Improving Network Integration Algorithms for Cancer Drug Predictions

In this project, we are going to use and build computational tools to analyse, compare and integrate molecular networks. We will work largely in R and Python and refine and build on an existing pipeline (see Fig. 1, right).
The analysis is based on public breast cancer data from a consortium (The Cancer Genome Atlas, TCGA) that measured thousands of molecular abundances for hundreds of patient tumours. We divided the patients into two clinically important breast cancer subtypes. The analysis starts with establishing **weighted molecular networks** for each molecule type. Second, the weighted networks are combined into an integrated network using biological knowledge and molecular interaction databases, or paired measurements. Further integration is performed by a semi-local integration technique. The integrated networks are compared between the patient groups, and the differential network is employed to predict differential effect of drugs.

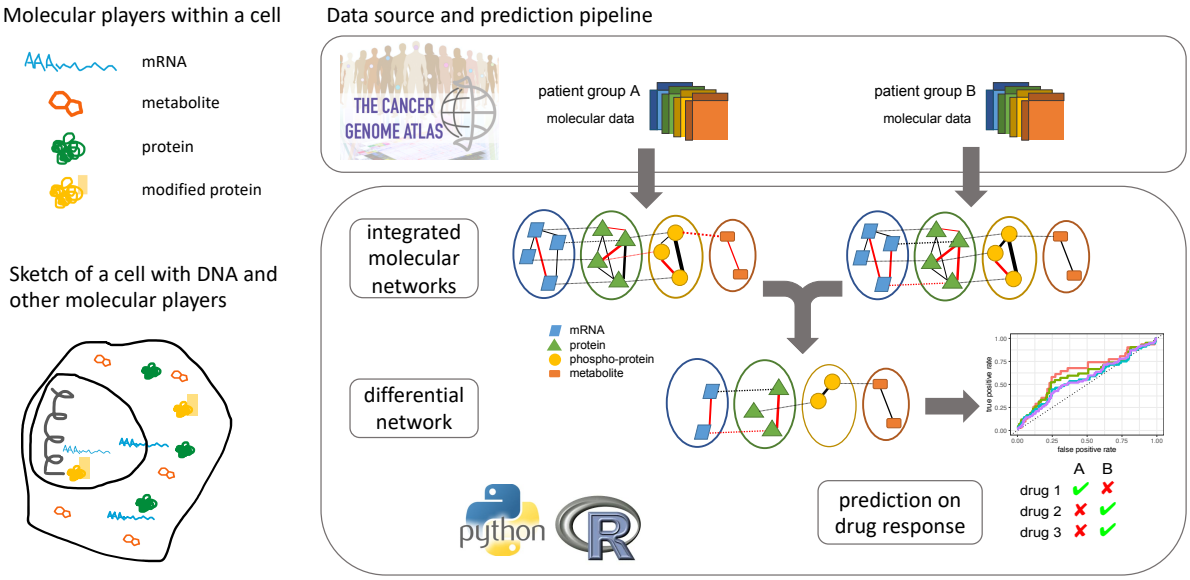

Fig. 1. Cellular molecules, and molecular network integration & analysis. © Katharina Baum, unpublished manuscript.

Infobox: Biological background

In human body cells, different types of molecules act together and determine cellular behaviour. They influence whether normal differentiation and tissue renewal prevails - or whether uncontrolled replication and migration mark cancerous tumour growth and metastases development.
Key cellular players are the DNA with its genetic information, and proteins created from it, but also the intermediates of protein synthesis (mRNA), products and educts of cellular metabolism (metabolites), and protein modifications (see Fig. 1, left). There are thousands of different genes, proteins and metabolites, and interactions within and between each layer of the same molecule type can exist. Consequently, the need emerges to consider and investigate the molecular players in their intricate network context.
Not only cancerous and normal cells differ, and different cancer types such as breast or colon cancer, but there are also different subtypes of cancer that require different medication. To push forward personalized, patient-specific drug recommendations, we aim to compare molecular network characteristics of different types of breast cancer.

A first task in the project is to cast the analysis pipeline into a usable data analysis tool. Furthermore, different perspectives could be explored.

