beta_i}, \Psi_0 = \prod_{i=0}^{n} \frac{P(y|\bar{x}_i)}{P(\bar{y}|\bar{x}_i)} \triangleq \exp^{\beta_0} .$$

The odds $P(y|\boldsymbol{x})/P(\bar{y}|\boldsymbol{x})$ for a given $\boldsymbol{x}$ is defined as

$$P(y|\boldsymbol{x})/P(\bar{y}|\boldsymbol{x}) = \prod_{i=0}^{n} \Psi_i^{x_i} \quad (18)$$

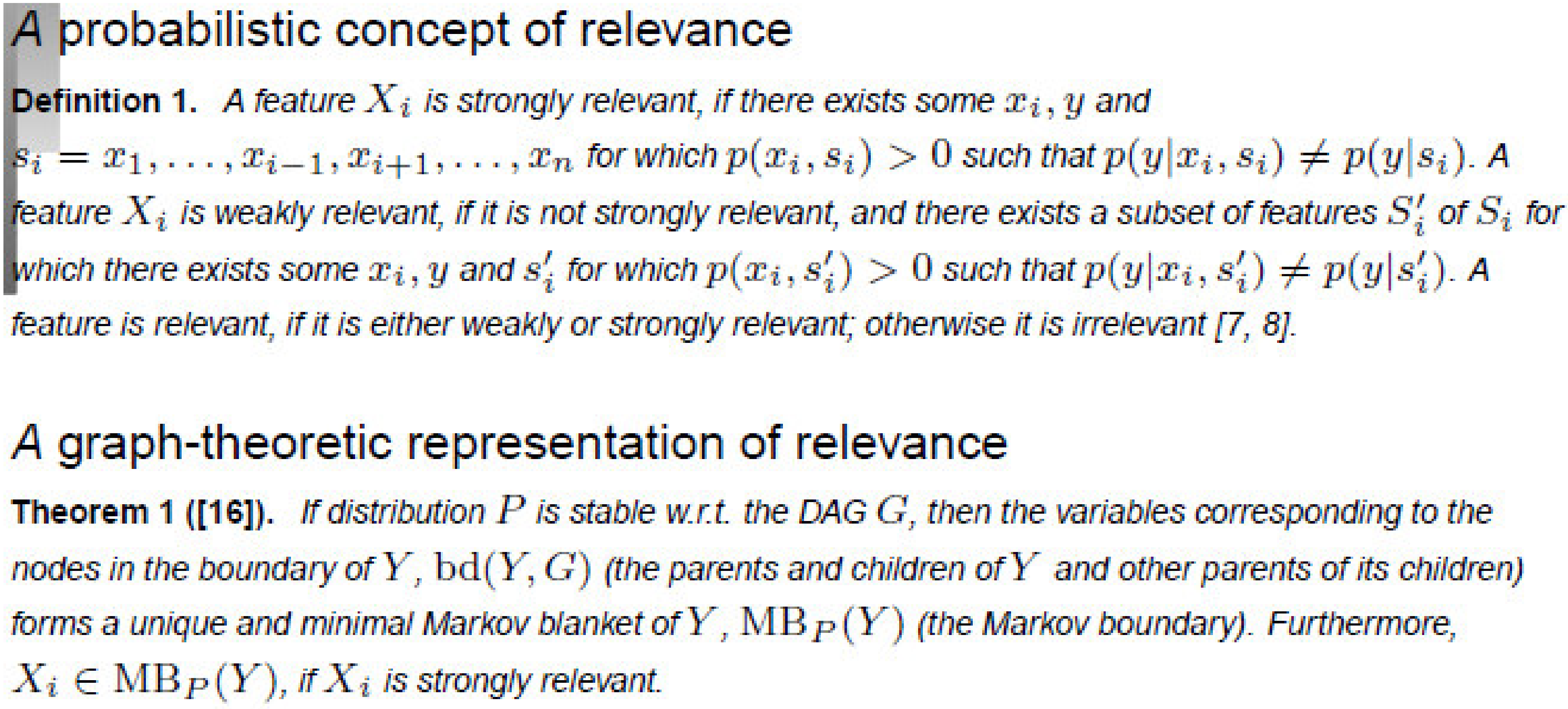
$$\log(P(y|\boldsymbol{x})/P(\bar{y}|\boldsymbol{x})) = \sum_{i=0}^{n} \beta_i x_i \quad (19)$$

$$P(y|\boldsymbol{x}) = \sigma(\sum_{i=0}^{n} \beta_i x_i), \quad (20)$$

where $\sigma()$ is the logistic sigmoid function$\sigma(x) = 1/(1 + e^{-x})$.

$$P(y|\boldsymbol{x}) = \sigma[\sum_{i=0}^{n}(\beta_i x_i + \sum_{j=1}^{n} (\beta_{i,j} x_i x_j + \sum_{k=1}^{n}(\beta_{i,j,k} x_i x_j x_k + \ldots)))],$$

Biomarkers and
the feature subset selection (FSS) problem



A probabilistic concept of relevance

Definition 1. *A feature X_i is strongly relevant, if there exists some x_i, y and $s_i = x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ for which $p(x_i, s_i) > 0$ such that $p(y|x_i, s_i) \neq p(y|s_i)$. A feature X_i is weakly relevant, if it is not strongly relevant, and there exists a subset of features S'_i of S_i for which there exists some x_i, y and s'_i for which $p(x_i, s'_i) > 0$ such that $p(y|x_i, s'_i) \neq p(y|s'_i)$. A feature is relevant, if it is either weakly or strongly relevant; otherwise it is irrelevant [7, 8].*

A graph-theoretic representation of relevance

Theorem 1 ([16]). *If distribution P is stable w.r.t. the DAG G, then the variables corresponding to the nodes in the boundary of Y, $\mathrm{bd}(Y, G)$ (the parents and children of Y and other parents of its children) forms a unique and minimal Markov blanket of Y, $\mathrm{MB}_P(Y)$ (the Markov boundary). Furthermore, $X_i \in \mathrm{MB}_P(Y)$, if X_i is strongly relevant.*

Bayesian networks

Directed acyclic graph (DAG)

