Metabolic pathway identification via unsupervised methods

Max Conway

Outline

- What a metabolic model is, and why you would want one
- How to make one
  - Basic data format
  - Steady state assumption
  - Biomass maximization assumption
- Controlling metabolism with gene expression
- Building up a multiplex network
- Collapsing it back down again with our take on Similarity Network Fusion
- Pathway labelling:
  - Linear approaches
  - Decision Trees
  - Restricted Boltzmann machine

Basic data format

- Input table or SBML file
- Can be transformed to stoichiometric matrix

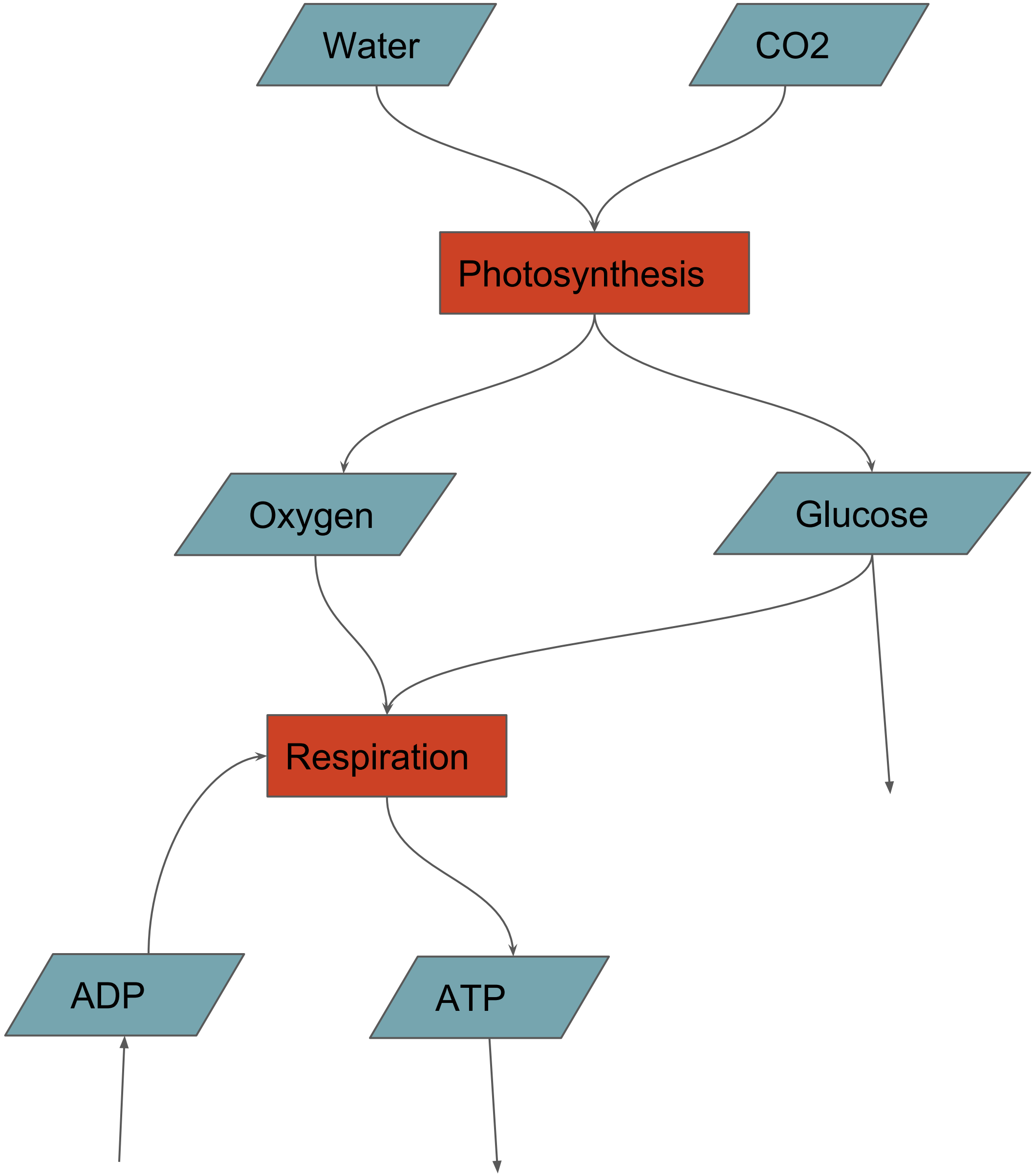
Name	Reaction	Min	Max
Respiration	$C_6H_{12}O_6 + 6\ O_2 \rightarrow 6\ CO_2 + 6\ H_2O$	0	100
Ex: Glucose	$\rightarrow C_6H_{12}O_6$	-100	1
Ex: Oxygen	$\rightarrow O_2$	-100	10
Ex: CO2	$\rightarrow CO_2$	-100	0
Ex: Water	$\rightarrow H_2O$	-100	10

