**W2 – Metabolomics and Flux analysis**

Divided into two parts: Metabolomics and flux analysis. Both have different techniques and approaches for analysis.

**Metabolomics**: About measuring concentrations. Two approaches: untargeted and targeted profiling.  
Untargeted = discovery, no a priori knowledge needed, inductive hypothesis generation, relative concentration of many metabolites.  
Targeted = absolute quantification of few metabolites, need underlying knowledge, bottom-up strategy, hypothesis testing.

Challenges: large concentration differences, matrix effects, chemical compounds (diversity vs similarity), number of compounds, biological dynamics.

**Workflow in metabolomics**: At each step, there are pitfalls: sample, sampling, extraction, separation, detection.

Detection has different methods: MS methods (LC-MS, GC-MS) or NMR.

**Working principle of MS based methods**: Separates molecules based on their mass-charge ratio. Needs an ion source (transforms molecules from liquid to gas phase and ionizes them), mass analyzer (gate where only specific m/z ratios can pass) and a detector.

**Working principle of NMR based methods**: Good for detecting isotopes like 13C and 12C. Can also detect if these C-atoms are neighbours.

**Flux analysis**: Flux data quantify integrated biochemical and regulatory response of the network. It makes heavy use of mathematical and statistical methods.

**Def. flux**: The turnover of a metabolite, that is the amount that is being produced and consumed over a given time.

Stoichiometric flux analysis fails when: parallel pathways without related flux measurement, certain metabolic futile cycles, bidirectional reactions, split pathways with cofactors when not balanced.

Normally, one performs a 13C labelling experiment.

Global fitting of flux to network by iterative isotopologue balancing: simulate experiment in silico (choose fluxes, stoichiometric matrix S) and compare with in vivo experimental data. If simulation not good enough, optimize until fit is good enough (basically, parameter fitting of mathematical model).

**W3 – Computational and statistical tools**

Three models to analyse networks: topological for gene interactions, DNA-protein interactions, synthetic lethality etc., constraint based with FBA methods (only need to know how many molecules are used, not the reaction rate), mechanism based when biochemistry well known (reaction rates needed). For a topological or stoichiometric network, steady state is assumed.

**Flux balance analysis**

Solve Sv = 0 under constrains a\_i <= v\_i <= b\_i for all i, so that v is maximized. This gives ambiguous results: for irreversible reactions or cases, define a\_i = 0, else if reversible define a\_i != 0. There are an infinite number of possible solutions for 0, since the system is underdetermined, and we have more fluxes/reactions than metabolites.

FBA provides us with no information about dynamics or metabolite concentrations.

This is computationally very fast since v is a linear objective and with linear constraints (we call this **linear programming**). Also, the function is convex. To find the optimal solution path, we can make use of the simplex algorithm which lies on a vortex of the allowable solution space.

**Non-linear programming**: <S,f(v)> = 0 under constraints, a\_i, b\_i. In case of f(x) = x\*\*2, we speak of **quadratic programming**, which is used for metabolic adjustment questions:  
minimize f(v) = ||v – v\_WT||\*\*2, such that S\_mut\*v = 0.

**Mechanistic models**

Mechanistic models find their application in non-stationary states. It is modelled through a set of differential equations. Also, we need to choose a suitable framework such as MM or MA. Thus, we can computationally simulate the system’s behaviour and optimize the parameters if they do not faithfully predict model data. For model selection, we can perform **cross validation** or use information criteria such as **BIC** or **AIC**.

In high dimensional spaces, every single reaction requires a parameter and is thus a dimension. Therefore, #dim = #reactions (mind non-convexity too in high dimensional spaces).

**Approaches**: grid search, MCMC, adaptive gradient.

**Grid search**: only feasible for small models, since one can define x settings for a parameter and thus, it searches for all possible combinations for k parameters, thus we get a computation time of: x\*\*k.

**MCMC**: MCMC finds the best parameter combination based on its posterior distribution. We simulate parameters from a probability distribution that approximate the data (distribution of data). Useful in high-dimensional spaces, but still computationally expensive due to ODE solving in every step and manual tuning is required (e.g. proposal distribution).