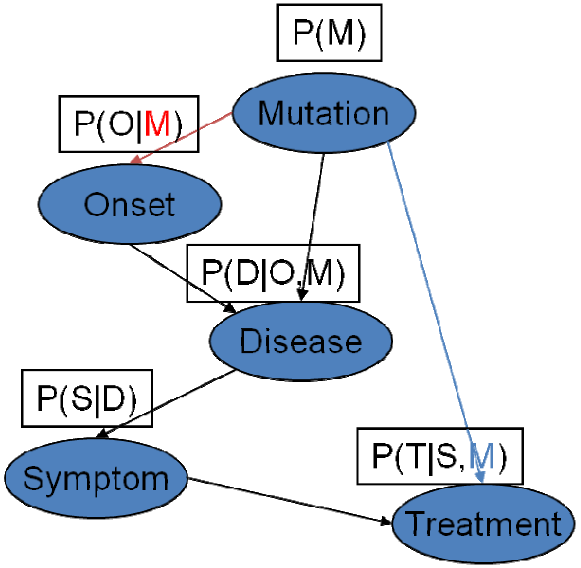
- nodes – random variables/domain entities
- edges – direct probabilistic dependencies
  (edges- causal relations

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

Three interpretations:

3. Concise representation of joint distributions

$$P(M,O,D,S,T) =$$
$$P(M)P(O|M)P(D|O,M)P(S|D)P(T|S,M)$$



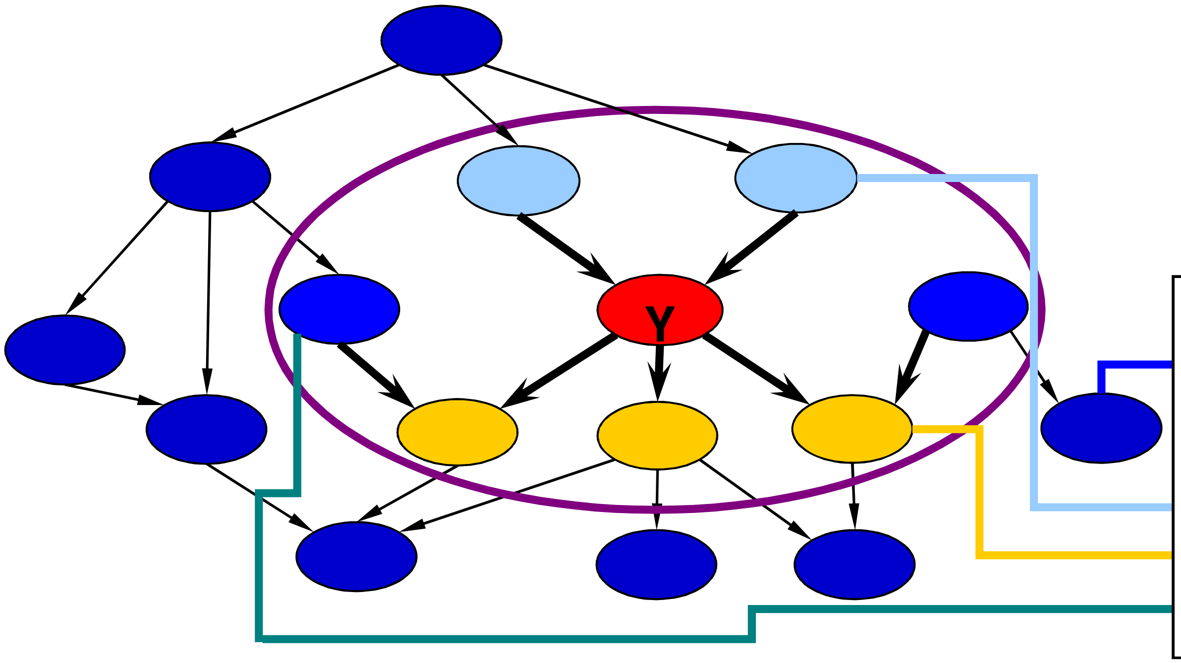
1. Causal model

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

2. Graphical representation of (in)dependencies

The Markov Blanket

A minimal sufficient set for prediction/diagnosis.



A variable can be:

- (1) non-occuring

Irrelevant (strongly)

- (2) parent of Y
- (3) child of Y
- (4) pure (other parent)

Relevant (strongly)

Markov Blanket Sets (MBS) the set of nodes which probabilistically isolate the target from the rest of the model

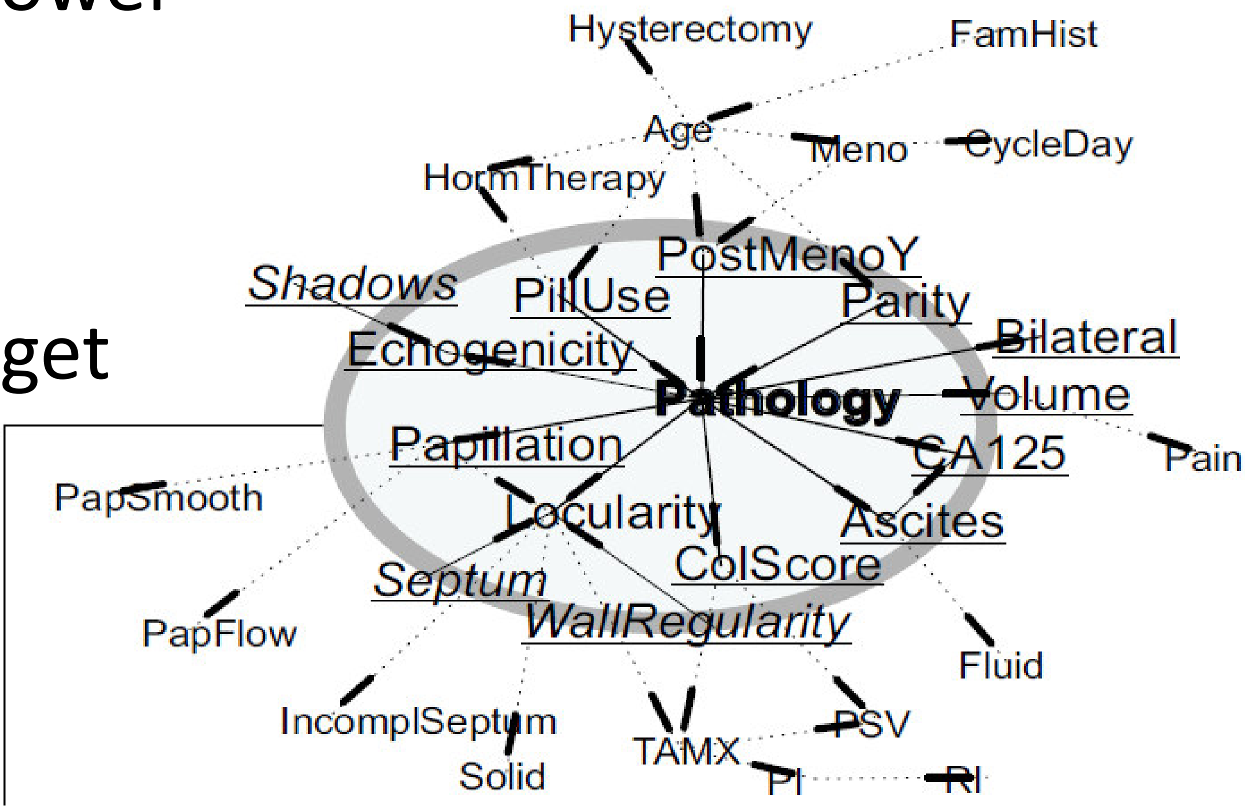
Markov Blanket Membership (MBM)

