Integrative omics for the discovery of biosynthetic pathways using MEANtools (MEtabolite ANticipation tools)

Kumar Saurabh Singh[1,2], Hernando G. Suárez Duran[1], Justin J. J. van der Hooft[1], Saskia C. M. van Wees[2], Marnix H. Medema[1]

[1] Bioinformatics group, Droevendaalsesteeg 1, 6708PB Wageningen University and Research, The Netherlands; [2] Plant-Microbe Interactions group, Padualaan 8, 3584CH Utrecht University, The Netherlands

Introduction

MEANtools is a **python-based** workflow that integrates genomic, transcriptomic, and metabolomic data with enzymatic reaction databases to predict metabolic pathways, by identifying mass differences between metabolites that are co-abundant with transcripts whose enzymatic products are capable of catalysing biosynthetic reactions.

MEANtools first identifies strongly correlated transcript-metabolite pairs and then annotates these pairs with reactions by querying the **RetroRules** (Duigou *et al.* 2019) and **Natural Product** (Stokman *et al.* Msc. thesis WUR) databases. The annotated pairs represent an enzyme-encoding transcript with a protein domain capable of catalysing an enzymatic reaction that has a correlated metabolite as a substrate or product.

To further expand the chemical transformation search space of MEANtools, the RetroRules database has been further expanded by including **KEGG** (Kyoto Encyclopedia of Genes and Genomes) orthology (Nakaya *et al.* 2013) and **Rhea** (Alcantara *et al*. 2012) reaction databases.

pandas SQLite matplotlib NetworkX Network Analysis in Python

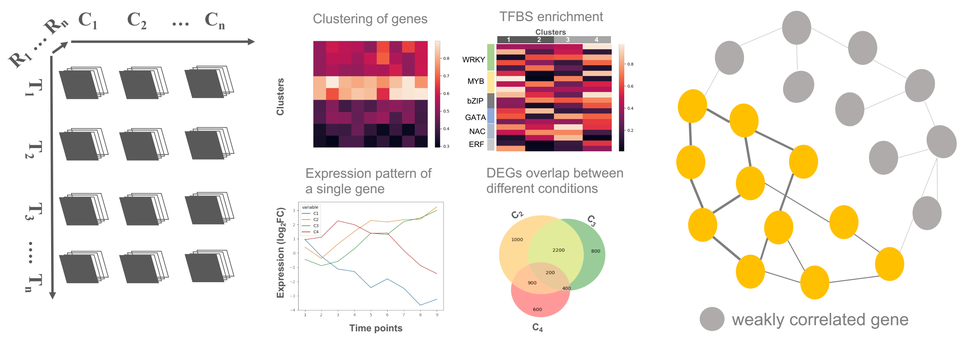
https://git.wur.nl/medema-group/meantools.git

