Hybrid models – combining mechanistic and statistical modeling approaches

Prof. Dr. Andreas Schuppert
RWTH Aachen University
Aachen, Germany
Content

- Why is both mechanistic modelling & machine learning not effective in biomedicine
- Hybrid modeling: Integrating knowledge & machine learning
- Hybrid modeling in medicine – an example
- Hybrid modeling – a tool for real world evidence?
- Summary
Challenge of Precision Medicine: complexity & heterogeneous data sources

From –omics to drug action: the most complex problem in science and engineering?

Confounders on clinical level are crucial for therapeutic success, but can hardly be identified and analyzed in lab

Precision medicine requires the integration of data & information from heterogeneous sources and system scales in models
Is the information hidden in Big Data? - the machine learning promise...

Der Spiegel, 8.11.2017
Is the information hidden in Big Data? – …and the first caveats

MD Anderson Benches IBM Watson In Setback For Artificial Intelligence In Medicine

Matthew Herper, FORBES STAFF
I cover science and medicine, and believe this is biology's century. FULL BIO

Watson
IBM's Supercomputer stellt sich dumm an

Supercomputer Watson sollte der klügste Computer der Welt werden und dem Unternehmen unter Chefin Virginia Rometty Milliarden bringen. Doch der Rechner ist begriffsstutzig. Von Matthias Kamp, Michael Kroker und Sven Prange. Mehr...

Wirtschaftswoche, 18.9.2017
Modeling in Precision Medicine: the threat of the curse of dimensionality in machine learning

Accuracy of patient classification

\[ n_{\text{data}}: \# \text{ of diagnostic parameters in ML - model} \]

the dream...

...the reality

All patients get “equally different”: “curse of dimensionality”

Missing hidden parameters
Modeling in precision medicine: mechanistic modeling is hampered by lacking knowledge of mechanisms

**Primary action:**
Mechanistic model available

**Secondary response:**
Lack in mechanistic knowledge
Modeling in Precision Medicine: between Scylla and Charybdis

- Modeling disease using mechanistic models
  - Requires full quantitative representation of all biological processes associated with the disease
  - Enables extrapolative simulations
  - Supports systems understanding

- Modeling outcome by machine learning from “big data” without mechanistic knowledge
  - Requires availability of “big data”, not “fat data”
  - Data must (implicitly) represent all relevant information
  - Curse of dimensionality
    - Adding more data improves coverage of information, but decreases performance of characterization
  - No extrapolation from training data set possible
Modeling in Precision Medicine: between Scylla and Charybdis

Lack of understanding

Curse of dimensionality

https://i.ytimg.com/vi/DaMs7dn6EJQ/hqdefault.jpg
Content

- Why is both mechanistic modelling & machine learning not effective in biomedicine
- Hybrid modeling: Integrating knowledge & machine learning
- Hybrid modeling in medicine – an example
- Hybrid modeling – a tool for real world evidence?
- Summary
Hybrid modeling – integrating knowledge and data

- Rationale:
  - Machine learning conceptually
    - neglects a priori knowledge
    - learns model from data by smart interpolation
  - all investments into (at least partial) understanding of the mechanisms driving the systems behavior are lost

**Hybrid modeling**: integrate the available knowledge into the machine learning process, such that the weaknesses of machine learning are cured by the integrated knowledge
Hybrid modeling – realization by functional networks

black box model → functional network model:
- subsystems ↔ nodes
- interactions between the subsystems ↔ edges

use mechanistic models where established

Approximation by e.g. (flat) NN
Hybrid modeling – driven by science & applications

Three communities with relation to hybrid modelling:

→ Chemical Engineering (coupling black box models for subprocesses (or residues) with mechanistic models):

→ Mathematics: can networks of nested functions represent all functions (Hilbert’s 13th problem)? Identifiability & extrapolation problem:

→ Deep Learning: Curse of dimensionality, learning of structures
  » DOI: 10.1007/s11633-017-1054-2
  » Henry W. Lin, Max Tegmark, and David Rolnick, Why does deep and cheap learning work so well? arXiv:1608.08225v3
Mathematical foundations: functional networks & black box


The following functional network is a UMM for $\mathbb{R}^n \Rightarrow \mathbb{R}^1$:

This is not true if $u_k \in C^1(\mathbb{R}^1 \Rightarrow \mathbb{R}^1)$

Mathematical foundations: the benefit of structural knowledge

The model can be (locally) identified from data \((x, H(x)), x \in \mathbb{R}^n\), if the data guarantee the well-posedness of the initial value problem of the structure-constraint equations on a low-dimensional sub-manifold \(D\) of \(\mathbb{R}^n\).
Mathematical foundations: the benefit of structural knowledge

- The data demand for training of a functional network is controlled by the #of inputs $n_{\text{max}} \ll n_{\text{data}}$ of the most complex black-box node in the network.

- Functional networks can be identified from correlated data around a $n_{\text{max}}$-dimensional surface in the $n_{\text{data}}$ dimensional data space (data base).

- Functional networks allow extrapolative simulations outside the data base.

Routes towards integrating knowledge and data

- **Bayesian learning**
  - Use standard machine learning tool (like Deep learning)
  - Implement a priori knowledge as priors for parameters into the machine learning process
  - Choose the bandwidth of the priors according to the reliability of the knowledge

- **Penalization**
  - Use standard machine learning tool (like Deep learning)
  - Implement a priori knowledge as penalty-terms into the training process
  - Choose the weights of the penalizations according to the reliability of the knowledge
Routes towards integrating knowledge and data

■ **Pro:**
  - Utilization of standardized ML-tools
  - Utilization of established training algorithms
  - High flexibility

■ **Con:**
  - Implementation of structural knowledge requires large-scale penalization / definition of priors
  - Performance of training algorithms may suffer from large-scale a priori constraints
  - Implementation of available mechanistic models not possible within existing software architectures

■ Solution in case of massive a priori knowledge: explicit implementation of functional networks, e.g. in KNIME
Why is both mechanistic modelling & machine learning not effective in biomedicine

Hybrid modeling: Integrating knowledge & machine learning

Hybrid modeling in medicine – an example

Hybrid modeling – a tool for real world evidence?

Summary
Example: modeling ADR of Statins

Challenge: Prediction of ADR, e.g. rhabdomyolysis (strongly increased CK - level) caused by statin therapy

Data:
- > 3000 SNP’s on 86 genes
- clinical parameters
- 312 patients

Goal: Identify

\[ [\text{CK}] = F(\text{genotype, ...}) \]

Problem:
standard approaches failed, because of huge network complexity
Why hybrid modeling? multi-scale modeling is essential

Impact of genotype on disease

Disease status

healthy
disease

Status of physiological / environmental marker
Example: modeling ADR of Statins

unsupervised patient stratification

physiology-based prediction model

to predict phenotype

integration of genotype data

CK₀, sGPT/sGPC, BMI

Joint Research Center for Computational Biomedicine
Integrating heterogeneous data sets & mechanisms under uncertainty

Example: modeling ADR of Statins

Heterogeneous set of impact parameters

SNP

Gene 1

0 1 0 1

Gene 2

1 1 0 1

Gene activity

0 1

0 1

Disease
Physiology
Nutrition

readout

Prediction
Example: modeling ADR of Statins

**Challenge:** model cooperative effects between genes *without a metric*

Gene 1

\[
\begin{array}{ccc}
0 & 1 & 0 \\
1 & 0 & 1 \\
0 & 1 & 0 \\
\end{array}
\]

\[
\{1, \ldots, 2^n\} \rightarrow \{1, 2\} \rightarrow 0 \rightarrow 0 \text{ functional status of gene pair}
\]