(symmetric) pairwise relationship induced by MBS

Aspects of biomarkers

„Maximum predictivity, minimum redundancy”

- Predictive power
- Directness
- Causality
- Multiple target
- Uncertainty



The wrapper approach to FSS

Assume our goal is an „efficient” predictor f(**X’**)=Y

1. Initialize set S with a priori good predictors
2. Cycle

   2.1 Modify S

   E.g. Greedy-univariate: select an additional predictor based on predictive power

   2.2 Improved predictive power?

   Misclassification rate, AUC, likelihood:P(Data|Model(S))
3. Terminate

The filter approach to FSS („local causal(?) discovery”)

1. Initialize set S with a priori good predictors
2. Cycle
    1. Expand S
        1. E.g. Greedy-univariate: select an additional predictor based on predictive power, which is still relevant to target Y given S
    2. Decrease S based on Markov blankets
        1. E.g. S-X shields X from target Y
3. Terminate

The hypothesis testing framework

- Terminology:
  - False/true x positive/negative
  - Null hypothesis: independence

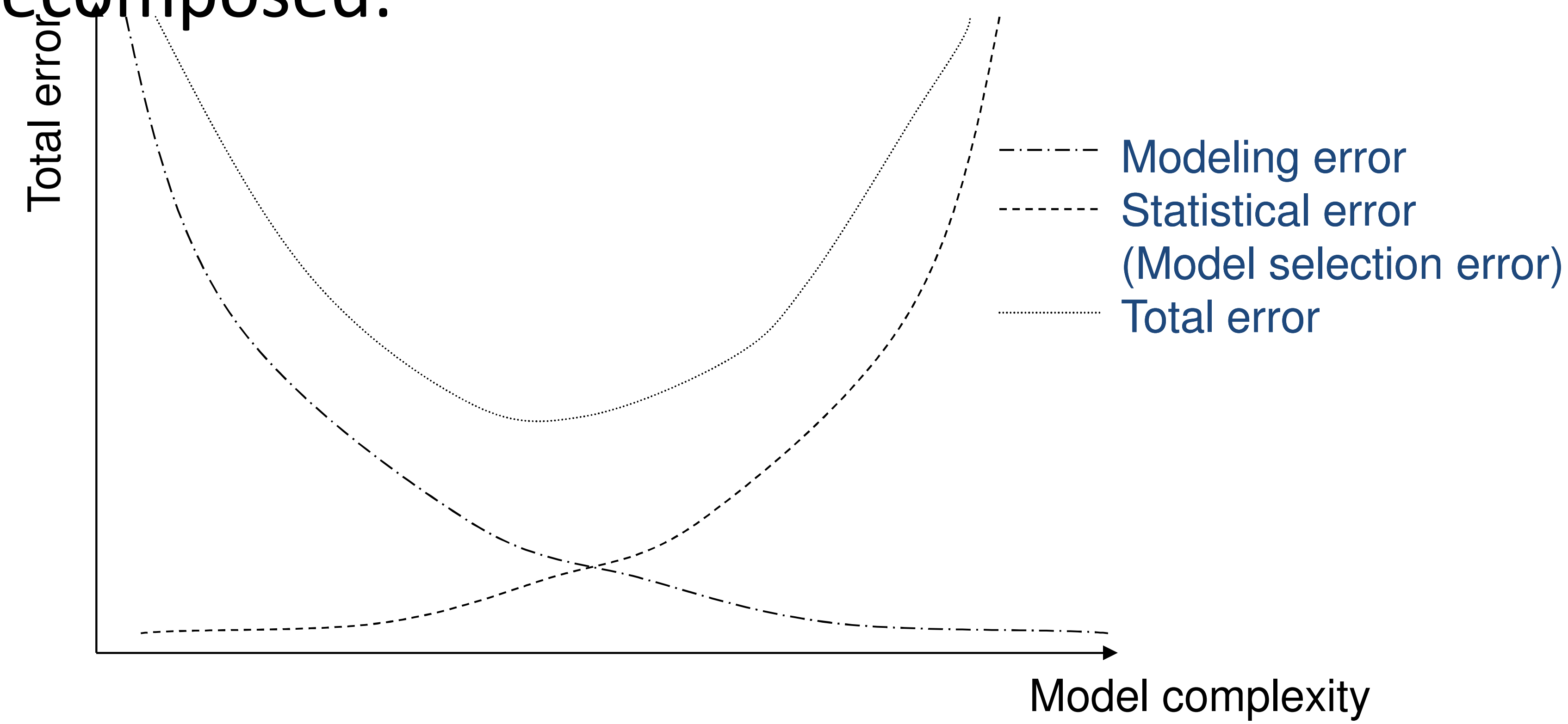
reported	Ref.:0/N	Ref.1/P
0/N	TN	FN
1/P	FP	TP

  - Type I error/error of the first kind/α error/FP: $p(\neg H_0|\underline{H}_0)$
    - Specificity: $p(H_0|\underline{H}_0) =1-\alpha$
    - Significance: α
    - p-value: „probability of more extreme observations in repeated experiments”
  - Type II error/error of the second kind/β error/FN: $p(H_0| \neg \underline{H}_0)$:
    - Power or sensitivity: $p(\neg H_0| \neg\underline{H}_0) = 1-\beta$

reported	Ref. $\underline{H}_0$	Ref.:$\neg\underline{H}_0$
H_0		Type II
$\neg H_0$	Type I („false rejection”)	

The bias-variance dilemma

- For a given sample size the error is decomposed:



Bayes rule, Bayesianism

„all models are wrong, but some are useful”

$$p(X \mid Y) = \frac{p(Y \mid X)\, p(X)}{p(Y)}$$

A scientific research paradigm

$$p(Model \mid Data) \propto p(Data \mid Model)\, p(Model)$$

A practical method for inverting causal knowledge to diagnostic tool.

$$p(Cause \mid Effect) \propto p(Effect \mid Cause) \times p(Cause)$$

Bayesian prediction

In the frequentist approach: Model identification (selection) is necessary

$$p(prediction \mid data) = p(prediction \mid BestModel(data))$$

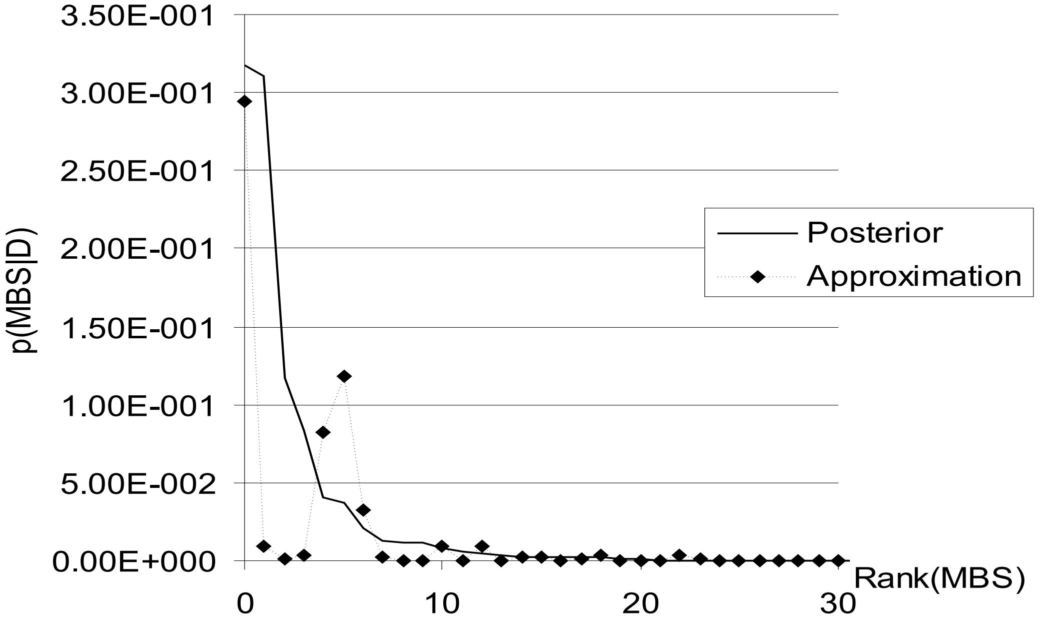
In the Bayesian approach models are weighted

$$p(prediction \mid data) = \sum_i p(pred. \mid Model_i)\, p(Model_i \mid data)$$

Note: in the Bayesian approach there is no need for model selection

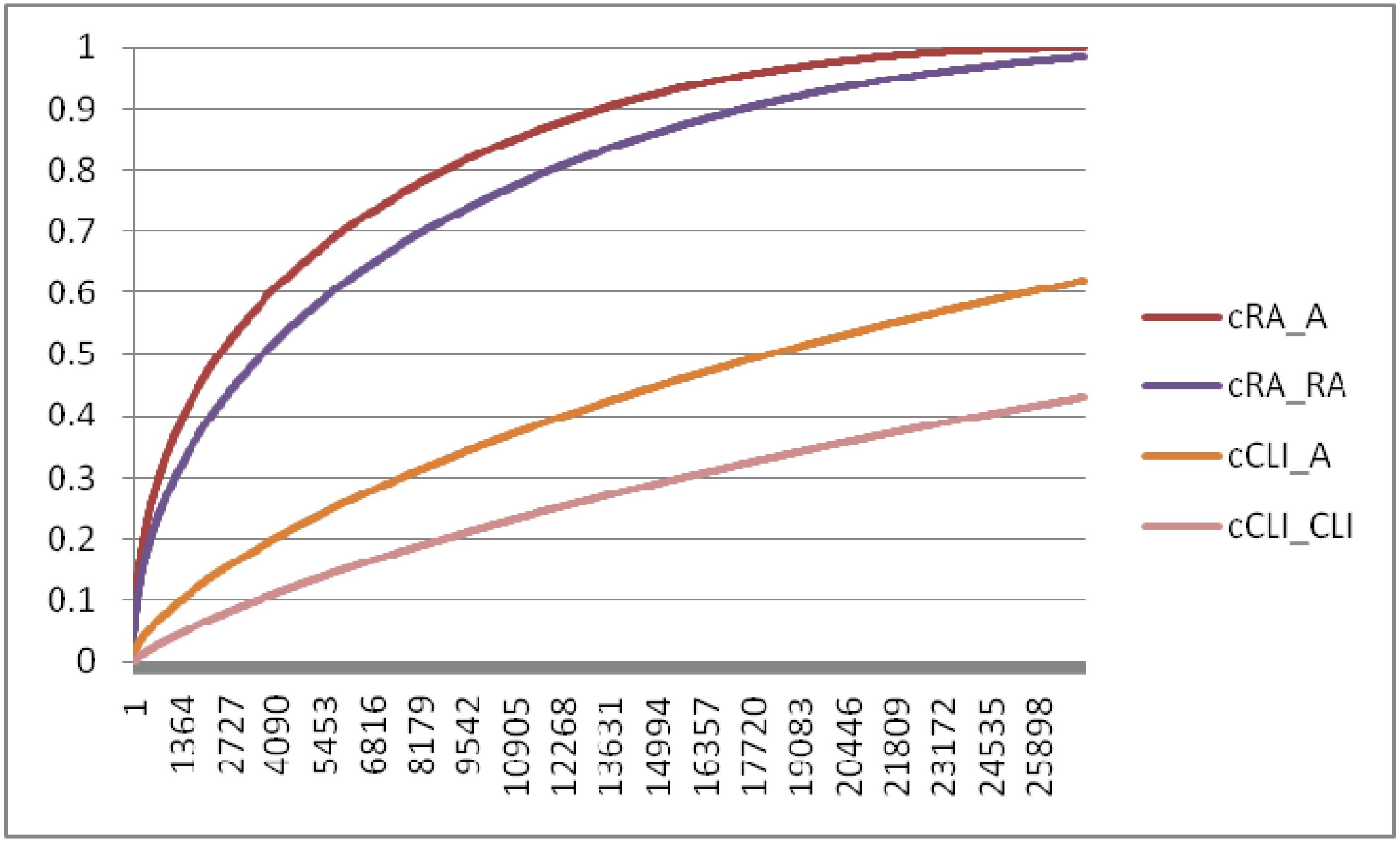
Posterior for complete sets

- High-level of uncertainty in multivariate analysis
- There are stable sub-parts (e.g., subset, subgraphs)
- Results for target variables and for certain SNPs could be aggregated