=

	$C_6H_{12}O_6$	O_2	CO_2	H_2O
Respiration	-1	-6	6	6
Ex: Glucose	1	0	0	0
Ex: Oxygen	0	1	0	0
Ex: CO2	0	0	1	0
Ex: Water	0	0	0	1

Steady State assumption

- The reaction table and stoichiometric matrix tell us what reactions exist, and rough speed limits, but we need stronger assumptions to better understand how reactions relate.
- Therefore, we assume that the network is in steady state.



Biomass maximization

We need more constraints:

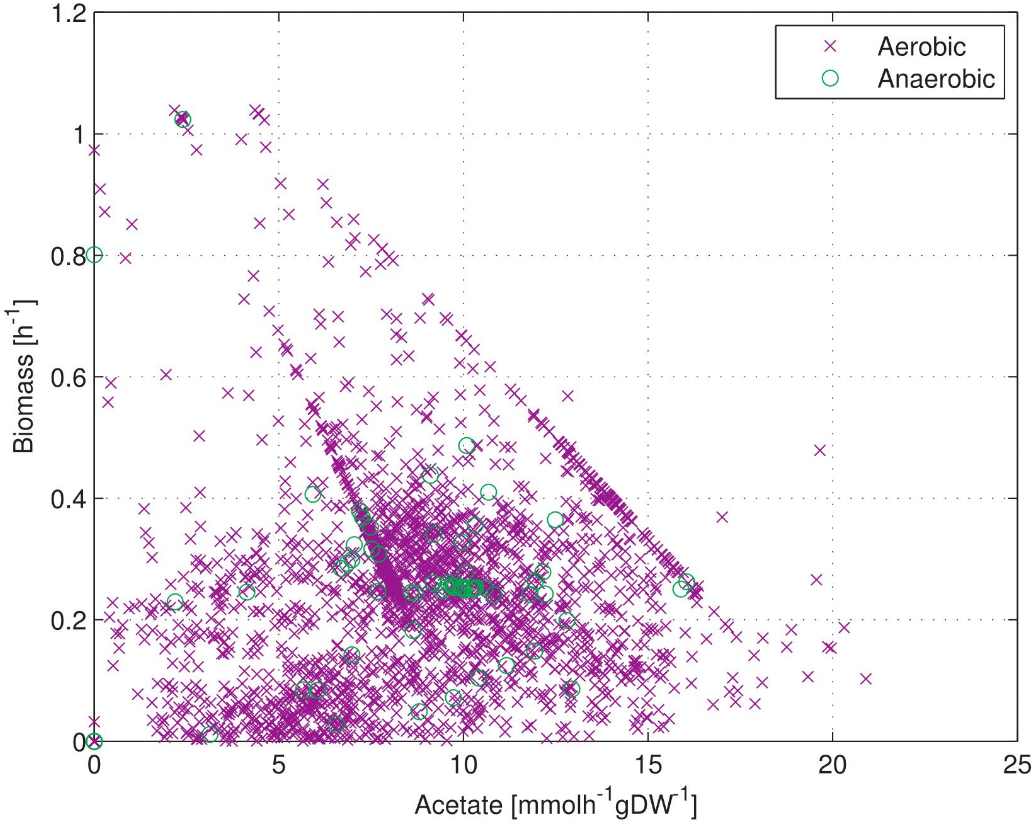
- Steady state constrains the model to possible phenotypes
- But which of these phenotypes is the one chosen by nature?
- The fittest one!
- We use linear programming on the constraints and stoichiometric matrix to find the model with highest biomass output.

