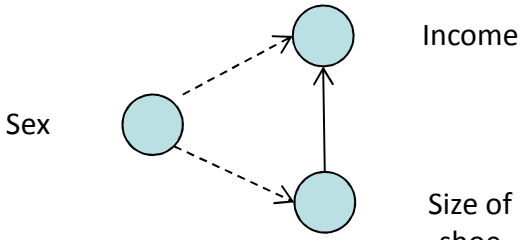
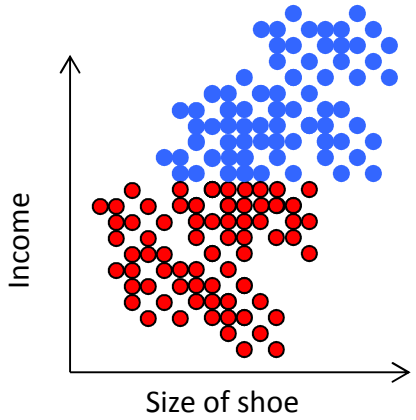
Enhancing Statistical Inference by Exploiting Knowledge Structures

Ulrich Mansmann,
IBE, Ludwig-Maximilian-Universität München
ulrich.mansmann@lmu.de

Enhancing Statistical Inference by Exploiting Knowledge Structures

Knowledge structures concept of sound analysis





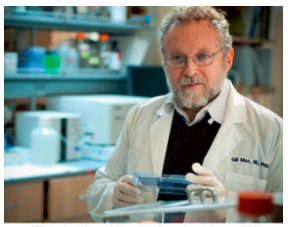
Knowledge as intrinsic part of the statistical method

Bayesian statistics

Prior + Data → Posterior

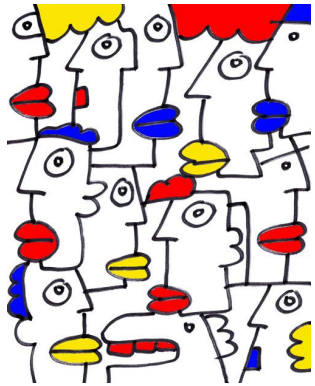
Predictive value of a diagnostic test

Knowledge of sensitivity and specificity are not enough to assess the value of a test environment of its application also matters



428 | NATURE | VOL 471 | 24 MARCH 2011

Taming uncertainty



- Control precision of unbiased estimates,
  avoid false positive and false negative statements;

- Data derived information has to be handled care;

- Explain variability by disclosing systematic structure and noise,
  Sources of variability: Systematic and random,
  Reduce bias and suppress random effects;

- Discovery – Validation:
  What are good candidates for a successful validation?
  Explore scope of explanatory constructs.

- Most people see statistical activities as a ritual, and not as a structured approach to create useful information (knowledge).

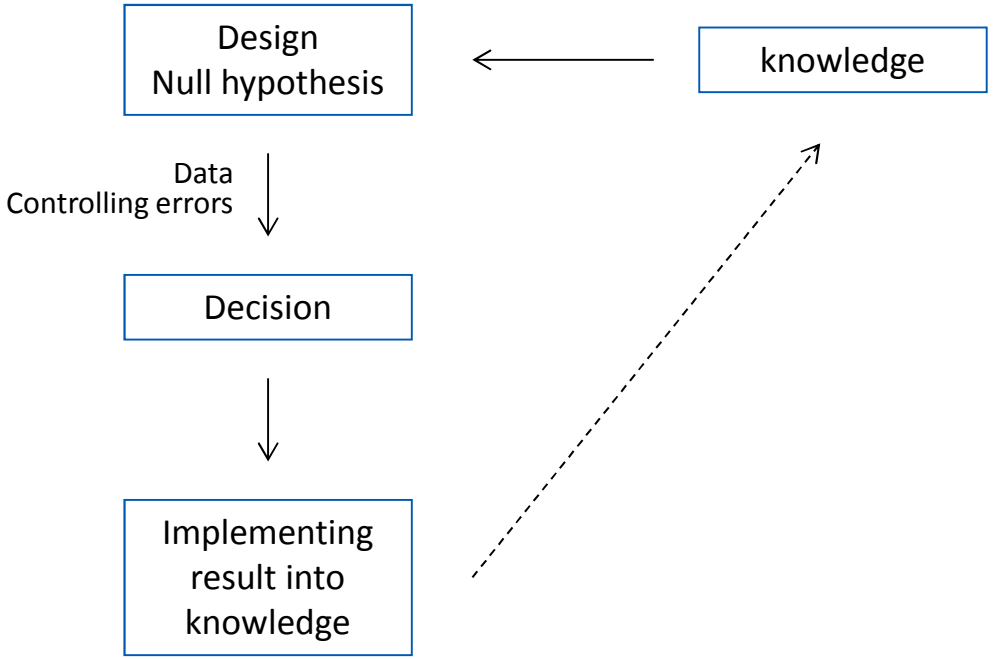
Statistic provides information but information is not knowledge

