Institut für Informatik
Praktische Informatik und Bioinformatik
Prof. Dr. Ralf Zimmer
Dr. Evi Berchtold, Markus Joppich, Felix Offensperger, Jewgenija Zhdanovich

Corona/Online Winter Term 2020/21
Computational Systems Biology

Assignments 2020-3
Applications of dynamical systems models for detailed quantitative predictions and Systems models for interpreting high-throughput data.

Working period: Three/Four weeks (22.12.-19.1.2021)
Hand-in anytime or in any exercise class
Please hand-in only reproducible results, answers, figures, tables, simulations, ...

Projects 1-3 apply dynamical systems models to make or validate precise quantitative biological predictions.

Many papers use systems models to support biological hypotheses and provide additional evidence. Often only a sketch (from a textbook or paper is given). Then the system has to be derived from the sketch. Parameters need to be specified, or learned from some data, or extracted from appropriate databases (maybe different versions for diverse contexts), or modifications need to be made to ready-made available systems models.

The solution of such models (and the visualisation of the results) depends on 1.) the specification of the model (i.e. are all rate constants and dependencies specified such that a (system of) ODE(s) can be fully stated and solved via integration, or is a partial model given such that only a crude partial simulation is possible) 2.) modelling choices (is a continuous time-resolved behaviour wanted, or the states in predefined discrete steps, the final states in a number of discrete conditions, the final outcome after convergence, the steady-state-behaviour, attractors, oscillations, ...) 3.) the methodological framework: ODE solvers, numerical solutions, stochastic simulation, Petri net semantics.

Simulations and solutions of dynamical systems models produce visualisations to provide additional evidence for biological hypotheses. But major problems are uncertain parameters, robustness of models and results, reproducibility of outcomes, and missing factors and incomplete models.

There are many databases around which collect prepared models in various formats (e.g. BioModels, https://www.ebi.ac.uk/biomodels). There are also many tools, which allow to specify reaction systems, both general (e.g. matlab, octave) or special (e.g. Gepasi, Copasi, CellDesigner, ...) modelling tools.

Applications of dynamical systems

Task 1 Circadian clocks and cancer

Recently, several papers and reviews have been published to propose a close connection between key circadian clock genes (such as CLK:BMAL1, Per2) and key cancer genes (e.g. p53 and MDM2). Both, circadian rhythms and cancer checkpoint regulation are well-studied processes with a wide variety of models and experimental evidence. Surprisingly, there is also mounting evidence close interconnection of the two processes via key well-known molecular factors.

This project reviews the models and the interconnections, resolves the chain of experimental data and facts and theoretical models, and tries to reproduce certain hypotheses and claims.

(a) Start with the recent review [Zou Xianlin, Kim Dae Wook, Gotoh Tetsuya, Liu Jingjing, Kim Jae Kyoung, Finkielstein Carla V., A Systems Biology Approach Identifies Hidden Regulatory Connections Between the Circadian and Cell-Cycle Checkpoints Frontiers in Physiology, 11, 327, 16.4.2020, https://doi.org/10.3389/fphys.2020.00327) and summarize the facts.

(b) Several network and dynamical models are proposed or reviewed in the paper. Discuss the models and whether there are conflicting hypotheses.

(c) The authors of the review have published several papers about the topic

- Selfridge JM, Gotoh T, Schiffhauer S, et al. Chronotherapy: Intuitive, Sound, Founded ... But Not Broadly Applied. Drugs. 2016;76(16):1507-1521. doi:10.1007/s40265-016-0646-4
- Liu J, Zou X, Gotoh T, et al. Distinct control of PERIOD2 degradation and circadian rhythms by the oncoprotein and ubiquitin ligase MDM2. Sci Signal. 2018;11(556):eaau0715. Published 2018 Nov 13. doi:10.1126/scisignal.aau0715
- Gotoh T, Vila-Caballer M, Liu J, Schiffhauer S, Finkielstein CV. Association of the circadian factor Period 2 to p53 influences p53's function in DNA-damage signaling. Mol Biol Cell. 2015;26(2):359-372. doi:10.1091/mbc.E14-05-0994
- Gotoh T, Vila-Caballer M, Santos CS, Liu J, Yang J, Finkielstein CV. The circadian factor Period 2 modulates p53 stability and transcriptional activity in unstressed cells. Mol Biol Cell. 2014;25(19):3081-3093. doi:10.1091/mbc.E14-05-0993

Review the progression of findings and models.

(d) The paper reviews the findings concerning the interconnection of circadian rhythms and cancer. Sketch the historical findings and facts about this connection from the published literature in addition to the publications from the Finkielstein lab. Who and which paper claims of ahs the priority for certain facts. Have the facts been reproduced or falsified later on?

