Machine Learning in Biomedicine:
From Basic Research to Clinical Translation

Lars Kaderali

Lübeck, November 11, 2019

Universitätsmedizin GREIFSWALD

Institute of Bioinformatics

Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

Before we start a few words about myself…

- Computer Science @ University of Cologne
- PhD on a high dimensional regression problem: Predicting Cancer Patient Survival from Gene Expression Data
- Subsequently worked at Los Alamos National Laboratory (USA), German Cancer Research Center dkfz, University of Heidelberg, University Hospital Dresden and now University Medicine Greifswald
- Professor for Bioinformatics at Medical Faculty
- Short stints in industry – CEL Corporation in Canada, ITK in Germany, Partner in ISL GmbH (IT Security Company), Consultant for Beiersdorf AG
- Worked on AI in Biomedicine since 2000, moving more and more into applications over the years

Universitätsmedizin GREIFSWALD

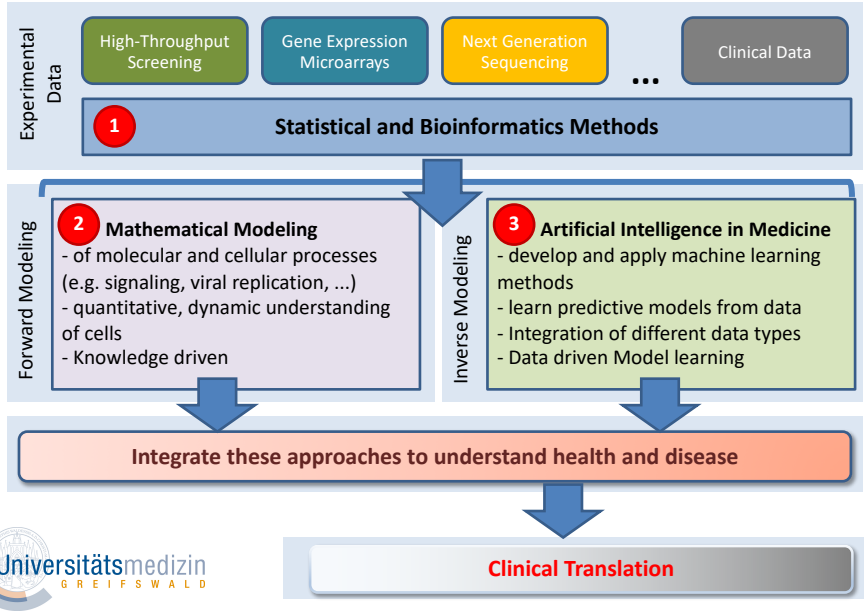
Institut für Bioinformatik

Institute of Bioinformatics

- Founded 2015/2016
- Part of the University Hospital / Medical Faculty
- Two Professors, Staff approx. 20 Persons (Postdocs, PhD Students, Bachelor/Master Students, Administrative Staff), mostly funded through third-party grants, mostly with a maths / computer science background
- Former Institute of Biometry and Medical Informatics
- Teaching: Biometry/Biostatistics, Bioinformatics
  - Medicine
  - Human Biology BSc & MSc
  - Biomathematics BSc

Institut für Bioinformatik

Research Overview



AI IN MEDICINE: HOPE OR HYPE?

Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

Top 10 AI Health Applications

APPLICATION	VALUE*
Robot-Assisted Surgery**	$40B
Virtual Nursing Assistants	$20B
Administrative Workflow Assistance	$18B
Fraud Detection	$17B
Dosage Error Reduction	$16B
Connected Machines	$14B
Clinical Trial Participant Identifier	$13B
Preliminary Diagnosis	$5B
Automated Image Diagnosis	$3B
Cybersecurity	$2B