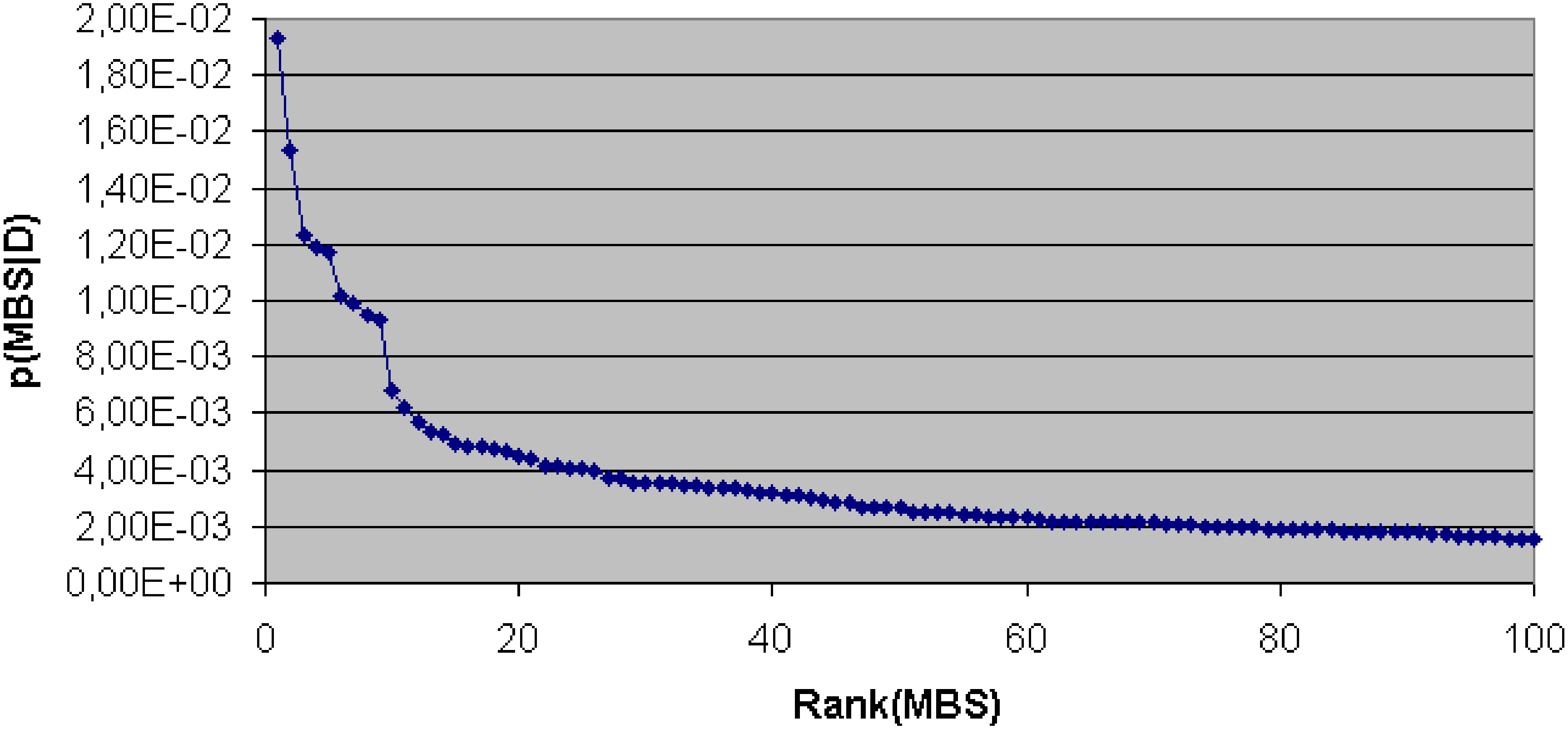


The peakness of the posteriors of the most probable MB sets and their MBM-based approximations.
(46 variables, 1000 samples)