(e) A recent manuscript ([C Wang, H Liu, Z Miao and J Zhou, Circadian rhythm regulated by tumor suppressor p53 and time delay in unstressed cells] see internal website) proposes a (new?) time-delayed model about circadian and cancer processes and the relation between Per2 and p53 in particular. What is the new contribution of this paper?

(f) What is known about the pathways in pathway databases such as BioModels, KEGG, or ConsensusPathDB? Anything about the connection of the two processes?

(g) To obtain complete information about the facts and relations of participating genes and proteins

in the pathways a systematic text mining could be considered. This involves prescreening of relevant publications and, then, a systematic information extraction from the papers.

Task 2 Parameter estimation of dynamical models

Typically, models obtained from the literature are not fully parametrized. In addition the actual parameters might depend on the respective condition and cellular context.

The paper [Fabian Fröhlich, Thomas Kessler, Daniel Weindl, Alexey Shadrin, Leonard Schmiester, Hendrik Hache, Artur Muradyan, Moritz Schütte, Ji-Hyun Lim, Matthias Heinig, Fabian J. Theis, Hans Lehrach, Christoph Wierling, Bodo Lange, Jan Hasenauer, Efficient Parameter Estimation Enables the Prediction of Drug Response Using a Mechanistic Pan-Cancer Pathway Model, Cell Systems, Volume 7, Issue 6, 2018, Pages 567-579.e6.] claims to obtain a reasonable parametrisation of large-scale mechanistic models via a highly efficient computational method. Explain the setup and the used method. Why does it achieve and guarantee a 10.000-fold speedup? For what purpose can the resulting model be used? What are the predicted outcomes and what are the evidences for their validity.

A recent manuscript ([Robin Schmucker, Gabriele Farina, James Faeder, Fabian Fröhlich, Ali Sinan Saglam, and Tuomas Sandholm, Multi-Drug Treatment Optimization Using a Pan-Cancer Pathway Model, 2020], available on internal web site) employs the method to predict drug responses for cancer treatment including combination therapies. Discuss the approach and the results.

Task 3 (from Projects 2): SBML, BioModels

SBML (https://sbml.org) is "a free and open interchange format for computer models of biological processes. SBML is useful for models of metabolism, cell signaling, and more" and constitutes a whole toolset of libraries, tools, editors, simulators, etc. including APIs and language bindings for several programming environments including python.

The BioModels resource (https://www.ebi.ac.uk/biomodels/) at EMBL-EBI provides thousands of curated and non-curated models, mainly based on ODEs from a variety of research, papers, and pathways. BioModels aims to provide a comprehensive collection of "existing literature-based physiologically and pharmaceutically relevant mechanistic models in standard formats. Our mission is to provide the systems modelling community with reproducible, high-quality, freely-accessible models published in the scientific literature."

(a) Install SBML and make use of SBML tools and the access to BioModels.

(b) make a statistic about the models and pathways! Which pathway is covered by the most models? Which model touches on the most pathways.

(c) Checkout the key proteins from Task 1. Are there models containing these proteins? Can the models contribute to the hypotheses and claims of Task 1?

Comprehensive systems models

The following projects try to build as comprehensive and as detailed as possible models of specific cellular systems for further use in model analysis (FBA, invariants) and interpretation of high throughput transcriptomic and proteomic data.

The goal is to produce useful network models for the respective purposes, but also to investigate methods to obtain such models from databases and literature (beyond manual collection and curation). This involves information extraction methods and data structures to represent such models.

For various purposes different types of information and networks need to be represented: (a) metabolic networks (enzymes transforming sets of substrates into sets of products) (b) regulatory networks (TF - target gene interactions) (c) signalling cascades (signal transmission via protein interactions or protein modifications (e.g. phosphorylations). In addition to these three major types, there are also specialised networks such as miRNA-target interactions, protein complexes, phosphorylations, epigenetic modifications, binding sites, splicing, ...

The most simple type of network are interaction networks, where interactions represent some connection between the two objects (typically of unknown function and unknown interaction type not to speak of dynamics). A very common rudimentary type are sets of objects somehow belonging to a certain class of genes/proteins, which probably belong to or contribute in some way to a common function or pathway (e.g. gene ontology (GO) sets and similar gene set definitions and databases). The common use of these 'ontology' or 'functional' sets (Molecular Function (MF), Biological Process (BP), Cellular Component (CC)) is to compute for expression data some enriched sets with the idea that the represented process is somehow regulated in the investigated system (over-representation analysis (ORA), gene set enrichment analysis (GSEA)). The identified sets should be more meaningful and interpretable than long lists of individual objects and, at the same time, more robust against fluctuations and noise in the data and sensitivities of the data analysis.