**Problem with polynomial fits (overfitting)**

A polynomial fit can fit the data perfectly if dim(polynom) = #data\_points. This leads to overfitting, since for another data sample, the fit will completely fail. Therefore, a simpler model (smaller dimension) is more desirable to describe the data. There is a trade-off between model complexity and information gain. **Rule of thumb**: 1 parameter for 10 data points (if data sample is large).

**Theoretical approach to determine model complexity: cross validation and information criteria**

In cross validation, we divide the data in a training and test data set. Obtain parameters (⬄ train model) on training data and test parameters on test set to estimate generalization error. Repeat the process with a different subset of training and test data set and take the model with the smallest generalization error.

Also, we plot the dimension of model complexity against the root-mean-square error and choose the number of dimensions for the model, where the “sweet spot” (last bend before the curve flattens. Also, when the error = 1, then overfitting occurred).

The information criteria are alternatively used in scarce data. We evaluate the square deviation between model prediction and sampled data and add a penalty term for parameters/data size.

AIC = 2k – 2\*ln(L). BIC = k\*ln(n) – 2\*ln(L). (take the number of parameters with the smallest value for both criteria).

**W4 – The relationship between fluxes and metabolites**

Pathways that regenerate NADPH in E.coli: oxidative PP pathway, CAC, another pathway via transhydrogenase.

**How to design an experiment which faithfully captures pathway contribution to NADPH metabolism**: No perturbation since we want to know this exact situation: label substrate (glucose). Cells need to grow in steady state, measure the labels with labelling experiment (MS pattern or NMR pattern), make a simple stoichiometric model that includes atom position (NO ODEs), computational approach to simulate measurement, then algorithm for comparison and parameter fitting, then iterate again until difference is 0. We calculate the carbon flux and multiply with the stoichiometric numbers.

**The relationship between fluxes and metabolites**

In **metabolic systems analysis**, we start with a fluxome to understand component behaviour/interactions (proteins, metabolites, transcripts) for which we need **functional data**. To make **predictions of network operations**, we start with the proteome/metabolome/transcriptome, interfere component interactions to derive a fluxome and phenotype. Here, we work with **variables that are predicted with an ODE model**.

Metabolites determine fluxes (use enzyme kinetics and MM), while fluxes determine metabolites using system of ODEs (matrix form).

FBA identifies a set of optimal flux distributions under a given condition computationally. 13C-based analysis is an actual in vivo experiment which identifies a flux (that occurs in the cell), which may or may not be optimal.

**Synthetic lethality screen in systems biology**: A systems biology approach reduces the actual number of experiment to be performed. Mind the following: From all possible reactions in an organism, remove those that do not occur under the optimization function (removes the most), are active on glucose from 13C experiments and are synthetically lethal (of the remaining reactions). Mind redundant/alternative reactions.

A model does not normally capture all conceivable cases in such an analysis, for example, it fails when a perturbation is added. For this another model is more appropriate.

One model describes a glucose experiment better than the other (the one with the smaller SSR is the better), while the other model describes the same experiment with a fructose input perturbation better (the second model has in this case the lower SSR).

Both models can be combined with the pareto surface. The **pareto surface** is the set of all points that have a low SSR value for both experiments. These points cluster together and are the basis for the combined model (in a graph with SSR exp 1 vs SSR exp 2, it is basically the lower left corner where the points cluster).

**W5 – Perturbation in biological systems**

DNA level (permanent): mutagenesis, CRISPR, transposon, deletion libraries.  
RNA level: RNAi ; Protein level: chemical genetics (good for eukaryotes due to high abundance of small molecules, less suitable for prokaryotes).

**Transposon mutagenesis**: Transposons (=: Tn) are viral DNAs that encode a transposase which allows the Tn to re-insert itself somewhere else in the genome (some are also capable of copy-pasting themselves). This leads to gene disruptions. A transposon usually contains a resist gene (for which we can screen to identify successful mutants) which is flanked by “inverted sequences” (the transposase recognizes this special sequence and removes it from its genomic location, such that the Tn can be reinserted somewhere else).