TOTAL = ~$150B

Source: Accenture Analysis

Universitätsmedizin GREIFSWALD

Potential of AI in Medicine

- AI as support / tool for medical doctors
  - Clinical Decision Making / Decision Support Systems
  - Complementation of human expertise, e.g. in Radiology
  - Access to current information (Scientific Literature, Medical Practice Guidelines etc)
  - Support for younger, less experienced doctors
  - 24x7 Availability
- Early Diagnosis
- Prediction of Disease Outbreak and Progression, Prediction of Treatment Outcome
- Feedback for patients and doctors, e.g. concerning patient's compliance
- Reduction of diagnostic and therapeutic error
- Increased patient safety and cost savings through AI
- AI can extract information from huge patient numbers, more than any single doctor will ever see during his lifetime

Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

Successful Applications of Pattern Recognition in Medicine

- Detection of diabetic Retinopathy
  -> FDA approval for AI-based retina scan diagnostics
- Classification of Geneder using Retina Scans and AI:
  Human doctor 50%, AI 97% Acc
- Processing of Images from Radiology (MRI / CT Scans)
- Examples from Bioinformatics, e.g. Gene Signatures for Breast Cancer Prognosis
- Critical Examples: Prediction of Suicidality / Depression from Social Media Profiles

Universitätsmedizin GREIFSWALD

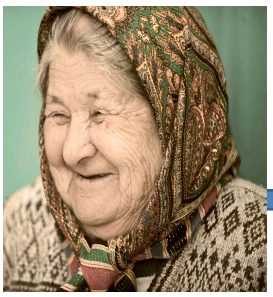
Institut für Bioinformatik

MOLECULAR MECHANISMS OF AGEING

Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

The Quest for Eternal Youth



Universitätsmedizin GREIFSWALD

Beiersdorf

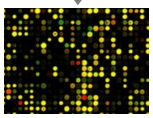
dkfz. GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION

Study Outline

Skin samples (epidermis from outer forearm) from 108 healthy, female, caucasian donors

24 Young (18-27 yr) | 60 Intermediate | 24 Old (61-78 yr)

Illumina Infinium Methylation Microarrays



Targets about 450,000 CpG sites

Gene Expression Profiles

Phenotypic Data

Deep Sequencing (RNA+ChIP)

Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

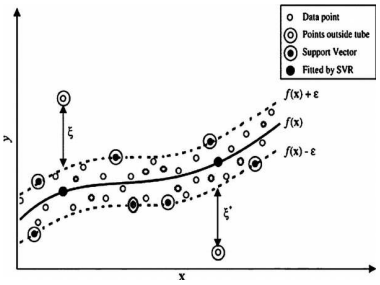
Prediction of "Biological Age"

- Methylation Data: 450.000 Features
- Chronological Age

- Linear SVM
- Train to predict age based on methylation profiles

- Apply to new participants
- Predicted Age is interpreted as "Biological Age"

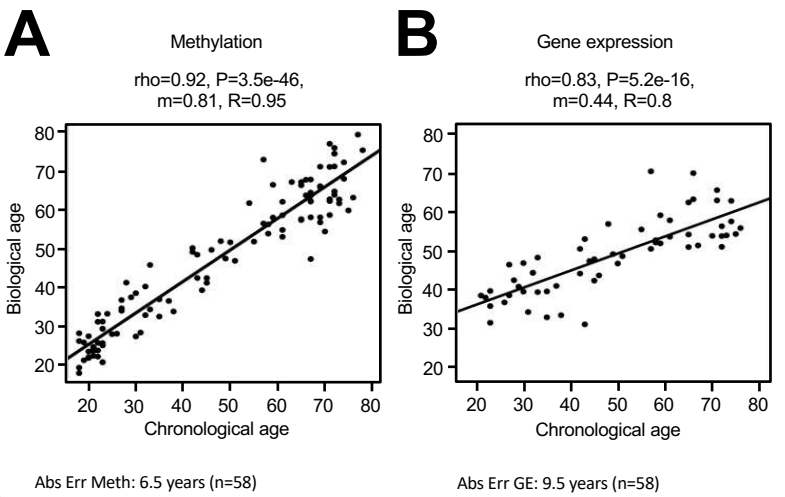
- Identify Outliers in Patients
- Identify Molecular Hallmarks of Ageing



Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

SVM Prediction of Biological Age



Abs Err Meth: 6.5 years (n=58)

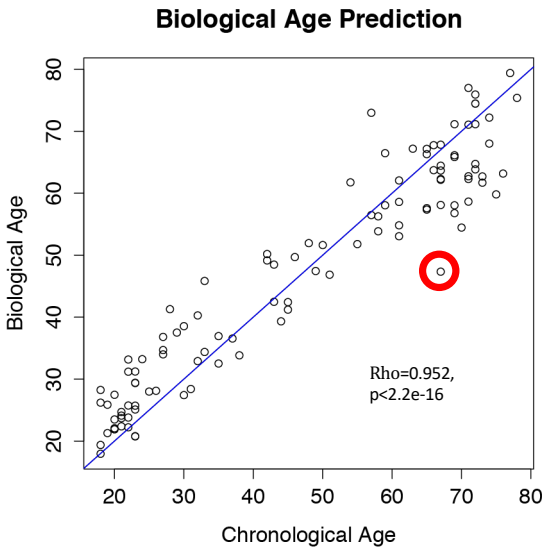
Abs Err GE: 9.5 years (n=58)

Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

Outliers in Ageing

- Of particular interest are individuals that have prematurely aged, or that have managed to "stay young"

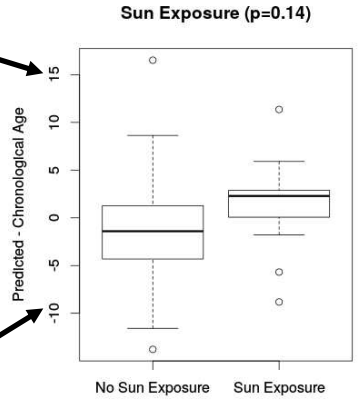


Universitätsmedizin GREIFSWALD

Premature Ageing and Phenotype Correlations

Prematurely Aged

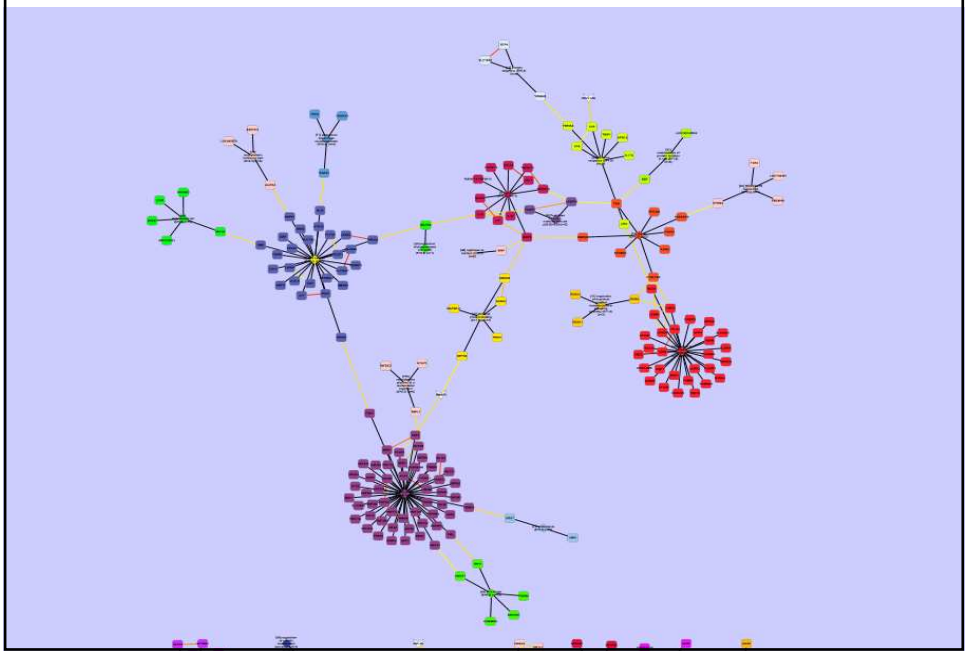
Staying Young…



Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

Multi-OMICs Signatures of „Oldies“ and „Youngsters“

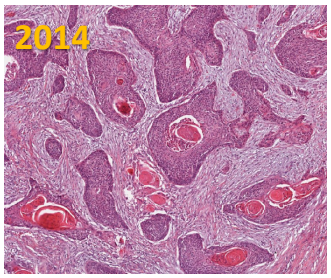


TREATMENT OF LUNG CANCER IN SMOKERS

Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

Treatment of Lung Cancer in Smokers

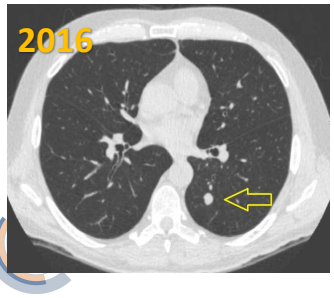


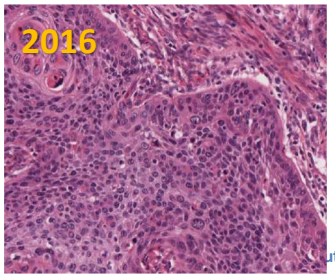
72 year old male, 50 packyears

2014 - Squamous cell carcinoma of the oral cavity

2016 - 1 cm solitary nodule in the lung with squamous cell carcinoma histology

1) Primary lung cancer ? -> T1aN0M0?
   -> curative surgery?
2) Metastasis? -> Stadium IV?
   -> palliative chemotherapy?