Yet, it is knowledge that leads to good decision-making and spurs progress.

What do we know about the effect of interest?
Group sequential versus fixed design
Adaptive designs

Internal and external validity

General interpretation of the results in the context of current evidence



Exploration - Confirmation

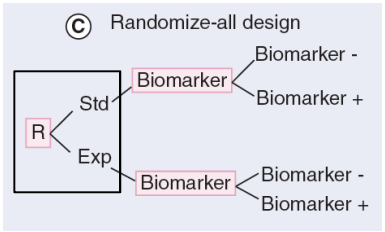
Two extremes of a wide spectrum

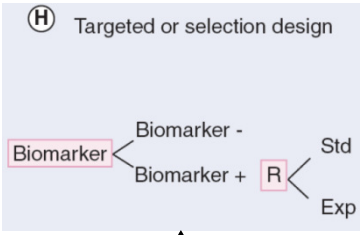
Predictive biomarker for a specific treatment

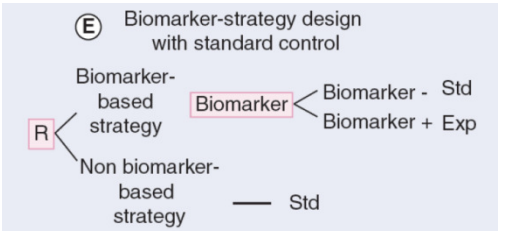
Screening

POC

Clinical relevance



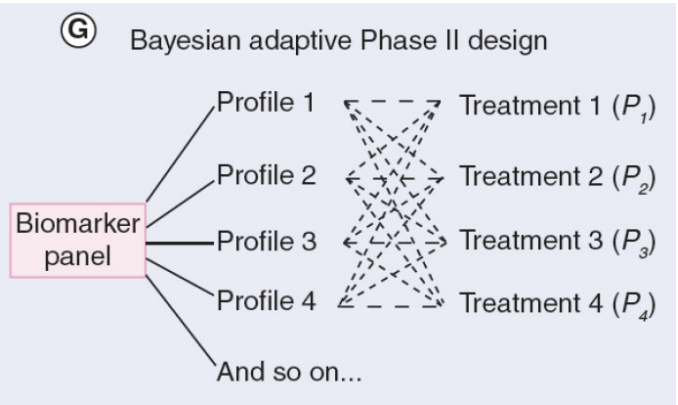




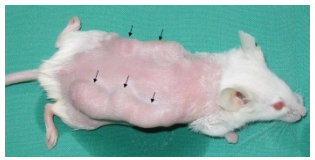
Large Phase III study with biobanking

Validate effect in biomarker positive group
Regression to the mean

Complex Exploration



preclinical

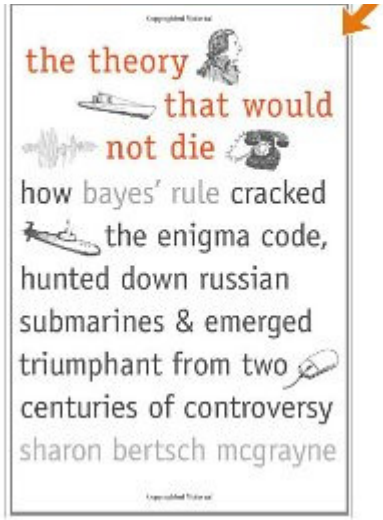


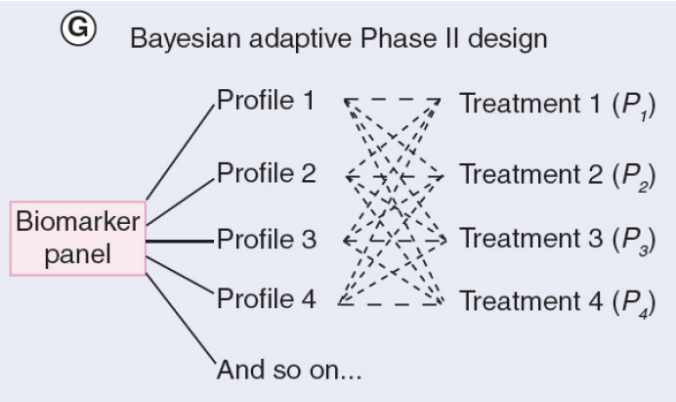
Clinical
Phase II
Multi arm, multi treatment trials

Challenge:

- Selection of biomarker candidates
- Selection of treatments
- Selection of promising combinations

Complex Exploration





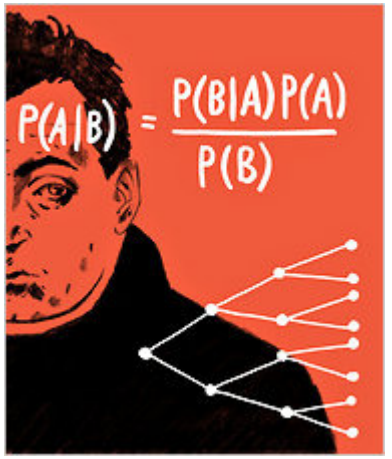


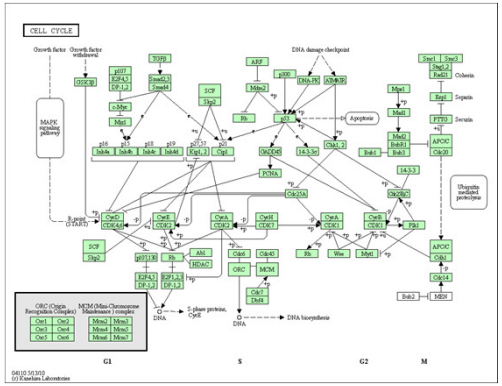
Illustration by Shannon May

- The theorem itself can be stated simply.