After successful reinsertion, the Tn sequence and the disrupted sequence are amplified, then, the DNA is sequenced such that the disrupted gene can be identified using bioinformatic tools (sequence easy to find, since Tn sequence is known).

**Side note**: tRNA is the smallest gene in all organism (universally conserved).

The Tn sequence can also contain another gene, which might influence a process X in the cell (for example, make it more fit etc.).

**How to write DNA: silicon-based DNA synthesis**

…

**FRET (protein level)**: Good for signalling networks in single cells. When two molecules are in close proximity and one of them emits energy (photons), the other can be excited by that and fluoresces. This is useful in proteins that change its structure when something binds to them temporarily or to trace the way of one molecule. Also, one can screen for small molecules that inhibit the protein, such that no fluorescence occurs. This approach is especially suitable for genes that are redundant, since all of the produced proteins can be targeted at once in a reversible fashion.

**W6 – Network biology**

**Def. complex disease**: Several SNPs and an unknown interaction with the environment determines the disease phenotype. A complex disease need not have 100% penetrance (environment can hyper-compensate) and it occurs in different degrees of severity. Also, it is **common**.

**Def. integral negative feedback**: Negative feedback caused by the accumulation of a metabolite (instead of input of a metabolite as in normal feedback). This feedback is more temporally more flexible and dampens the amplitude of reaction (less/no oscillatory behaviour).

Association study ⬄ relationship between (surrogate) marker and disease.  
Epidemiological study ⬄ relationship between environment and disease.  
GWAS ⬄ relationship between genotype and disease.

**Network biology in complex diseases**

necessary: ¬A => ¬B ; sufficient: A => B

General approach: define network (nodes and edges), perturb network while measuring response, integrate data into network (network model + experimental data generates Boolean ODE models), validation.

Defining the nodes means to find correlations between the actors (genes, proteins, metabolites, receptors etc.). The edges are feedback arrows (and types). A network can be perturbed with drugs or siRNA (one can also do several perturbations on one network and use that data). The nodes are basically the number of a measured species, which has or has no relationship to other species.

**W7 – Graphical models for complex diseases**

GWAS: Best tool to link genomic variations to diseases. It provides no information about which genes are important, their interaction, cooperation and working mechanisms. It only identifies genes, although we cannot say they are causally linked to the disease (actual disease genes can be inherited as a haplotype with other genes, which have same p-value, but they do not play a role in the disease mechanism) – the identified genes are only correlated in one way or another with the disease.

**Note**: Metabolic networks usually flow into one direction, feedback loops are rare. Signalling networks have mostly feedback loops, since the ligand has to come off its target.

**Graphical models: nodes and edges (interaction-based models)**

Graphical models are useful to describe complex diseases or complex biological networks. The edges have a value (corresponds to strength or significance of an interaction approximately), which is defined by perturbing the involved nodes using statistical association through repeated perturbations.

**Workflow for edges**: Perturb the edge(s) with gene knockdown/overexpression, RNAi, drugs. Observe the population of interest (screen for it) and their proteome. On the other hand, perform mathematical modelling of the model parameters and generate proteome simulated data (prediction). From observed and predicted data, determine goodness of fit (repeat if fit is not good enough).

If an edge cannot be measured directly with the current techniques, infer a link with correlation (zero correlation means no link).

**W8 – Network biology and hierarchical networks**

Shortest path algorithms are used to reproduce a metabolic pathway. The pathway always starts by the source/sink compound and the compounds are weighted according to number of reactions.

Protein complexes: protein networks which interact with other proteins. Some components of a network will appear in more than just one network. These common components are called **cliques** and they connect networks together (in a spatial sense, such that one can picture a protein complex, e.g. a ribosome).

Real life cellular networks are often **scale-free networks** that is their edges are added iteratively to “the edge before” based on probability. The **diameter** of a network is the average node length of a network. In biological networks, they appear to be small and constant (removal of hubs increases diameter). Biological networks exhibit **small world** properties.