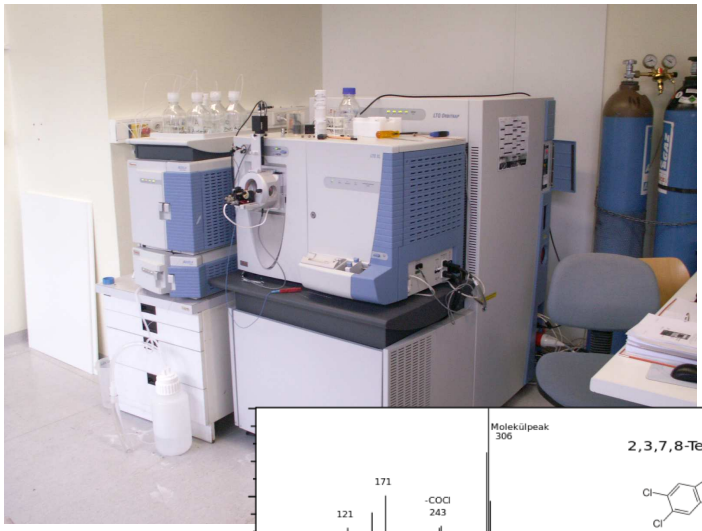


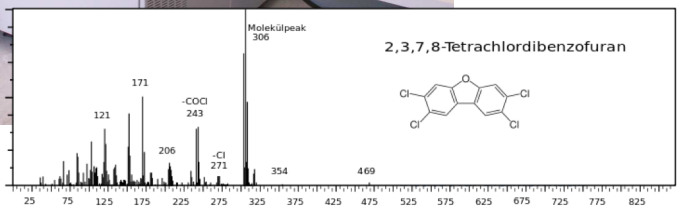


Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

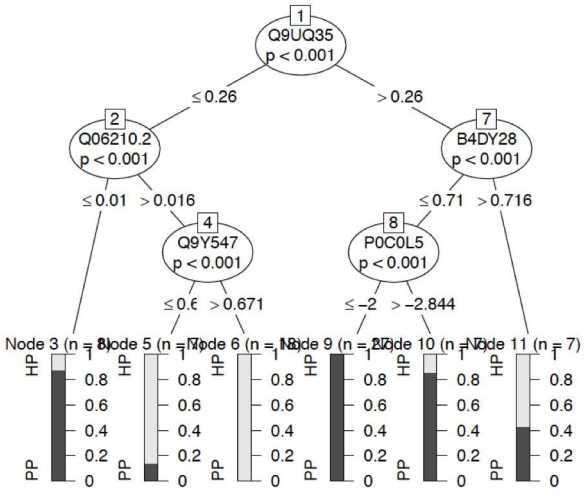
SILAC Protein Quantification





Universitätsmedizin GREIFSWALD

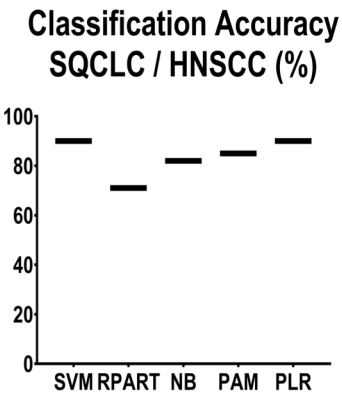
Classification of lung cancers

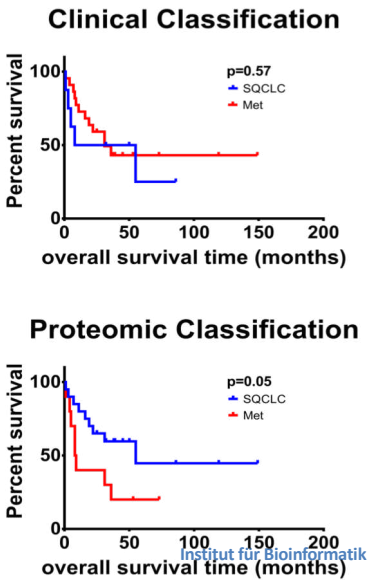


Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

Classification by Machine Learning





Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

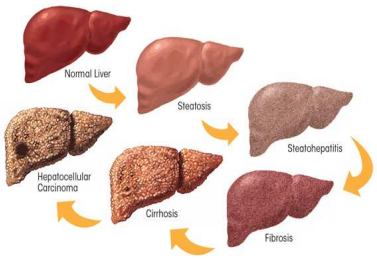
BAYESIAN NETWORKS AND RISK FACTORS FOR LIVER DISEASE

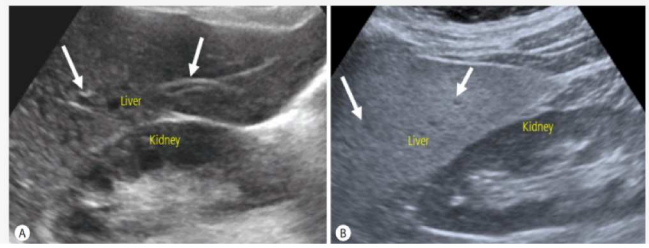
Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

Non-Alcoholic Fatty Liver Disease

- starts with **accumulation of fat** in the liver, progresses in 10-30% of the cases
- usually **no signs and symptoms** in early stages of NAFLD (sometimes fatigue or discomfort in the upper right abdomen)
- Leads to severe liver disease, cirrhosis & cancer
- Disease is underdiagnosed
- No noninvasive diagnostic method suitable for large-scale screening, diagnosis requires liver biopsy