- Beginning with a provisional hypothesis about the world (there are, of course, no other kinds), we assign to it an initial probability called the prior probability or simply the prior.
- After actively collecting or happening upon some potentially relevant evidence, we use Bayes's theorem to recalculate the probability of the hypothesis in light of the new evidence.
- This revised probability is called the posterior probability or simply the posterior.

Computable Knowledge

Machine-readable knowledge bases store knowledge in a computer-readable form, usually for the purpose of having automated deductive reasoning applied to them. They contain a set of data, often in the form of rules that describe the knowledge in a logically consistent manner. An *ontology* can define the structure of stored data - what types of entities are recorded and what their relationships are.

Computable knowledge: machine-readable knowledge base can be accessed by algorithms.



KEGG Pathway: Cell Cycle

Advances in computable knowledge

Gene Ontology

- Biological Process
- Cellular Component
- Molecular Function

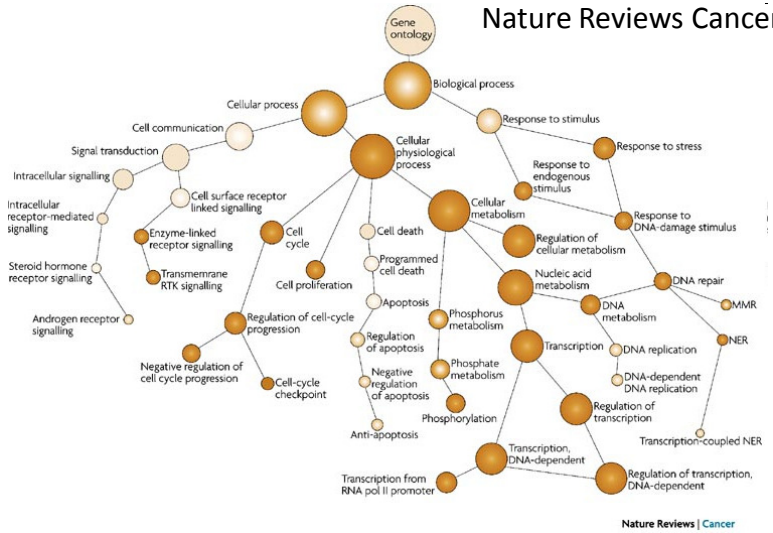
(organigrams of a complex company)