**Def. hub (systems biology)**: A molecule (protein, metabolite etc.) with a large number of potential interaction partners. Such molecules tend to be essential.

**Hierarchical networks**

**Network motifs** are subgraphs (patterns) that occur significantly higher than in random graphs. **Motif learning** is the counting of certain topology types and the enumeration of all subgraphs.

Interaction inference is carried out by clustering using algorithms such as k-means or hierarchical clustering. They are based on similarity (or dissimilarity) measures such as Euclidean distance, Pearson/Spearman correlation etc. **Note**: Different similarity measures account for different properties (we get different results). Therefore, it is important to know which measure is the most suitable. Euclidean distance is the shortest path between two components and it takes its absolute distance into account while the correlation coefficients look for similarities (the shape of the response) and don’t use the actual values.

Hierarchical graphs are useful for expression profiling.

**k-means procedure**: Define number of clusters (= k). Randomly pick k elements as the initial center points. Associate all other points to a center with a distance measure (e.g. Euclidean distance). Calculate the new mean point of the k clusters and repeat the previous step until convergence (center does not change anymore).

Clustering algorithms do not allow finding regulatory interactions among genes. Solution: Bayesian networks and module networks.

**W9 – MS in proteomics**

In MS, the protein is digested into peptides and the peptides are used for further analysis. Every cycle will produce different results from the same underlying protein sample, since not all peptides can be sequenced or analysed in one single trial. Thus, we have to deal with undersampling (incomplete data matrix).

An ionized protein can generate several peaks (since it can be split differently into its peptides) and a peak can be given by different isobaric species. Also, a protein may have multiple charges or contain isotopes (isotopes typically shift the peak by 1 m/z unit).

In MS/MS, we have a fragmentation stage after the first mass filter, which fragments the ion of interest into smaller ionized peptides through collision with gases (argon, nitrogen, helium) such that some kinetic energy is converted into intramolecular vibration that disrupts bonds. After this follows a second mass filter. This is done because there are far too many ions after the first filter which cannot all be analysed.

**Def. isobaric species**: Molecular species that have the same mass.

**Discovery MS**: Shotgun method on protein sample. A proteome is digested by peptides (10-20 amino acids, then prefractioned with HPLC) and then analysed. One can also choose a subset of the proteome (via mass filtering) to analyse a subsample. Reproducibility is low, because a proteome has many different proteins in different abundances. Only a small subset can be analysed, therefore not all proteins can be covered (especially those who are unabundant in the cell).

**Targeted MS**: Prior knowledge about protein and its fragments are necessary. Per injection, only a limited number of proteins can be analysed. It has a high reproducibility and repeatability as well as a high sensitivity.

The intensity of a peak correlates poorly with actual peptide abundance (depends on efficiency of ionization, fragmentation and competition of other peptides with a limited ion source).

**MS and modified proteins/peptides: phosphorylation**

**Workflow**: protein extraction from the sample containing phosphoprotein 🡪 digestion into phosphopeptides and normal peptides 🡪 phosphopeptide enrichment (remove normal peptides) 🡪 shotgun MS 🡪 quantification. The main ion fragment will often be the peptide with the phosphate removed.

Attachment of phosphates at a specific site is separable in retention time.

**W10 – Boolean networks and Boolean ODEs**

Regulatory networks (interaction-based) can be modelled with Boolean functions. A Boolean network P is optimized through: theta(P) = fit\_MSE + alpha\*complexity, alpha is size penalty for huge models.

**Genetic algorithms** are also used for optimizing Boolean networks. A sequence of truth values (corresponding to activity state of a molecule) is mutated by flipping one value in the sequence and evaluating the sequence with the objective function. Good partial solutions are preserved (similar as in protein folding).

Boolean networks are useful to study biological systems with oscillatory behaviour, since Boolean networks have an inherent time delay (thus oscillation is an inherent feature of Boolean networks).