Healthy Liver | Fatty Liver Disease

Universitätsmedizin GREIFSWALD

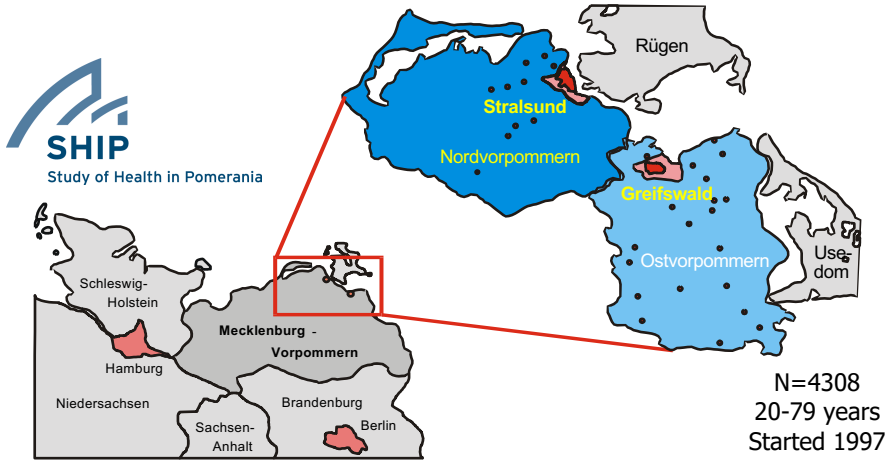
Institut für Bioinformatik

SHIP – the Study of Health in Pomerania

Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

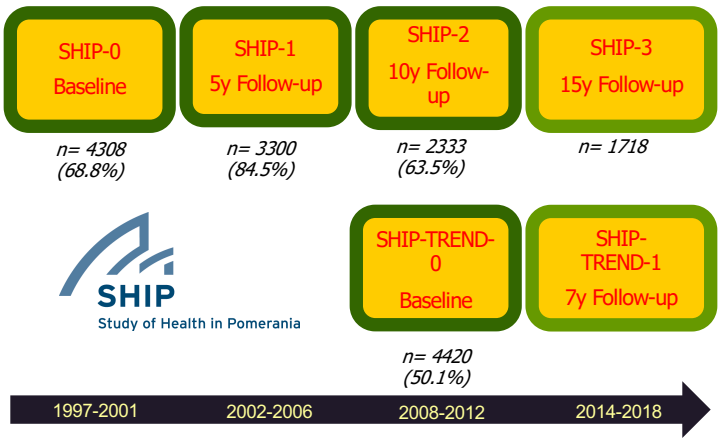
SHIP – the Study of Health in Pomerania



Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

SHIP – the Study of Health in Pomerania



Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

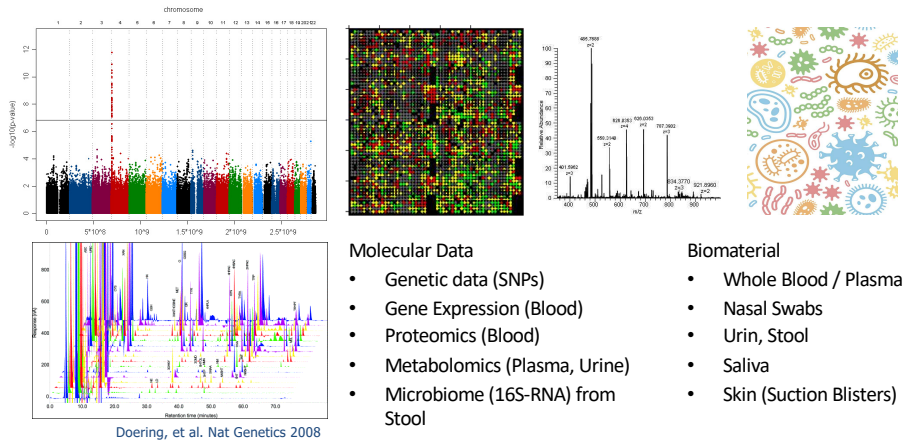
SHIP – the Study of Health in Pomerania

Völzke et al.; Int J Epidemiol 2011

Universitätsmedizin Greifswald

Institut für Bioinformatik

Molecular Data in SHIP-Trend / Biomaterial



Doering, et al. Nat Genetics 2008

Molecular Data
- Genetic data (SNPs)
- Gene Expression (Blood)
- Proteomics (Blood)
- Metabolomics (Plasma, Urine)
- Microbiome (16S-RNA) from Stool

Biomaterial
- Whole Blood / Plasma
- Nasal Swabs
- Urin, Stool
- Saliva
- Skin (Suction Blisters)