Open Biomedical Ontology (OBO)

Transcription factors

Functional gene sets

Nature Reviews Cancer (2007) 7, 23-34



MSigDB
Molecular Signatures Database

THE NATIONAL CENTER FOR BIOMEDICAL ONTOLOGY

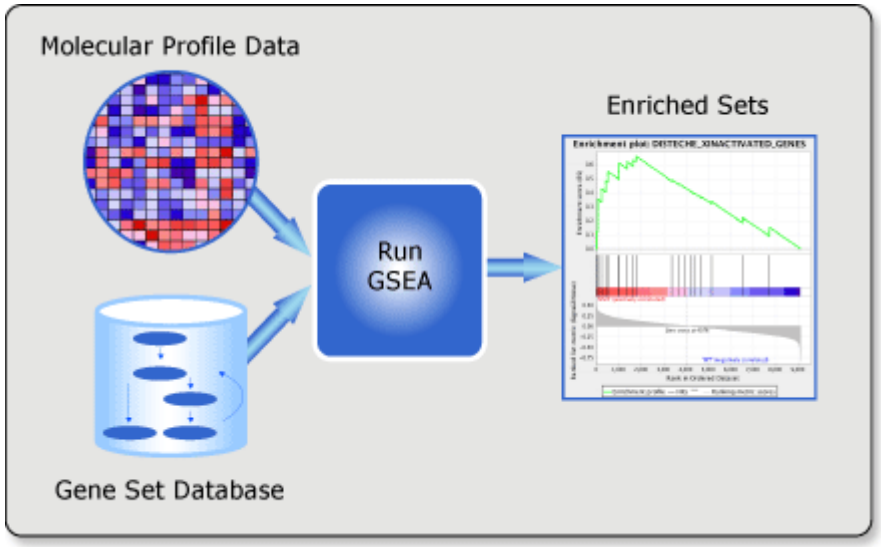
Integrating knowledge into statistical algorithms

- Gene set enrichment analysis
- Focus level analysis
- Global tests and the search in the haystack
- Graphs and networks

Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

Subramanian, Tamayo, et al. (2005, PNAS 102, 15545-15550) and Mootha, Lindgren, et al. (2003, Nat Genet 34, 267-273)

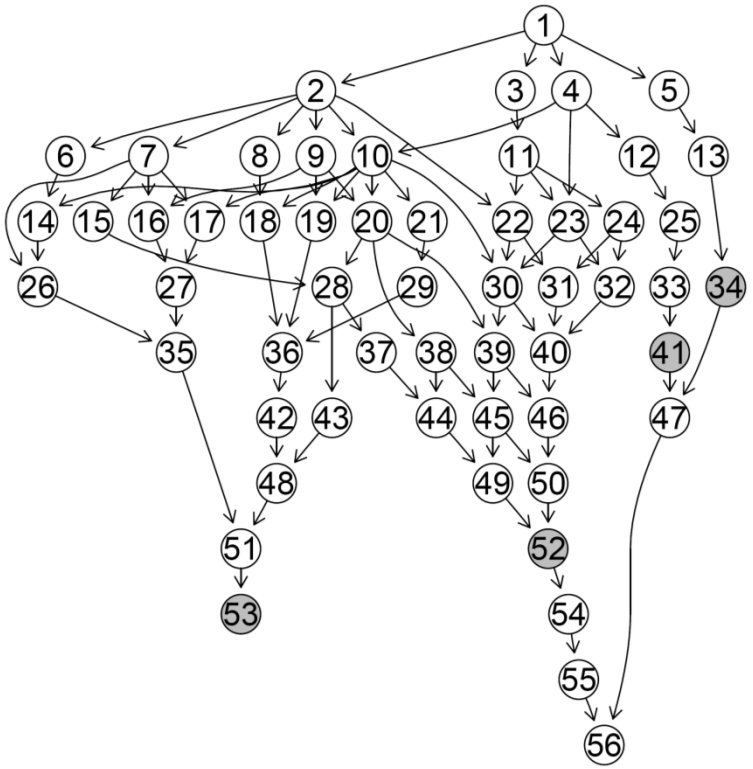


Focus level method

Make use of GO's logical structure to improve power of statistical multiple testing.