1. We will deal with different methods to establish and reduce weighted networks from entity measurements. We can use measures such as **mutual information**. The networks consist of thousands of nodes, and reducing them to the relevant ones requires techniques from graph theory such as the **topological overlap matrix** to ensure certain topological criteria (e.g. **scale-freeness**) or stability in specific network properties. Alternatively, reduction to relevant biological pathways could be explored.
2. Uncertainty is an important factor when dealing with biological data that stem from measurements. It occurs not only for the molecular abundance measurements, but also within the interaction databases. Using algorithms for **spurious or missing edge detection**, e.g. by fitting to generative network models such as the **stochastic block model**, would be an option, or incorporating uncertainty as a second edge weight function, and determine its influence on the prediction.
3. Further uncertainty occurs when comparing networks derived from only few or even **single samples** (i.e. single patients). Approaching this personalized, precision medicine setting for a network-based analysis could be an option during this project.
4. The semi-local integration requires efficient **path detection** in large graphs for which algorithms could be further developed or adapted. A comparison of the integrated networks when using **network diffusion** or other **global algorithms** is possible.
5. Finally, how to model interactions not only of molecules but also of different drugs when administered in combination is an open question that could be pursued. Thereby, different options how **information** can **flow** through a network can be assessed, and what role differential interaction strengths play.

Infobox: Some mathematical concepts

A **weighted network** is a tuple (V, E, d) with V the set of vertices (nodes), $E \subset \{(x, y)|x, y \in V\}$ the set of edges, and $d : E \to \mathbb{R}$ the real-valued edge weight function.

Mutual information between two random variables X, Y is defined as the divergence between joint distribution and product of marginal distributions, $I(X;Y) = D(P_{(X,Y)}||P_X \otimes P_Y)$. If zero, X and Y are independent.

In **scale free** networks, the distribution given by the node degrees $D_x = |\{y|(x, y) \in E\}|$ of the nodes follow a power law for large degrees k, $P(D_x = k) \sim k^{-\gamma}$.

Network diffusion deals with the process of propagation of commodity ψ of the vertices along a network. The change in time of ψ at each node can be described by

$$\frac{d\psi}{dt} = C(A - D)\psi,$$

where $C \in \mathbb{R}$ the diffusion constant,

$$A_{ij} = \begin{cases} 1 & \text{if } (x_i, x_j) \in E \\ 0 & else \end{cases}$$

the adjacency matrix of the graph, and $D = diag((D_{x_i})_i)$ the diagonal matrix of the node degrees. Steady state as well as intermediate states of ψ can be considered.

Why should you join this project?

You will deal with large, densely connected, weighted networks (up to 30 000 nodes) and graph theoretic algorithms for path detection. There are many potential avenues this project can take, and we can adapt to your interest.

You will work with real-world molecular data from breast cancer patients and solve questions on prevailing molecular mechanisms. You will predict which drugs may be the best cure for which patient subgroup. Exploring other cancer types is always an option!

If you are familiar with R and Python and proficient in at least one of the languages, and interested in clinically relevant, biological questions, this project is for you.

Look forward to an engaged supervision by Prof. Dr. Bernhard Renard (bernhard.renard@hpi.de) and Dr. Katharina Baum (katharina.baum@hpi.de) in your upcoming master project! Please do not hesitate to contact us in case of questions.

Further reading

Topological overlap matrix in weighted network analysis (Zhang and Horvath 2005)
Spurious edge detection with stochastic block models e.g. (Baum, Rajapakse et al. 2019)
Single sample estimations e.g. (Koh, Fermin et al. 2019)
Path detection algorithms (Sedgewick 2002)
Network diffusion for network integration (Di Nanni, Bersanelli et al. 2020)
Information flow in complex networks (Harush and Barzel 2017)

Baum, K., J. C. Rajapakse and F. Azuaje (2019). "Analysis of correlation-based biomolecular networks from different omics data by fitting stochastic block models." F1000Res **8**: 465.

Di Nanni, N., M. Bersanelli, L. Milanesi and E. Mosca (2020). "Network Diffusion Promotes the Integrative Analysis of Multiple Omics." Frontiers in Genetics **11**: 106.

Harush, U. and B. Barzel (2017). "Dynamic patterns of information flow in complex networks." Nature Communications **8**(1): 2181.

Koh, H. W. L., D. Fermin, C. Vogel, K. P. Choi, R. M. Ewing and H. Choi (2019). "iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery." NPJ Syst Biol Appl **5**: 22.

Sedgewick, R. “Algorithms in C, Part 5: Graph Algorithms”, Addison Wesley Professional, 3rd ed., 2002.

Zhang, B. and S. Horvath (2005). "A general framework for weighted gene co-expression network analysis." Stat Appl Genet Mol Biol **4**: Article17.