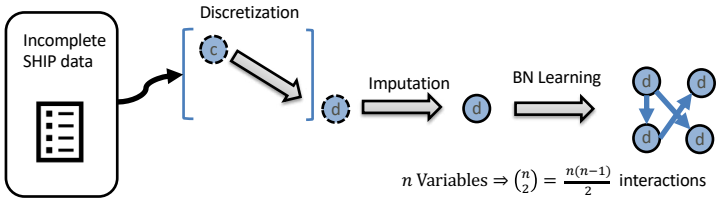
Universitätsmedizin Greifswald

Institut für Bioinformatik

Bayesian Network Idea

Idea: Learn a Bayesian Network from SHIP data to model all interactions of processes inside and outside of the liver, which possibly drive NAFLD

Steps of the Learning Process:

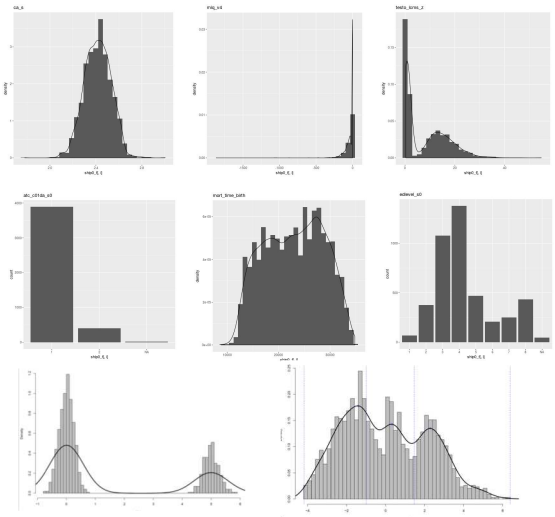


n Variables $\Rightarrow \binom{n}{2} = \frac{n(n-1)}{2}$ interactions

Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

What do our data look like?



- Large number of variables, small number of patients (p>>n)
- Mix of discrete and continuously distributed variables
- Variables at different scales
- Large number of missing values
- Highly correlated variables
- High levels of noise

Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

Inferring a Bayesian Network from Data

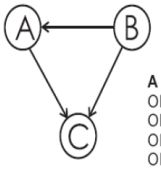
A	B	C
0	1	0
0	1	1
1	0	1
1	1	0

p(A|B)

B	ON	OFF
ON	0.2	0.8
OFF	0.01	0.99

p(B)

ON	OFF
0.2	0.8



P(C|A,B)

A	B	ON	OFF
ON	ON	0.0	1.0
ON	OFF	0.8	0.2
OFF	ON	0.9	0.1
OFF	OFF	0.99	0.01

- Two interdependent questions to solve: Find optimal model topology, and optimal model parameters
- Maximum Likelihood: max p(D|M) / MCMC Sampling from the likelihood
- Due to low number of samples and noisy data usually does not work
- Instead, work with the posterior distribution

$$p(M|D) = \frac{p(D|M)p(M)}{\int_M p(D|M)p(M)dM}$$

- P(M) can be used to include prior knowledge

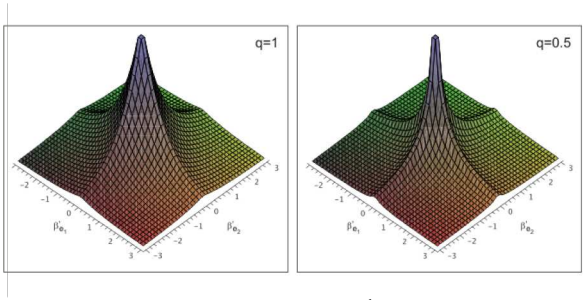
Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

Prior Distribution

Prior Distribution:

- If biological knowledge is available, can be included
- Simplest assumption is expectation of „sparse network“



$$p(M) = Ce^{-\frac{1}{qs^q}|w|^q}$$

Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

Automatic Adaptive Grid Refinement (Hydrodynamics)

= Method of adapting the accuracy of a solution within certain sensitive or turbulent regions of simulation, dynamically and during the time the solution is being calculated

Start computing a solution on a global grid

1. Apply an Error Estimator at every point of the grid
2. Flag points with high estimate
3. Create a finer grid at the parts of the grid where the estimate is high, so that every flagged point is contained in the finer grid
4. The subgrid is nested in the coarse grid