238 AML patients with normal karyotype, Recently, mutations in two genes, namely the fms-like tyrosine kinase 3 (FLT3) and nucleophosmin 1 (NPM1) were shown to be relevant for prognosis of AML patients with normal karyotype.

Results on the pure FLT3 effect. The pure effect of one mutation is defined by adjusting the analysis for the effect of the second mutation.

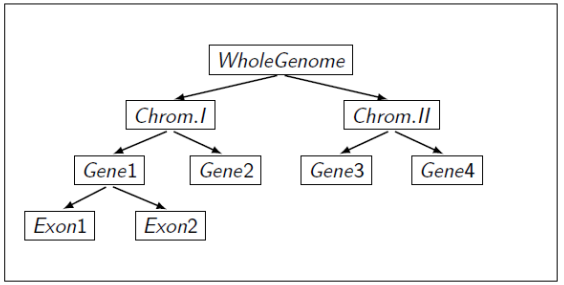


Node 53 (GO:0006427, histidyl-tRNA aminoacylation)
Node 56 (GO:0006994, sterol depletion response, SREBP target gene transcriptional activtion)

Goeman, Mansmann (2008) Bioninformatics, 537–544

Global tests and the search in the haystack

- Meinshausen (2008)
- Goeman, Solari, Finos (2011)
- Etc.



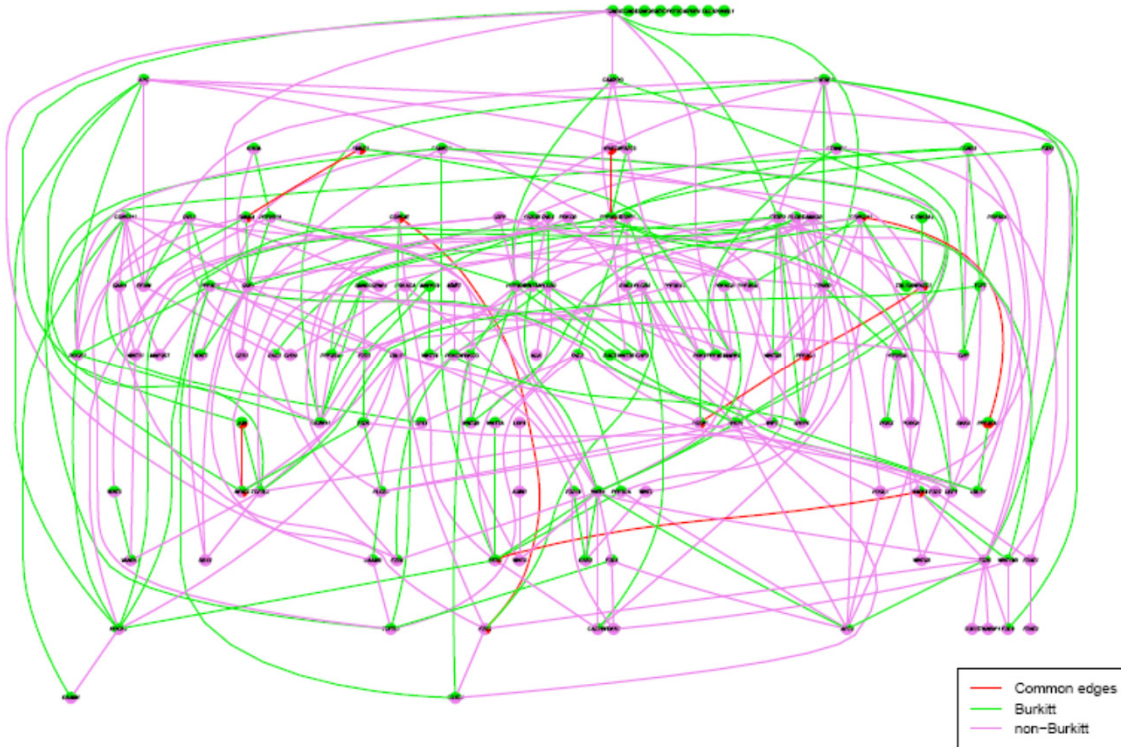
Global test: Is in a specific context something of interest?