Gene 2

\[
\begin{array}{ccc}
1 & 1 & 0 \\
0 & 0 & 0 \\
\end{array}
\]

\[
\{1, \ldots, 2^m\} \rightarrow \{1, 2\} \rightarrow 0 \rightarrow 1 \text{ functional status of gene pair}
\]

2x2-table
Example: modeling ADR of Statins

<table>
<thead>
<tr>
<th></th>
<th>standard biomarker</th>
<th>hybrid biomarker</th>
</tr>
</thead>
<tbody>
<tr>
<td>genotype only</td>
<td>70%</td>
<td>97.7%</td>
</tr>
<tr>
<td>and black box</td>
<td></td>
<td></td>
</tr>
<tr>
<td>clinical parameters</td>
<td>70%</td>
<td></td>
</tr>
<tr>
<td>and genotype</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Example:
Adverse drug reaction of cerivastatin
Content

- Why is both mechanistic modelling & machine learning not effective in biomedicine
- Hybrid modeling: Integrating knowledge & machine learning
- Hybrid modeling in medicine – an example
- Hybrid modeling – a tool for real world evidence?
- Summary
Biomedicine in the data-rich world: massive data enable a new tier of quality

- Single input, multiple output
- Targeted system perturbation

What we observe for all downstream genes / proteins if we stimulate the cells

What the systems mechanism may look like
Real World Evidence: utilization of massive unbiased data

- Real World Evidence – towards precision medicine
  - Clinical trials are designed to answer specific questions
  - Hence they may miss essential confounders
  - Complement trials by massive data analysis

Real World Evidence: A New Era for Health Care Innovation

RWE is not just “Big Data” – it’s the integration of multiple sources of data

RWE promises to transform patient outcomes, but it also threatens to upend long-established norms in the generation and use of health care evidence. RWE challenges the traditional paradigm in which the only authoritative medical evidence is generated through prospective randomized clinical trials (RCTs), validated through peer reviewed publication in reputable journals and incorporated into broadly applied clinical practice guidelines.
Disease modeling: hierarchical networks & observability

- Common mechanisms result in apparent co-regulation network on the data level:
  - Data on “clinical level” are available as “Big Data”, but do not provide direct insight into underlying mechanisms
  - Analysis of the correlation structure may reveal the respective common mechanisms and their interaction
Reengineering of mechanisms: the inverse problem

- **Algorithm 1**: identify all triplets of input variable \( \{x_i, x_j, x_k\} \), s.t.
  \[
  \partial_x^2 y = 0 \quad \text{(structure-constraint equations)}
  \]
  then the tree structure can be identified \( O(n^3) \) algorithm, but requires designed input data combinations
  → Problem: algorithm is ill-posed, not feasible in categorical, real life data

- **Algorithm 2**: identify the set of all correlations in high throughput experiment data and reconstruct the network
  → can be used for categorical data
  → Robust
  → NP – problem (max. \( O(2^n) \))
  → Must be adapted to continuous data (efficient algorithm open problem)

- **Algorithm 3**: identify the set of all (generalized) correlations by means of intrinsic noise (manuscript in preparation)
Reengineering of mechanisms: the inverse problem

Example: Algorithm 2
Content

- Why is both mechanistic modelling & machine learning not effective in biomedicine
- Hybrid modeling: Integrating knowledge & machine learning
- Hybrid modeling in medicine – an example
- Hybrid modeling – a tool for real world evidence?
- Summary
Program: Functional networks & next generation deep learning

- Functional (hybrid) network theory guarantees

  → When to avoid curse of dimensionality (in FF-networks)
  → Dimensionality of data base (rigorous sparsity tests)
  → Extrapolation from sparse data distributions
  → Ill-posedness
  → Structural identifiability
  → Provides additional flexibility to integrate non-standard models

- Utilize the rigorous results for generic functional networks to develop next generation (bayesian) deep learning
- Combine with generic state-space descriptors to identify critical state
JRC for Computational Biomedicine @ RWTH Aachen