- The procedure is iteratively performed until all error estimates fall below a fixed threshold

Application to Bayesian Networks:

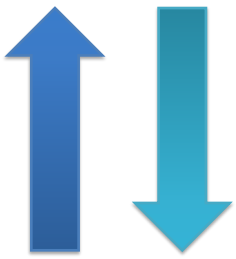
Start Learning a network between a small number of coarse clusters

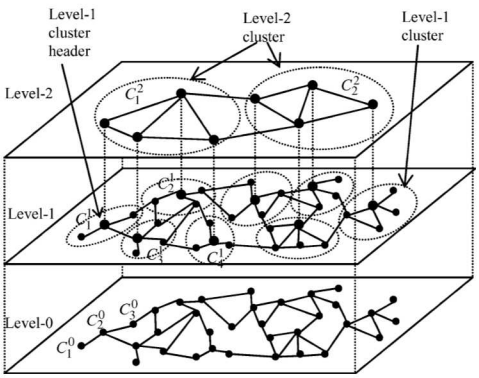
1. Error Estimation for every cluster and flag clusters with high errors
2. Split clusters with high error
3. Add new clusters to the data and delete old ones
4. Refine network around new clusters

Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

Hierarchical Learning of Large Bayesian Networks





Idea: Learn hierarchy from data (via hierarchical clustering)

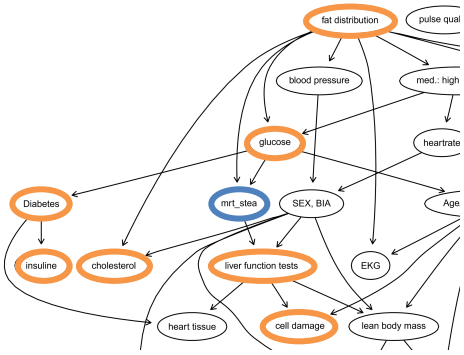
Learn networks between groups, not single variables

Refinement done taking relevant output variable (liver disease!) into account

Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

Results NAFLD



Fat distribution:
BMI
whtr
Koerperfett_in_prozent
etc.

Liver Function Tests:
Alat, Asat, GGT

Cell damage:
Lactatdehydrogenase
Creatinkinase

Cholesterol:
Chol_hdl (HDL-Quotient)
Triglycerides
LDL Cholesterol
Cholesterol (in total)
HDL Cholesterol

Next steps: Refine Methods, Include Molecular Data

Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

Further Examples from our Work

- Intraoperative classification of tissue type during brain cancer operations
- Prediction of Mortality using Deep Learning on SHIP data
- Prediction of Mortality based on Electronic Health Records
- Using neuronal networks to predict cardiovascular disease on ECG data
- Predicting therapy response and survival in cancer patients using penalized survival models
- Risk factors for thyroid disease
- Prediction of therapy response for prediabetic patients
- Prediction of depression based on social media data

Universitätsmedizin GREIFSWALD

Institut für Bioinformatik

Summer School “AI in Medicine“

- Zielgruppe: Naturwissenschaftliche Doktoranden und junge Ärzte
- 15 Mediziner + 15 Informatiker/Mathematiker
- 8. bis 12. Juni 2020 auf Hiddensee
- Dozenten:
  Tim Beissbarth, Unimedizin Göttingen
  Tim Friede, Unimedizin Göttingen
  Christoph Lippert, HPI Potsdam
  Mario Stanke, Greifswald
  Olaf Wolkenhauer, Rostock
  Harald Binder, Uniklinik Freiburg
  Nils Grabe, Uni Heidelberg
  Andreas Stahl, Greifswald
  Fabian Theis, Helmholtz München
  Lars Kaderali, Greifswald
- Mehr Infos unter
  http://www.kaderali.org

Universitätsmedizin GREIFSWALD

JOACHIM HERZ STIFTUNG

THANK YOU FOR YOUR ATTENTION

Universitätsmedizin GREIFSWALD

Institut für Bioinformatik