Once we’ve got the fittest phenotype, we can find out what other properties it has:

- How would it respond to changes of condition?
- What metabolites would it produce?
- What can we do to make it produce more of the metabolites we’d like?

Adding Gene Expression

- Map gene expressions to flux bounds
- Use Colombos gene expression compendium
- Create a set of 2369 flux distributions with associated gene expressions



Building up a multiplex network

2369 individuals, each with:

- 4280 Gene expressions
- 1260 internal fluxes
- ~10 external fluxes

How do we interpret all this information?

Pivot the network:

- Before:
  - Nodes are reactions and metabolites
  - Edges are fluxes
  - Layers are individuals
- After:
  - Nodes are individuals
  - Edges are correlations
  - Layers are datasets (fluxes or genes)

Similarity Network Fusion

Basic similarity network fusion:

- First transform to similarity network (vs distance)
- Iteratively move each edge similarity closer to the mean of the parallel edges in other layers
- Wait for convergence

We used a weighted mean, rather than an unweighted mean.

This makes sense because our layers are not equivalent to each other.

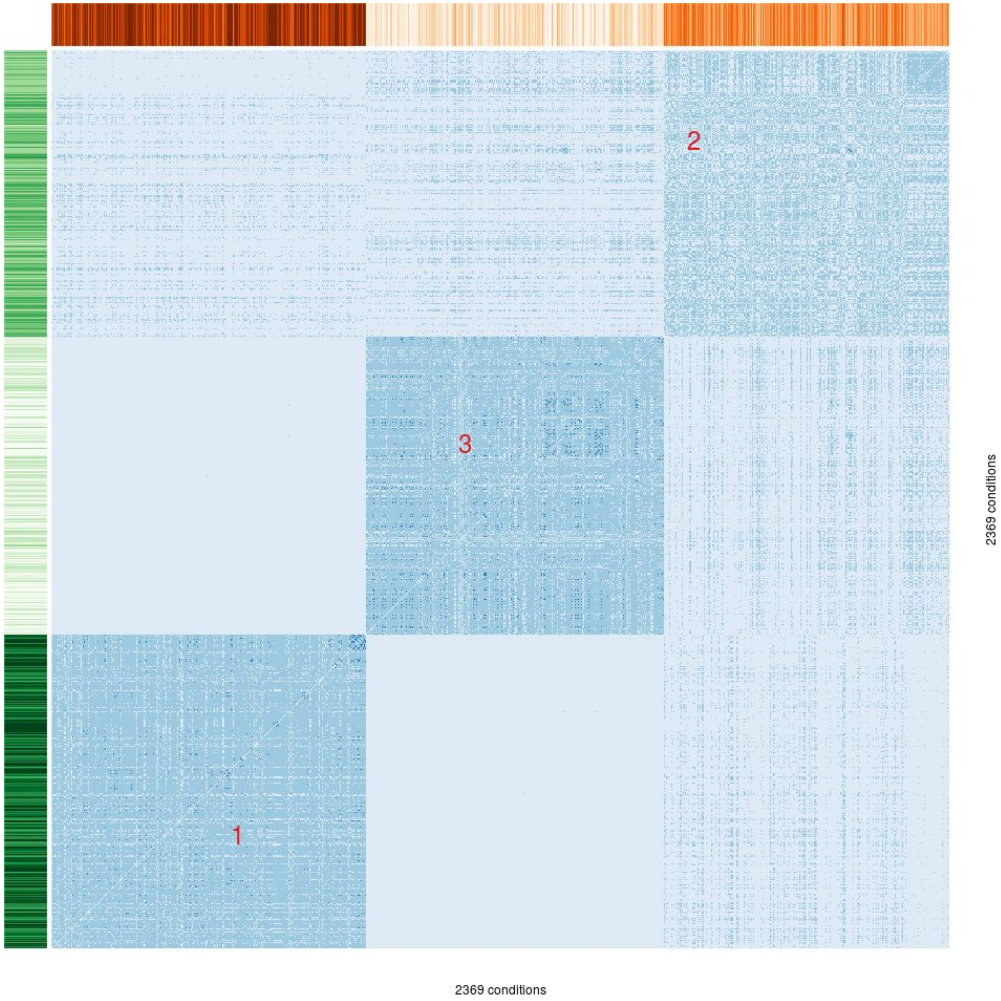
Results

Heat map of spectral clustering of fused network

- Orange top bar: 5-deoxyribose exchange
- Green side bar: biomass

X and Y axes are individuals, blue colour intensity is similarity.

But what does it mean?



What does it mean/what next?

Network clusterings are often hard to interpret

Implicit model in network algorithms is often less obvious than in tabular algorithms

Want to look at identifying structure within networks, such as subsystems

Labelling pathways

- Multiple valid labellings
- Subsystem annotations exist, but don’t tell us much
- A good model should be able to predict fluxes from other fluxes
- The structure of the model gives us the pathways
- We need an interpretable model

