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 2020-4
Applications of dynamical systems models for detailed quantitative predictions and Systems models for interpreting high-throughput data.

Working period: Three/Four weeks (19.1.2021-9.2.2021)
Hand-in anytime or in any exercise class
Please hand-in only reproducible results, answers, figures, tables, simulations, ...

Comprehensive systems models for interpreting large scale high-throughput data

Projects 3 built (systems) models and collected facts, network data and hypotheses of specific cellular systems.

In this project these models and networks will be used to analyse, interpret and validate high throughput transcriptomics and proteomics data. They will serve to generate hypotheses which will be checked against the 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 ... Many and quite comprehensive (meta-)databases on interaction and regulatory relations are available (Yeastract, ISMARA, BioGrid, ConsensusPathDB[1], Pathway Commons [2]), several of them including extensive documentation on formats (SBML, BioPAX), workflows, and analysis/visualization tools [3].

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 are the "model" systems (typically animals serving as model for the human case). 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, perturbation...).

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

[1] http://cpdb.molgen.mpg.de/: unique physical entities: 172,559 unique interactions: 660,318 gene regulations: 17,447 protein interactions: 448,725 genetic interactions: 4,793 biochemical reactions: 23,487 drug-target interactions: 165,866 pathways: 5,436

[2] Rodchenkov, I, et al., Pathway Commons 2019 Update: integration, analysis and exploration of pathway data, Nucleic Acids Research, Volume 48, Issue D1, p. D489-D497, https://doi.org/10.1093, https://www.pathwaycommons.org/, Version 12: 5,772 Pathways – 2,424,055 Interactions – 22 Databases.

[3] https://www.pathwaycommons.org/guide/

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 1 Yeast Transcriptomics: Heat shock gene expression time series with knock-outs

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!).

The project tries to understand yeast heat shock as a very typical and well studied stress condition. Heat shock employs major 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, MSN2, and MSN4). The exact role and function of theses TFs in yeast heat shock response and beyond is still not fully understood. So models should focus on transcriptional regulation and these TFs in particular.

Different data types on the system have been measured, i.e. transcriptomics, translatomics, proteomics, and regulatory data and an initial systems models has been proposed to analyse and interpret the measurements, in particular to understand the regulation "downstream" of the gene expression.[4]

Many and quite comprehensive databases on interaction and regulatory relations are available (Yeastract, ISMARA, BioGrid, ConsensusPathDB, Pathway Commons). A task in projects 3 was to make these kind of information available for the use in this project.

Our experimental collaboration partners measured the gene expression of the yeast heat shock response for several transcription factor knock-out strains. For each strain a time series with measurements after 10 and 30 min of heat shock at 37°C, as well as at 42°C was measured. Additionally, for each strain a control measurement at 25°C (this is the preferred normal growth condition for yeast) was done. Overall, this results for each strain in 5 measurements (control, 37°C 10 min, 37°C 30min, 42°C 10min and 42°C 30min). All measurements have been done in three replicates each. This yields $\{normal, \{10min, 30min\} \times \{37°C, 42°C\}\} \times \{r1, r2, r3\} = 15$ measured samples.

The transcription factors Hsf1 and Msn2/4 are known to be important for the heat shock response in yeast and, thus, can be assumed to be important in the experiment. Hsf1 is a heat shock specific transcription factor, while Msn2/4 regulate a more general stress response that is also activated in other types of stress such as salt or oxidative stress. We have time series measurements for the wild type, Msn2/4 knock-out, Hsf1 knock-out and a combination of Msn2/4 and Hsf1 knock-out. So overall, we have $15 * 4 = 60$ samples.

These 60 samples with different KO, different temperature and different time after heat shock allows a number of differential analyses which can be done fairly reliably using the information from the three replicates and appropriate grouping of the samples.

The data set has been collected and measured to learn about the stress response mechanisms in yeast and in the specific heat stress response in particular. Specifically it is unknown whether there are master regulators of the heat stress response or whether there are subsequent waves of TF activation

[4]Mühlhofer M., E. Berchtold, C. Stratil, G. Csaba, E. Kunold, N. Bach, S.Sieber, M. Haslbeck, R. Zimmer, J. Buchner, The Heat Shock Response in Yeast Maintains Protein Homeostasis by Chaperoning and Replenishing Proteins, Cell Rep. 2019 Dec 24;29(13):4593-4607.e8. doi: 10.1016/j.celrep.2019.11.109

leading to the overall stress response. It could also be, that heat is sensed by the yeast system and a more holistic response is triggered to do the necessary responses in order to return to the normal homeostasis. Even in this case, whether a return to homeostasis is possible, whether a TF-based specific heat shock response is induced, or whether the cell ultimately is doomed to apoptosis and cell death, could be dependent on the length or strength of the heat shock.

Differential analysis can be performed in a variety of ways for a range of scientific questions and hypotheses on the systems behaviour.

First of all, there is a variety of computational methods which process samples or sample groups and return a differential analysis between your specified samples or your groups of samples. We will provide differential expression results (such as fold changes and p-values) for different computational methods. This enables to investigate the differences of the methods and to obtain robust estimates, which are consistently produced by all or a sub-group of methods.

Lets assume that all methods somehow pre-process (mapping, normalization) the raw sequencing data and use replicates to obtain reliable fold change and significance estimates (a whole range of different methods have been proposed for these purposes but their effects and impacts are not a topic of investigation here).

In order to address various open research questions you could analyse

- the difference between individual samples, i.e. 42°C after 30 min vs control
- the difference between knock-out strains
  - between individual samples
  - for a time series
  - for a certain temperature
  - for all temperatures (heat) vs control (no heat)

For all of the differential analyses you could try to identify significant changes

- are the changes consistent or specific?
- are the functionally enriched in certain sets?
  - does it depend on the set definition (GO, KEGG, Reactome)?
  - the size of the individual sets?
  - the enrichment method?
- can network/pathway information be used?

Maybe genes could be clustered across the heat shock times and temperatures to identify groups of genes with the same behavioural pattern.

- Do the clusters of similar genes differ in their heat shock response between the different (knock-out) strains

- Which transcription factors are associated with these groups (e.g. by gene set enrichment using target sets of TF)?
- Can TF participating in gene regulatory networks be identified to be associated to these groups? Are these networks "striking" given the clusters from the data?
- How do these results differ for different computational differential expression methods or regulatory networks?
- Which transcription factors are active in the different time series? Use tools that predict transcription factor activities from gene expression data and compare the results for the knock-outs and different temperatures.
- Are there subsequent waves of transcription factors that are regulated by other transcription factors?

Whatever findings on differential genes, TFs, and pathways:

- is there evidence for findings already known?
- are there also new findings?
- Are there findings as significant and relevant than known results?
- Is there evidence and support for known hypotheses?
- Can anything be said about the above mentioned questions and goals of the study?

Can any result be cross-checked with other related data on yeast?

Are other relevant data types available as well, e.g. ATAC-seq, single-cell, or proteomics data? Do they support your claims?

Task 2 Human cell line Proteomics: Analysis of protein expression on SARS-CoV2 infection

Task 3 Human cancer Transcriptomics/Proteomics: Analysis of differential expression

Task 4: Human Macrophages, monocytes, neutrophils

Task 5: Mouse Macrophages, monocytes, neutrophils