Cumulative posterior of the most probable strongly relevant sets

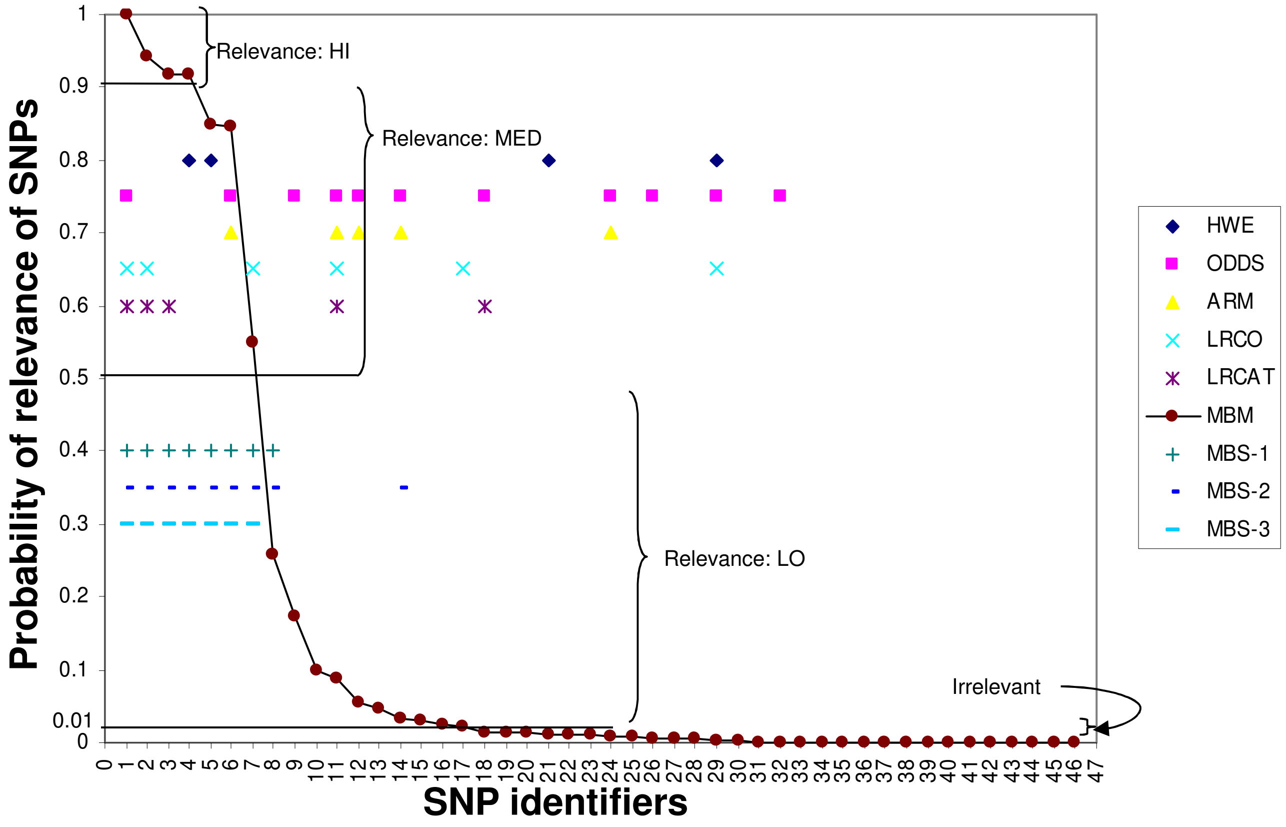


MBS posteriors in Asthma



Posteriors of strong relevance

HWE – Hardy-Weinberg equilibrium test, *ODDS* – odds ratio, *ARM* – Cochran-Armitage trend test, *LRCO* – logistic regression (continuous case), *LRCAT* – logistic regression (categorical case), *MBM* – Bayesian pairwise relevance, *MBS-1–9* relevant sets by Bayesian analysis. (only MBM values are numeric, others are arbitrary values for visualization)

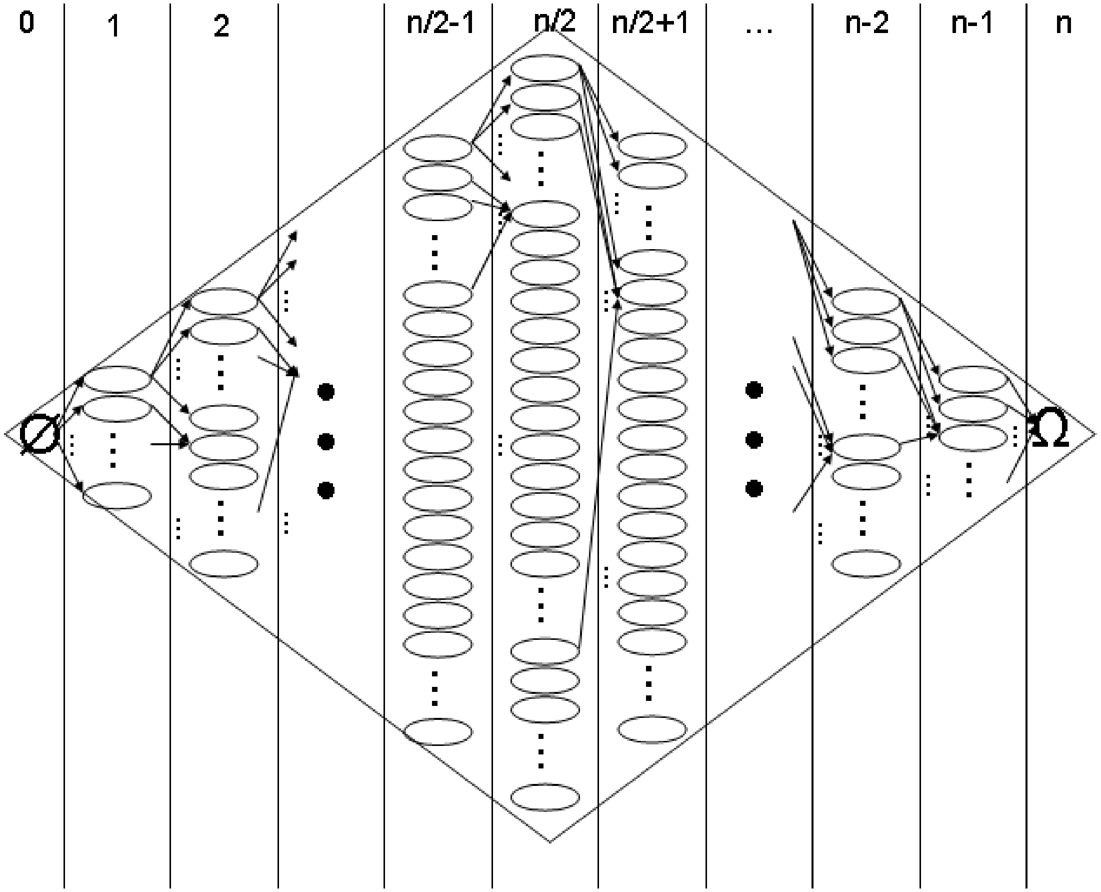


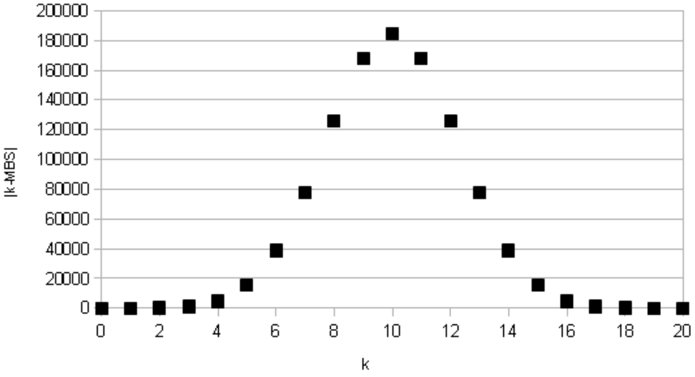
Frequentist vs Bayesian statistics

Frequentist	Bayesian
-	Prior probabilities
Null hypothesis	-
Indirect: proving by refutation	Direct
Model selection	Model averaging
Likelihood ratio test	Bayes factor
p-value	-!
-!	Posterior probabilities
Confidence interval	Credible region
Significance level	Optimal decision based on Exp.Util.
Multiple testing problem	Remains, so ➔ **complex model**
Model complexity dilemma	**Best achievable alternative**

