Omics データ解析の最前線：
トランスクリプトームデータの新しい解析法 Upstream Analysis

トランスクリプトーム解析は、疾患発症の機序・薬理作用・細胞分化等の理解のために汎用される解析方法ながら、一般的なデータ解析ソフトでの結果に不満な研究者も多い。
本セッションでは、BIOBASE データベースと geneXplain アルゴリズムを用いた新規データ解析法等、多様な発現プロファイルの解析方法を説明する。生命現象に関わる重要制御因子の予測を行う解析方法 Upstream Analysis については、詳細に解説する。

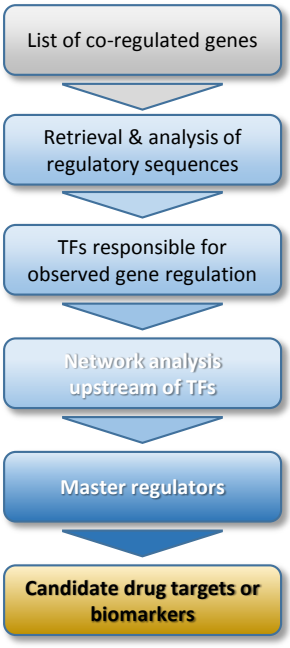
Upstream Analysis of gene signatures by geneXplain platform and BIOBASE databases

~ How to identify master regulators of the biological process of interest ~

Edgar Wingender
CEO, geneXplain GmbH, Wolfenbüttel, Germany
Professor of Bioinformatics and Director of the Department of Bioinformatics, University of Göttingen, Medical School (since 2002)

Generating high-throughput data sets became a standard approach to characterize the status of a biological system. For instance, monitoring the transcriptomic signature of a tissue is the typical means to characterize a certain disease state. But still today, the most of evaluation application of the data provide merely descriptive pictures. To overcome this, we have developed our proprietary "upstream analysis", which is a knowledge-based bioinformatic analysis of gene expression data and aims at identifying master regulators of the biological process under study.

The "upstream analysis" has been implemented in the geneXplain platform, an online workbench to run the daily computer applications in life sciences. It comprises state-of-the-art statistical methods, bioinformatic and systems biological methods, integrating public domain, own proprietary as well as commercial third-party products. The geneXplain platform also contains a user-friendly workflow management systems, which allows the user to compose recurrent tasks including their parameterization and store them for later re-use.



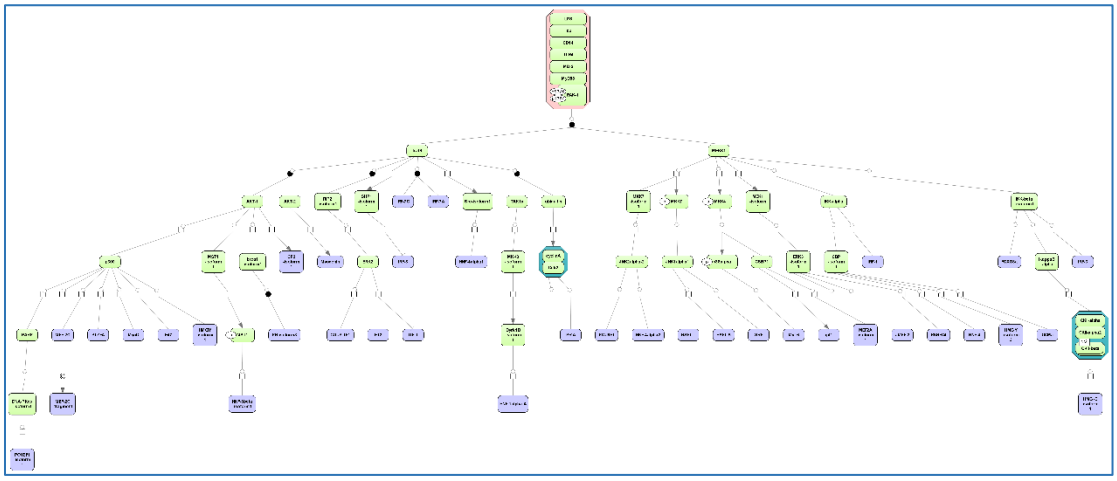
The "upstream analysis" applies a systematic and sophisticated analysis of the promoter sequences of co-regulated genes. For this, the BIOBASE TRANSFAC® Professional database, the gold standard for the identification of transcription factor (TF) binding sites, is used. Together with geneXplain algorithms, promoter models are generated that are usually very specific for the gene set under study. From these models, the TFs involved are automatically inferred. In a next step, the pathways controlling these TFs are generated with TRANSPATH®. Usually, these pathways converge at few molecules, so-called master regulators, which are promising targets for further drug development, or which by themselves (or their effectors) may serve as biomarkers for the process under study.

To validate our upstream analysis concept, we took a publicly available data set of IFN type III induced genes in human hepatocytes as a model of HCV infection (GEO dataset GSE31193). Among the TFs regulating these genes, we identified known regulators such as STAT and IRF factors, but also

several novel factors, interesting research targets. Starting from these TFs and applying a graph-analyzing algorithm to the signaling network represented in TRANSPATH®, we found several master regulators, such as, TLR3, TLR4, MYD88, RANK, TICAM1 and PIK3CA. Interestingly, the genes of these potential master regulators were also the IFN type III induced genes. This suggests positive feed-back loops operating during HCV infection and the subsequent host response, which is known to involve pathways where the identified master regulators play important roles in.

As an example, we show a network for one of the master regulators, TLR4, which is part of a complex with several other proteins (top). The blue symbols represent TFs, upstream of which the network has been analyzed, and the green color is for the intermediate molecules suggested by the graph-analyzing algorithm to trace back the paths from TFs to the master regulators.

Master regulatory molecules can be considered as potential biomarkers or drug targets.



A specific module, *DrugExpress*, in the geneXplain platform can be used to check whether the molecules in focus are regulated by any known drug.

DrugExpress is a collection of genome-wide transcriptional signatures of drug response. It contains sets of genes that significantly change their expression in response to treatment by different drugs. It originates from the Connectivity Map (also known as cmap) project developed at Broad Institute, USA (http://www.broadinstitute.org/cmap/).

At present, DrugExpress provides 642 sets of up- or down-regulated genes for 321 compounds grouped according to the classification of the respective drugs (for example, “anti-inflammatory drugs”).

Any molecule or list of molecules can be mapped to the DrugExpress ontology to be compared with the gene signatures of drug responses. The found matches may result in intriguing suggestions which of the 321 drugs can affect the expression of the molecules in focus. For example, mapping of disease-specific gene signatures to the gene signatures of drug responses may lead to suggestions which drugs can be potentially used for disease treatment, which may help to generate hypotheses for drug re-positioning.

In the use case described here, we have mapped the list of obtained master molecules onto the DrugExpress collection and found that many of them are regulated by one or even several drugs.