Linear approaches

Correlation with important fluxes

- Choose some important exchange fluxes (e.g. biomass, O2 excretion)
- See which reactions correlate with them
- Choosing more exchange fluxes gives us more information

Principal Component Analysis

- Natural conclusion of correlation based approach
- Look at every pair
- Loadings give us the amount of influence of each reaction

But:

- Can’t deal with nonlinearity
- Can only tell us average coefficient over all conditions

Decision tree

Regression tree, using R’s Cubist package.

- Build a decision tree
- Break it down into a set of rules
- Group the observations by the rules
- Interpolate using a regression model based on the remaining variables

Pros:

- Fast to build and run
- Piecewise-linear model makes sense given the structure of the dataset
- Highly accurate: cross-validated correlation > 0.99

Cons:

- Only predicts one flux at a time
- No obvious way to have one model predict all fluxes

Restricted Boltzmann Machine

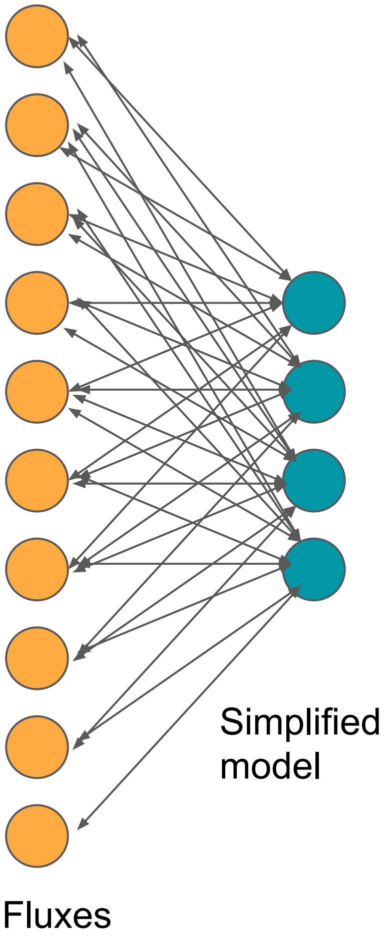
A neural network that predicts its own inputs

Pros:

- Simple change from classification network
- Adjustable model complexity (depth and width)
- Nonlinear

Cons:

- Slow to train



Summary

- Flux balance analysis metabolic models are detailed, steady state network models
- We estimate how continuous gene expression values affect them
- Looking at many gene expression vectors gives us a large multiplex network
- Similarity Network Fusion can help simplify this, but we still need more interpretability
- Linear dimension reduction can only take us so far
- Decision trees model the data well, but are not well suited to unsupervised use
- RBMs are more appropriate for nonlinear unsupervised learning

Thanks!

Questions?

Max Conway,
Claudio Angione,
Pietro Lio’

conway.max1@gmail.com
github.com/maxconway

EPSRC
Engineering and Physical Sciences
Research Council