Sequential rejection of structured null hypotheses

- FWER control procedures are often sequential:
  Critical values depend on previous rejections
- More and less interesting hypotheses:
  Test more interesting hypotheses before others
- Logically related hypotheses:
  Logical relationships induce testing order
- Naturally (generic) sequential procedures:
  Holm as a sequential improvement of Bonferroni

Network and graphs

Complementary merging of biological knowledge and statistical methods promises to be productive in helping unravel the functional complexity underlying cancer cell functions.



Wnt Pathway

Change of interaction structure induced by MYC translocation in lymphoma patients

Need of an objective and structured way how to add a biological interpretation to these findings.

Mansmann, Jurinovic (2011) Brief Bioinform. 12:230-44

Computational advances to infer *network* or interaction between genes

Networks inferred from gene expression data

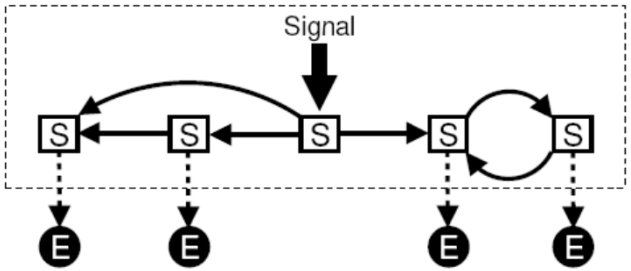
based on statistical reasoning: glasso, PC, GeneNet

Bayesian networks

based on bioinformatics techniques: ARACNE, C3NET

Networks inferred from gene silencing data (interventional data)

Nested effects models



S-genes (for "signaling" or "silenced"): candidate pathway genes where intervention takes place

E-genes (for "effects"): reporters for S-gene activity.

F. Markowetz, J. Bloch, R. Spang (2005) Non-transcriptional Pathway Features Reconstructed from Secondary Effects of RNA Interference, *Bioinformatics 21: 4026-4032*

Challenges and Visions

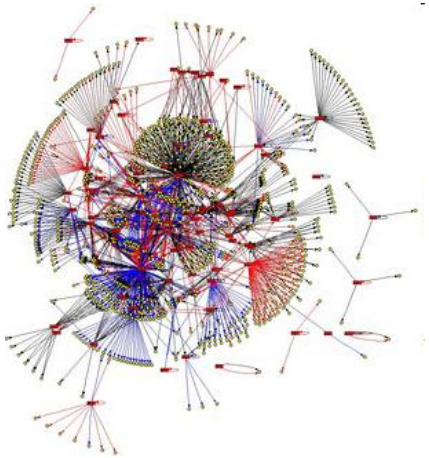
How good is the knowledge we use?

How transparent and objective is the way we incorporate knowledge into analysis strategies and designs: protocols and algorithms?

Is the knowledge construct tailored to our results:
Bradford-Hill criteria for causality: *plausibility* and *coherence*

Knowledge-free analysis: top-down systems approach
Knowledge based analysis: bottom-up

Knowledge-free statistical strategies – claimed to be the most objective way to do science. Even if the analysis seems to be know-ledge free, the knowledge is implicitely entered into the study design.



Knowledge is always inherently intertwined with statistical analysis – there is the need to make these relationships visible and accessible.

Activities within M4

- Support the design of efficient biomarker validation studies
- Integrate knowledge management into the process of protocol development
- Use Biomax knowledge tools to develop projects
- Implement procedures to make knowledge implementation transparent and reproducible.
- Provide use cases do develop a good practice of knowledge implementation in complex biomarker discovery and validation studies.

Thanks for your attention