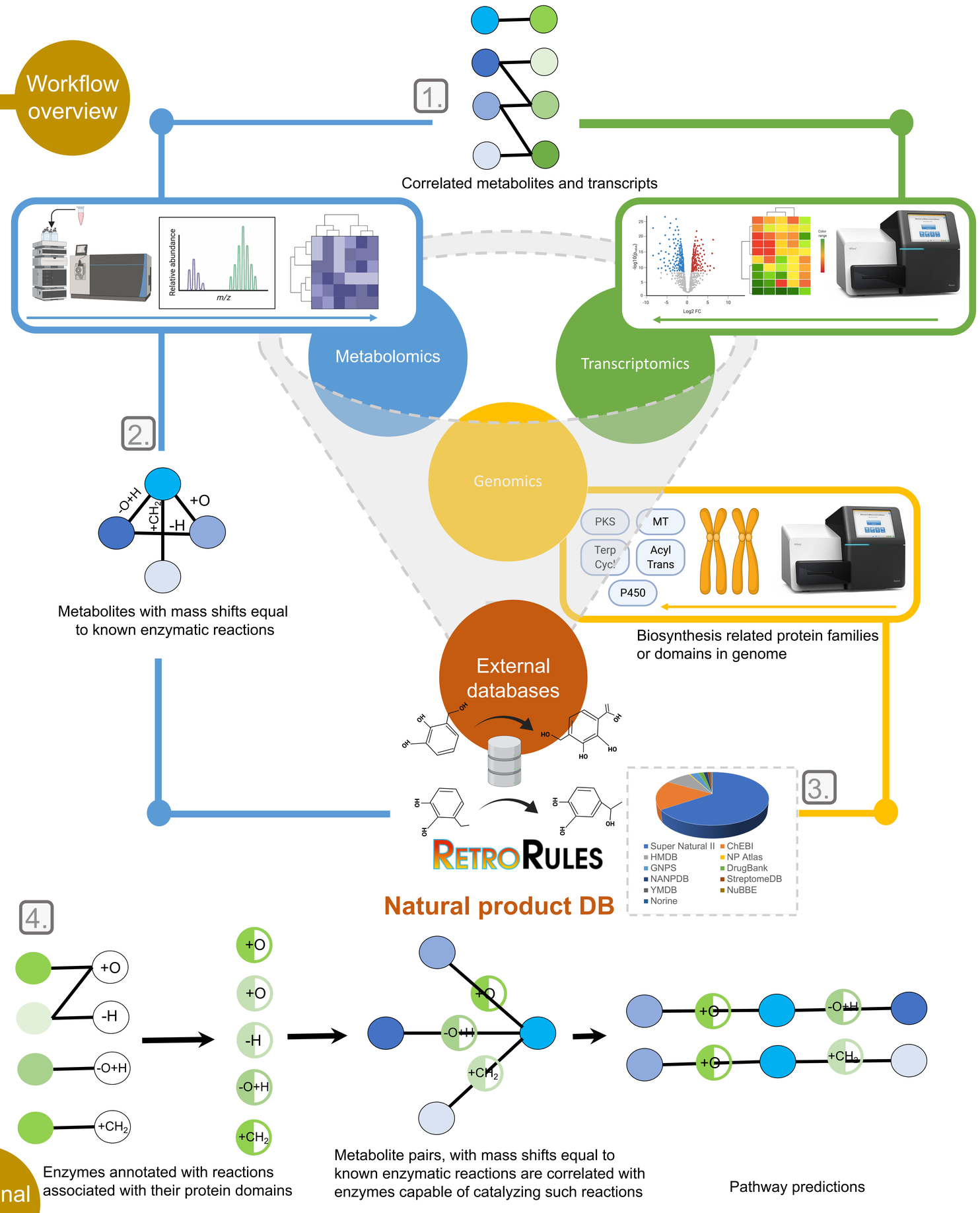
1. Correlations are computed between transcripts and metabolites based on gene expression and metabolite abundances.
2. Mass signatures from the metabolomic data are converted to Molecular Families (MF). MFs are then mapped with enzymatic reaction databases to identify pairs with mass differences associated with known enzymatic reactions.
3. Protein families/domains encoded by the genes are used to query enzymatic reaction databases.
4. MEANtools then identify cases in which metabolite pairs are correlated to a transcript that encodes an enzyme capable of catalyzing a reaction that explains their mutual mass difference.
5. The product of these reactions are then mapped to other mass signatures in the metabolome and the procedure is then repeated multiple times to generate pathway predictions.

Coexpression networks

Reconstruction of coexpression modules using time-based transcriptomics data is essential to identify groups of genes involved in the same metabolic pathway. MEANtools identifies coexpression clusters and allows selection of genes present in modules which improves confidence in the overall biosynthetic pathway predictions.