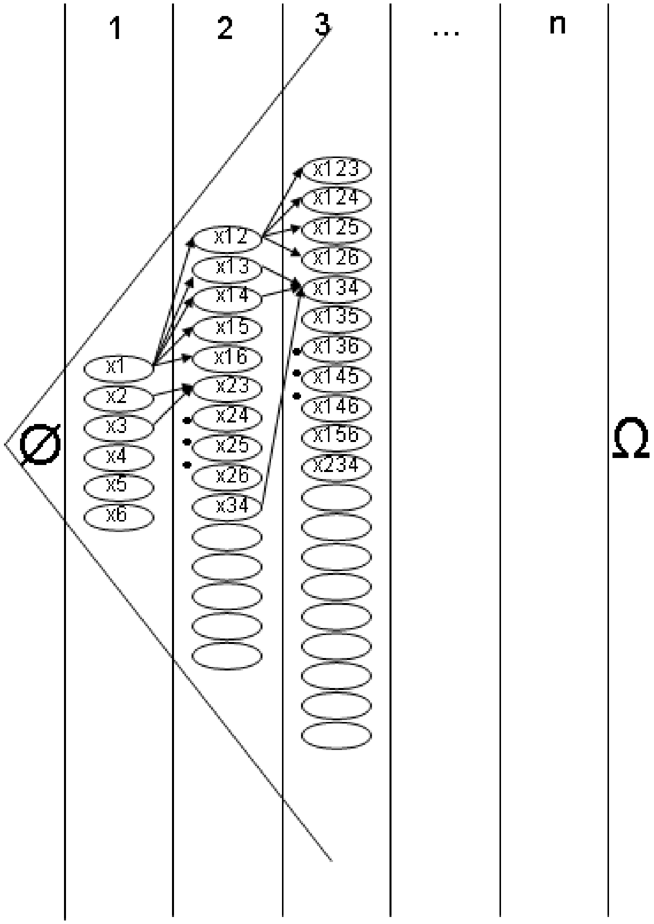
- Note: direct probabilistic statement!

The subset space

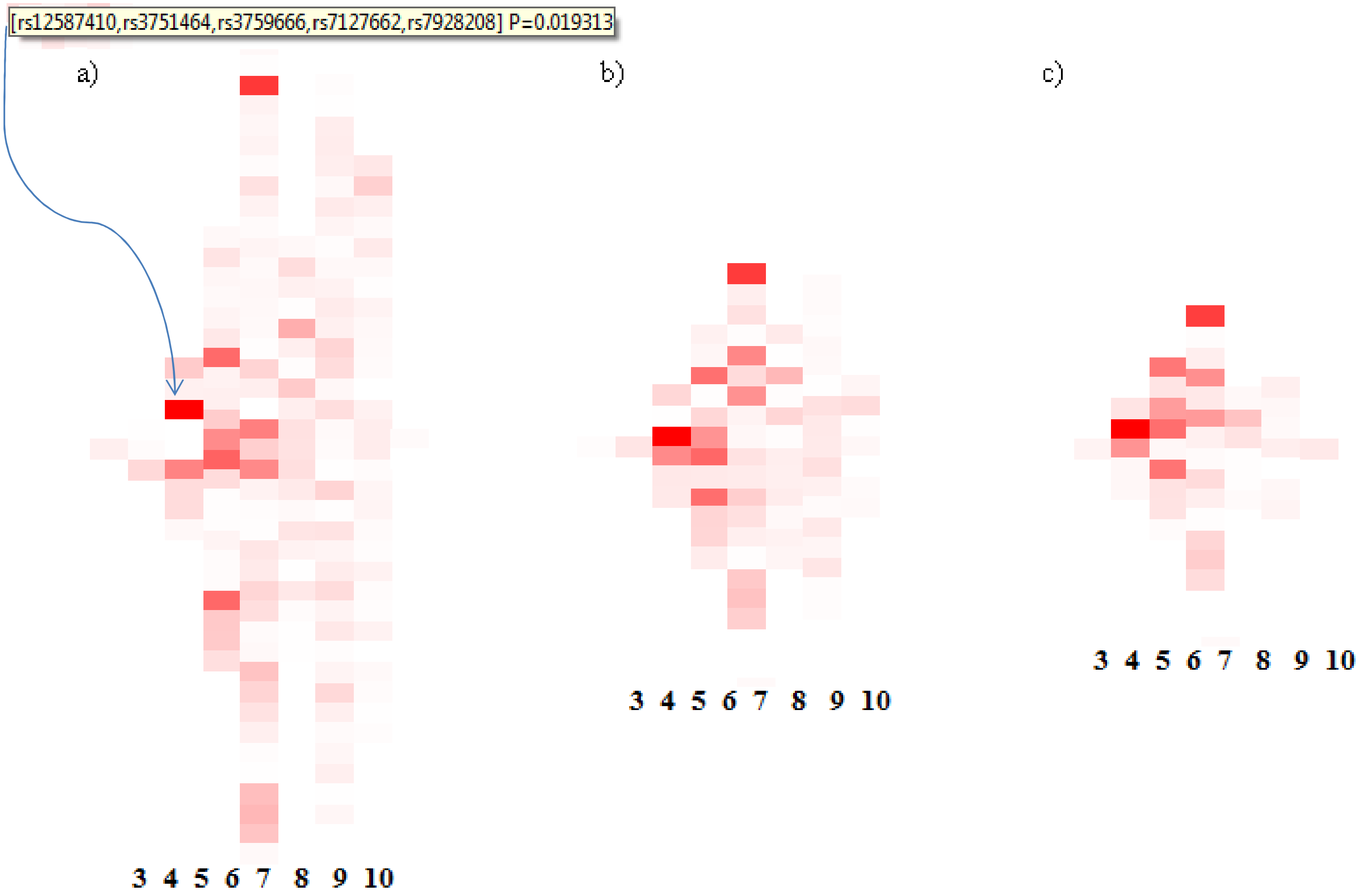




The subset space II.