For this purpose a number of functional sets and ontologies have been defined and collected in databases. This includes the Gene Ontology (GO) with its three subdivisions ML/BP/CC, pathway databases such as KEGG, Reactome, WikiPathways, and also collections of ontology-based or experimentally derived sets of molecular/gene/protein/disease biomarkers database (such as MsigDB, GenSigDB, ...) which are indicative for certain cell-types, cellular states, diseases ...

Typically set and pathways definitions are not immediately appropriate for the problem and data at hand. Moreover the definitions are not context-specific, i.e. not tailored for a specific cell types, cell state, tissue or disease type. E.g. only rarely it is known or annotated that certain interactions or regulations occur only in specific contexts but not in others. Of course any knowledge on this context-specificity is essential for the use of systems models to explain experimental data or even for doing straight-forward validation checks with the use of the systems models. This applies even more for pathway data and hampers the usefulness of pathway and network enrichment methods (e.g. gene graph enrichment analysis (GGEA)).

One particular important context is the model systems, i.e. it is not clear whether and to which extent data and models from animals or cell-lines or sorted cells or single-cells can be transferred to the real

human situation, say in patients.

Often different data types (e.g. sequencing data (expression, translation, regulation), mass spec data, binding data, epigenomics data) need to be integrated and be interpreted with the help of systems models and networks and functional classes.

Often different stages of a system (developmental stages, disease progression, time series after treatment, infection, perturbartion...).

Moreover, different systems have been measured, e.g. 'bulk' data from tissues, data from sorted cells (ideally just one cell type or one cell type in one particular state), single cell data, ...

Sometimes spatial information is available which can be combined with microscopy and physiological data.

Very often, experiments are performed on model systems as experimentation with human cells or patients is not possible. Therefore, genetically modified and knock-out animals are used for data generation. Large-scale experiments are mostly done on cell-lines, which are more homogeneous and much easier and cheaper to work with.

All of these issues could be addressed, but should certainly be kept in mind in collecting systems model data on various systems. In the following, projects will review the resources for the respective systems, collect network data from the available resources, curate the derived data, and prepare 'systems models' for the interpretation of experimental high-throughput data and the visualization of results.

Task 4 Yeast

This project develops systems models for yeast. Yeast is probably the best studied model system of all. It is a fungus and a quite simple eukaryotic single-celled microorganism, which can easily handled and modified in the lab. Moreover many cellular and regulation mechanisms can be studied in yeast in its quite basic form. In particular thousands of KO and double KO strains of yeast are available for experimentation (in fact all single and all (36 Mio) double knock outs!). tries to understand yeast heat shock as a very typical and well studied stress condition.

The project tries to understand yeast heat shock as a very typical and well studied stress condition. Heat shock employs main stress response mechanisms, but maybe also heat specific reactions. The project investigates general and specific molecular stress response mechanisms based on various experimental measurements in particular the effects and impacts of the knockout of important transcription factors (HSF1, MSN1, and MSN4). So models should focus on transcriptional regulation and these TFs in particular.

Different data types on the system have been measured, i.e. transcriptomic, translatomic, proteomic, and regulatory data.

Many and quite comprehensive databases on interaction and regulatory relations are available (Yeastract, ISMARA, BioGrid, ConsensusPathDB).

Task 5 Human cell lines As an example of a human cell line, this project investigates the human colon epithelial carcinoma cell line Caco-2 for the interpretation of COVID-19 / Sars-CoV-2 proteomics data. Here general information and systems models should be derived, but maybe also distinguished context-specific information on the Caco-2 cell line in contrast to other human cell lines.

Task 6 Human cancer TCGA and CPTAC are the major databases and resources for sequencing and proteomics data, respectively, for a number (about 30) of different human cancer types. Typically for a larger number of patients normal and diseased tissue has been measured with high-throughput sequencing and mass-spec.

Task 7 Human Macrophages, monocytes, neutrophils For the interpretation of SFB1123 data on atherosclerosis and coronary artery diseases (CAD) systems models and information on artery plaque and the involved cell-types are required. The project focusses on one of the most prominent cell types, i.e. macrophages which play an important rule in the disease aetiology and progression. Several data types have been measured, bulk data, sorted cell data, cell-type data, single cell data, miRNA data, mass spec data, microscopy pictures, and spatially resolved data. microRNAs are important post-translational regulators, which are hypothesized to play an important role in various disease processes. Specialised resources on miRNA and their targets have been compiled.

Task 8, Mouse Macrophages, monocytes, neutrophils Repeat Task 7 for the respective mouse cells and tissues. In particular, it is interesting to investigate specific differences and similarities between the mouse and human systems.