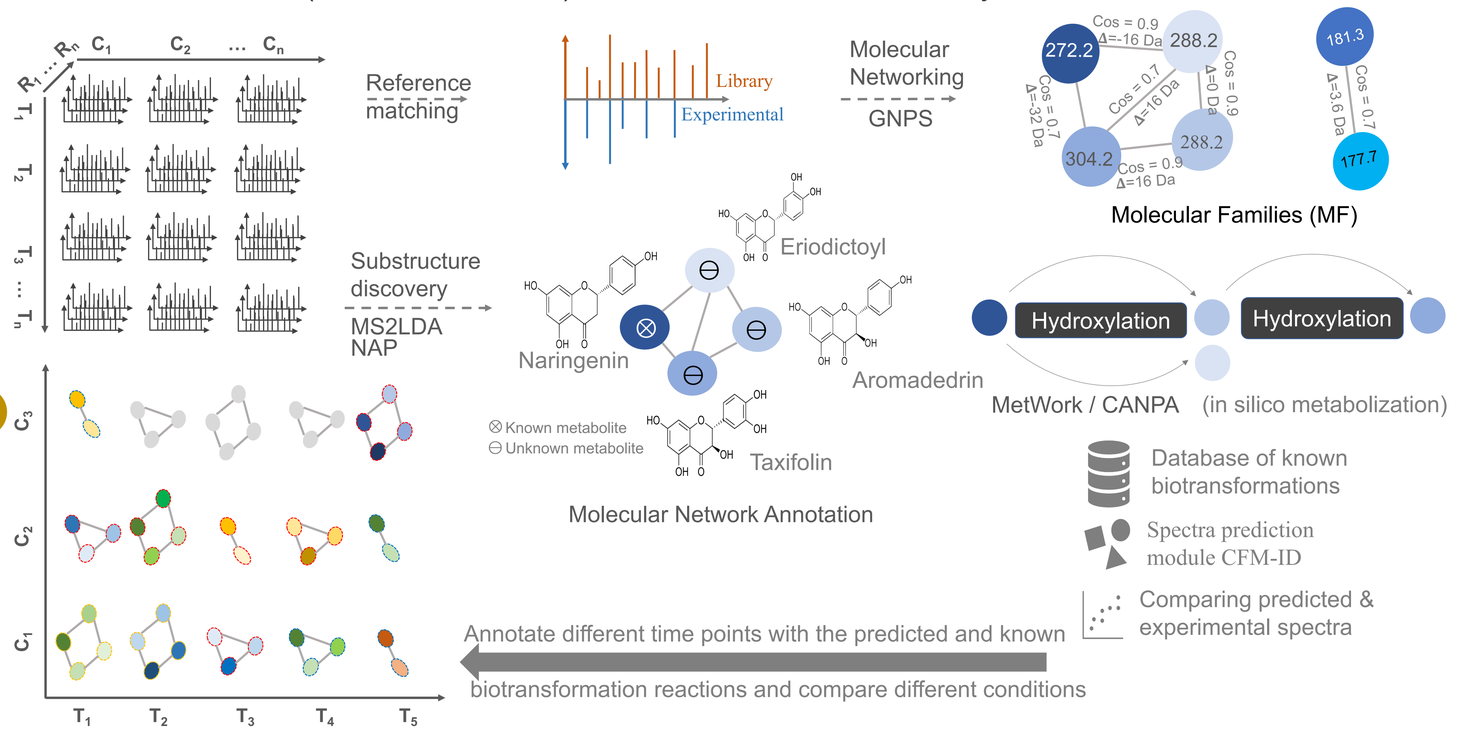
Transcriptomics with time-points | Temporal analysis of transcriptome data | Coexpression network





Time-based metabolomics

Time-based analysis has enormous potential to unveil transient metabolites both in terms of concentration and availability. Using time-series, spectral similarities can also be exploited to group several spectra together to form networks of fragmented features at individual time points. Here, tools like MetWork/CANPA (Ramos *et al.* 2019) can be useful for the discovery of unknown metabolites.



WAGENINGEN UNIVERSITY & RESEARCH

100years 1918 — 2018

Utrecht University

NWO Netherlands Organisation for Scientific Research

Dr. Kumar Saurabh Singh
Postdoctoral researcher
kumar.singh@wur.nl
k.s.singh@uu.nl
https://kumarsaurabhsingh.com
@Kumar_S_Singh