An MBS heatmap in the subset space

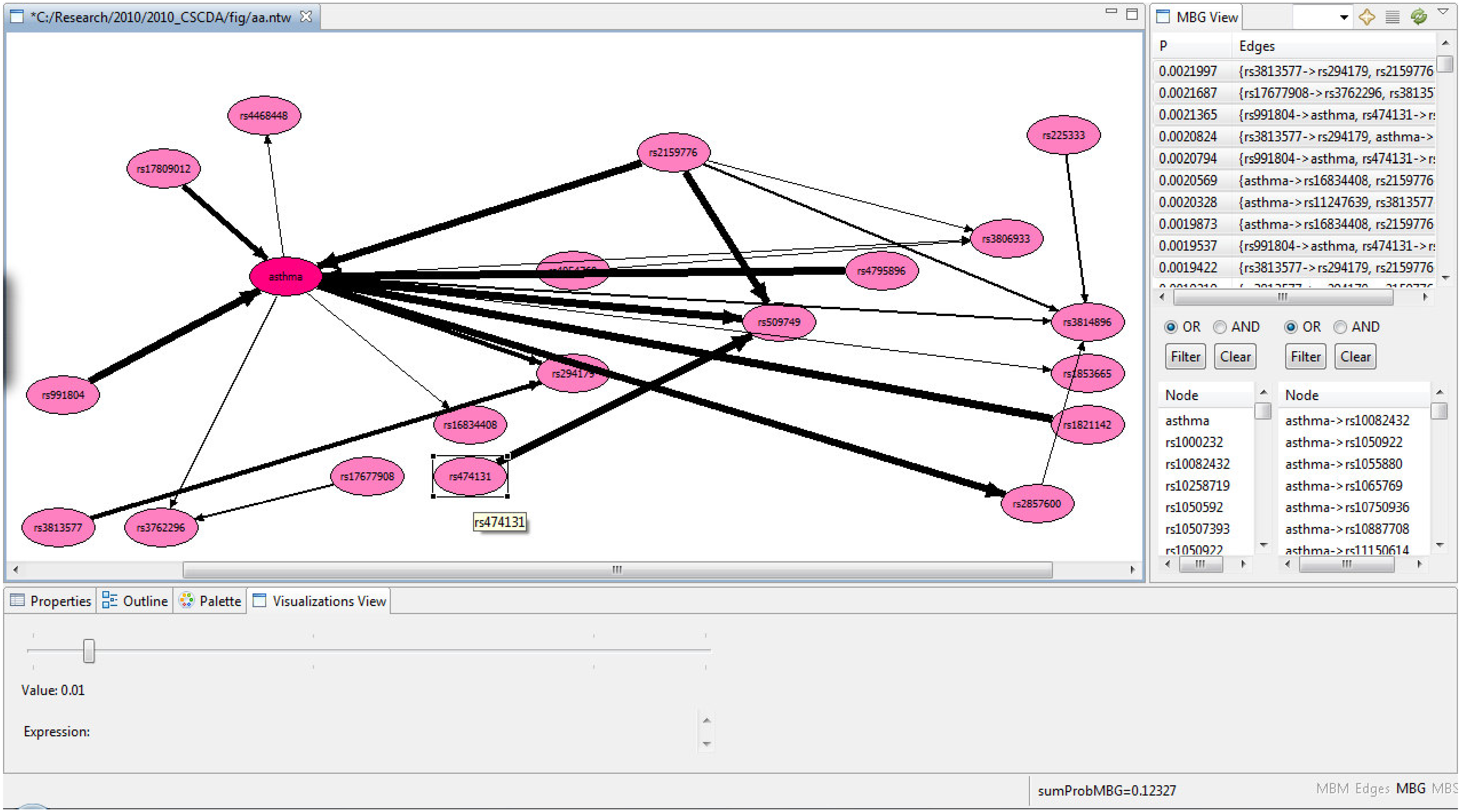


[rs12587410,rs3751464,rs3759666,rs7127662,rs7928208] P=0.019313

Genagrid

- SGI Altix ICE
  - – 5 TFLOPS
  - – 1TB memory
  - – 64x8 cores
  - – FPGAs

BayesEye



P	Edges
0.0021997	{rs3813577->rs294179, rs2159776
0.0021687	{rs17677908->rs3762296, rs38135
0.0021365	{rs991804->asthma, rs474131->r
0.0020824	{rs3813577->rs294179, asthma->
0.0020794	{rs991804->asthma, rs474131->r
0.0020569	{asthma->rs16834408, rs2159776
0.0020328	{asthma->rs11247639, rs3813577
0.0019873	{asthma->rs16834408, rs2159776
0.0019537	{rs991804->asthma, rs474131->r
0.0019422	{rs3813577->rs294179, rs2159776

● OR ○ AND ● OR ○ AND

Filter Clear Filter Clear

Node
asthma
rs1000232
rs10082432
rs10258719
rs1050592
rs10507393
rs1050922

Node
asthma->rs10082432
asthma->rs1050922
asthma->rs1055880
asthma->rs1065769
asthma->rs10750936
asthma->rs10887708
asthma->rs11150614

Value: 0.01

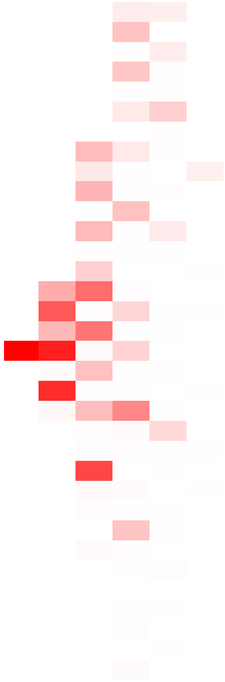
Expression:

sumProbMBG=0.12327

Marginal multivariate posteriors in the subset space?

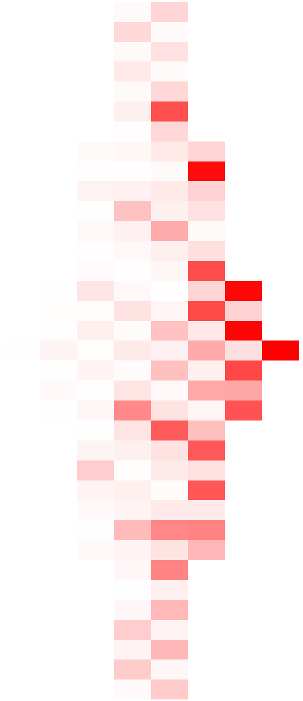
k-MBS-sub

$p(G : s \subseteq MBS(G) \mid D_N)$



k-MBS-sup

$p(G : s \supseteq MBS(G) \mid D_N)$



Summary

- Feature relevance
- The feature subset selection problem
- Identification of biomarkers
  - Methods
- Challenges
  - Interpretation➔ Bayesian networks
  - Causality➔ Bayesian networks
  - Uncertainty ➔ Bayesian statistics
- A Bayesian network based Bayesian approach to biomarker analysis