Causal discovery to reveal complex gene regulation networks from vast transcriptome data sets (SILS)

Intended promotor: *prof. dr. Leendert Hamoen (SILS); dr. Joris Mooij (IvI)*
Intended co-promotors: *dr. Martijs Jonker (SILS); lab manager*

Background:

The life sciences Omics revolution is creating vast amounts of biological data. The key challenge is to find new methods to analyze these data. Artificial intelligence, in particular machine learning, has been hailed as the solution to find patterns in these continuously and rapidly expanding data sets. However, despite its promises, progress in in this field has so far been limited. One of the main reasons for this is a lack of collaborative projects partially due to the fact that AI students are not trained in biology, and that AI-focused projects do not have the resources to experimentally test the outcomes of their data analysis. AI4Science provides a unique opportunity to tackle these shortcomings and to train AI students with additional biology and bioinformatics skills that will perfectly position them for industry and research positions related to the life sciences.

Transcriptome analysis is one of the most powerful deep-sequencing techniques and measures the expression of all the genes of an organism under certain conditions. The number of genes measured varies between a few thousand for bacteria to tens of thousands for more complex cells, including human cells. These transcriptome datasets are collected at the GEO repository curated by the NCBI. This massive and quickly growing amount of transcriptome data is a treasure trove awaiting automated scientific analysis to reveal unknown complex regulatory networks and other biological important information that is necessary to fully understand a living cell. Currently, even simple scientific questions such as “what causes the expression of this particular gene?” are still very challenging to answer in an automated, data-driven fashion. There is a huge potential for developing and improving modern machine learning methods to analyze these large omics data sets.

In this project, we will develop AI based tools, in particular causal discovery methods, to infer gene regulation networks from large transcriptome datasets. We will use a relative simple model system, the bacterium *Bacillus subtilis*, which has a genome comprising approximately 4100 genes. This is one of the best studied bacterial model systems and there are now about 12300 different *B. subtilis* transcriptome profiles in the GEO collection (each containing information on ~4100 genes), and this number is growing rapidly. However, so far, there has not been any AI-driven study to explore this data collection. This is a real misfortune since we are still in the dark when it comes to the regulation and function of many of the genes of *B. subtilis*, despite the fact that this organism has been studied in great detail. The *B. subtilis* GEO collection provides a perfect data set to explore the possibility to use machine learning, and more in particular, causal discovery, to identify regulation pathways and other unknown biological traits. The advantage of using *B. subtilis* is that many regulation pathways have been experimentally determined and therefore biologically relevant control sets are available. In addition, the Hamoen lab is an expert in *B. subtilis* gene regulation and deep-sequencing tools, including RNA-seq for transcriptome analysis, and will be able to experimentally verify novel causal relations emerging from AI analyses. Importantly, successful AI analysis methods can be applied to other bacteria, including notorious pathogens and those residing in microbiomes. The methods developed in this study are at least partially also applicable for the analysis of large omics datasets from eukaryotic microorganisms and more complex eukaryotic cells, including those of humans.

Aims:

The scientific aim is to further develop modern machine learning methods for causal discovery to enable reconstruction of gene regulatory networks from the wealth of available transcriptome data, and to validate the findings experimentally. The unique training aim is to teach the AI student simple basics of biology and bioinformatics. In summary the following **objectives** are envisioned:

1. Teach PhD student the fundamentals of molecular biology and bioinformatics.
2. Process *B. subtilis* GEO transcriptome profiles for automated analyses.
3. Apply recently developed causal discovery methods to infer gene regulation pathways from the processed GEO transcriptome database and compare results and with simple baselines.
4. Improve causal discovery results by systematically implementing biological knowledge and optimizing statistical and computational properties of the methods to match the domain of interest.
5. Experimentally assess (e.g. RNA-seq) whether regulation pathways discovered in Objective 4 are biologically true.
6. Combine all ingredients to perform active learning of the causal structure of a smaller (reasonably isolated) part of the gene regulatory network, by iterating: (i) analysis of available data; (ii) proposal of new experiments; (iii) performing new experiments in wet lab; (iv) extend available data with new measurements.
7. Assess the general applicability of the best performing developed causal discovery methods on transcriptome databases from other bacteria.

Approach:

We propose to build on recent causal discovery methods that have already been used on yeast transcriptome data [M++16]. In particular, we plan to further develop versions of the Joint Causal Inference approach [MMC18], adapting them to the type of data at hand. JCI is the first causal discovery approach that allows combining different data sets, can handle cyclic causal relations, under partial observability - exactly what is needed to make sense of this type of transcriptomics data. However, a remaining challenge that needs to be addressed is how to perform the necessary computations in a computationally feasible and statistically reliable way. Regulation pathway structures emanating from these analysis will be first verified against known *B. subtilis* pathways, and novel pathways emerging from this analysis will be experimentally tested by the Hamoen lab using RNA-seq and different mutants and/or growth conditions.

Feasibility: In the spring of 2018, we, L. Hamoen, M. Jonkers and J. Mooij, already initiated a collaboration on this topic by jointly supervising a group of B.Sc. AI students for the Project AI course. That project was aimed at making the *B. subtilis* GEO database available for in depth analyses, including machine learning approaches. This already yielded some encouraging first results and we are now keen to continue this collaboration. The AI4science initiative would provide the ideal support to push this collaboration forward.

Impact:

AI-based approaches can revolutionize the way that transcriptome data is analyzed, resulting in a better understanding of the workings of complex mechanisms in cells. Importantly, when algorithms have been developed on the *B. subtilis* dataset, they should in theory be able to efficiently tackle the transcriptome dataset of the many other organisms that are available, therefore providing an extremely powerful analytic tool for the microbial research community. The added value for the AI PhD student is the added biology and bioinformatics training, which will make her/him a very strong candidate for data driven life science positions. Finally, the added value for the AI group is the frequent feedback from a molecular biologist and bioinformaticist and experimental verification of results, assuring relevance and impact.

References:

[MMC18] J.M. Mooij, S. Magliacane & T. Claassen (2018): “Joint Causal Inference from Multiple Contexts”. arxiv.org preprint 1611.10351v3

[M++16] N. Meinshausen, A. Hauser, J. M. Mooij, J. Peters, P. Versteeg, P. Bühlmann (2016): “Methods for causal inference from gene perturbation experiments and validation”. Proceedings of the National Academy of Sciences of the USA 113:27 pp. 7361-7368