Discretization of activity states is a source of bad fits. Boolean ODEs solve this problem with a continuous update rule.

Strengths of Boolean systems: simple, computationally inexpensive and fast, no need for stoichiometry, no need for different time scales to take into account. Good to use for networks with approximately constant time scales (protein networks only etc.).

Weaknesses of Boolean systems: simplicity does not cover whole biological reality, cannot incorporate different time scales (with feedback loops of different time scales, like in signalling networks), cannot be used to understand fine-tuning of system components.

**W11 – Design of a SB experiment**

**Design of a systems biology experiment: Com.**: “choose model organism. define network and nodes. define how to perturb the system. define data gathering method: applications of proteomics, transcriptomics, genomics, metabolomics etc.  
=> which method and how exactly would you extract the object of interest? (perhaps, you can also use cell biological methods such as FACS or microscopy-based approaches, but mainly, use systems biology approaches).  
how to analyze the obtained data? what to do with that data, how to proceed with experiment. dont forget to mention the controls. how can we generate a model from the data? compare measured data with predicted data using statistics and ML. data gathering and evaluation can also be done using databases.”

**Perturbation of a network**: There are different methods for different levels: CRISPR for DNA, RNAi for RNA, small molecules for enzymes, proteins, metabolites. On a more general note, one can change the pH, temperature and other external factors.

It is important to keep the time scales, reversibility and redundancy in mind when performing a perturbation. Time scales: seconds for small molecule perturbations (reversible), hours for de novo protein transcription and translation.

A drug perturbation sometimes depends on the cell line and other present media. Drugs are always dosage dependent. Thus, a drug can have a different or no effect than intended in different cells (mammalian cells, prokaryotic cells, cancer cells and all their different cell lines).

**Predictability of a model or tests**

Area under ROC (=: AUC) is the representative method: plot false positive rate vs sensitivity. If the curve approaches a very early steep increase and early saturation (area under ROC is big), then it is good for prediction. On the other hand, if it has a linear increase (45° line), then it is random (bad prediction).

Aliter: AUC = 1 ⬄ good prediction; AUC = 0 ⬄ useless model (random) ⬄ linear graph (45° angle)

**W12 – Noise and Chemical Master Equation**

Extrinsic noise: fluctuations in reaction rates/initial conditions, varying ion strength affecting ligand binding constant, initial cell-to-cell receptor count.  
Intrinsic noise: intrinsic stochasticity of chemical reactions, averaging effects with increasing count.

Noise can be modelled with ODEs: **Chemical master equation (=: CME)**

General form: dP/dt = f(P(x), t) – right: models stochastic behaviour; left: deterministic.

CME predicts data from rate constant. Moment equations are a convenient relationship between data and parameters. This allows to solve the inverse problem: predict rate constants from data. CME also implicitly describes a sequence of species moments and a sequence of distributions of markers.

To predict rate constants, we can do a nonlinear fit (Gaussian processes) or do a least-squares fit.

Sometimes, parameters are not unique due to structural non-identifiability or practical non-identifiability (gather more data to fix this problem).

**Sources of noise**: “random” movement of molecules, thus randomized interactions, translational bursting, finite number effects, promoter de/activation.

There is an intricate relationship between system size and noise. The smaller the system size the bigger the noise and vice versa (**finite number effect**). Let N denote average amount of molecules of a system, then: eta = sigma/N is proportional to 1/sqrt(N), with eta is the coefficient of variation.

Translational bursting is the phenomenon of increased mRNA translation into proteins when there are few amounts of mRNA such that there is equally many proteins present compared to another protein species, where the mRNA is more abundant (high vs. low translational efficiency). Thus, the less abundant mRNA has to be translated a lot more in order to keep the protein abundance of that protein equal to the other protein. This leads to an increase in noise during translational bursting (increase in fluctuation).

**Def. noise strength**: phi = sigma\*\*2/N. phi predicts that N (protein abundance) linearly increases with higher translational efficiency but remains constant with increased transcription. Phi can give insights into qualitative dependencies where eta cannot. Increase in phi gives no increase in eta.