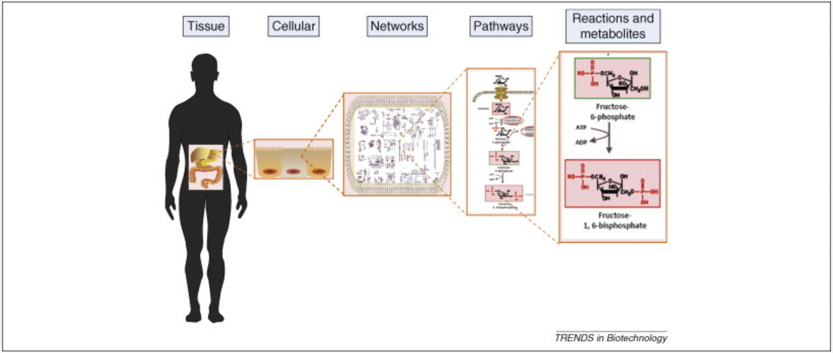
Phenomics validation of the Escherichia coli underground metabolic reconstruction

Claudio Angione

UNIVERSITY OF CAMBRIDGE

Studying a physiological system in silico [Mo and Palsson, 2009]



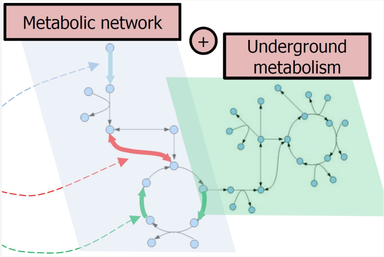
- Genome-scale models
- Huge variable space and objective space
- Computationally expensive to explore

Underground metabolism

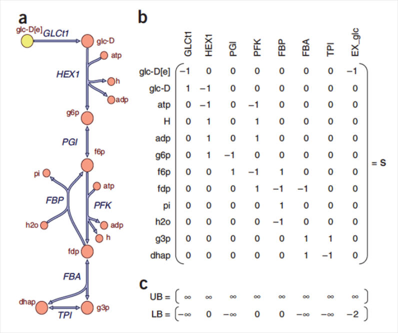
- Metabolism = set of chemical reactions taking place in living cells, with the aim of maintaining cellular functions
- Once considered only a passive result of the state of a cell
- Now widely recognized as a main contributor to cell behavior

Underground metabolism = reactions catalyzed with less catalytic efficiency, result of the weak side activity of preexisting enzymes

- 1380 genes
- 3027 reactions
- 2151 metabolites



Mathematical interpretation of FBA [Orth et al., 2010]



$$
\begin{array}{c|cccccccc}
 & \text{GLCt1} & \text{HEX1} & \text{PGI} & \text{PFK} & \text{FBP} & \text{FBA} & \text{TPI} & \text{EX\_glc} \\
\hline
\text{glc-D[e]} & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\
\text{glc-D} & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{atp} & 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\
\text{H} & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
\text{adp} & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
\text{g6p} & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\
\text{f6p} & 0 & 0 & 1 & -1 & 1 & 0 & 0 & 0 \\
\text{fdp} & 0 & 0 & 0 & 1 & -1 & -1 & 0 & 0 \\
\text{pi} & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
\text{h2o} & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
\text{g3p} & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
\text{dhap} & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0
\end{array} = \mathbf{S}
$$

$$
\text{UB} = \begin{pmatrix} \infty & \infty & \infty & \infty & \infty & \infty & \infty & \infty \end{pmatrix}
$$

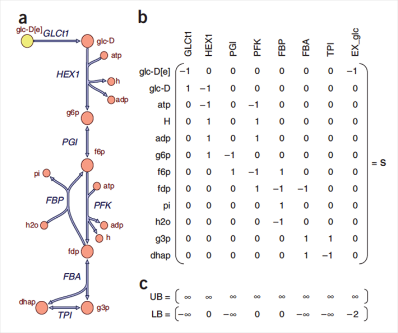
$$
\text{LB} = \begin{pmatrix} -\infty & 0 & -\infty & 0 & 0 & -\infty & -\infty & -2 \end{pmatrix}
$$

For every metabolite X_i, $i = 1, \dots, m$ a material balance is $\dfrac{dX_i}{dt} = \displaystyle\sum_{j=1}^{n} S_{ij} v_j$.

FBA: Steady state conditions $dX/dt = 0 \Rightarrow Sv = 0$

$$
\begin{aligned}
\text{maximise (or minimise)} \quad & f^{\mathsf{T}} v \\
\text{such that } & Sv = 0 \qquad (1) \\
& V_j^{min} \le v_j \le V_j^{max}, \quad j = 1, \dots, n
\end{aligned}
$$

Mathematical interpretation of FBA [Orth et al., 2010]



$$\begin{array}{r|cccccccc|}
 & \text{GLCt1} & \text{HEX1} & \text{PGI} & \text{PFK} & \text{FBP} & \text{FBA} & \text{TPI} & \text{EX\_glc} \\
\text{glc-D[e]} & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\
\text{glc-D} & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{atp} & 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\
\text{H} & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
\text{adp} & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
\text{g6p} & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\
\text{f6p} & 0 & 0 & 1 & -1 & 1 & 0 & 0 & 0 \\
\text{fdp} & 0 & 0 & 0 & 1 & -1 & -1 & 0 & 0 \\
\text{pi} & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
\text{h2o} & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
\text{g3p} & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
\text{dhap} & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0
\end{array} = \text{S}$$

$$\text{UB} = \begin{bmatrix} \infty & \infty & \infty & \infty & \infty & \infty & \infty & \infty \end{bmatrix}$$

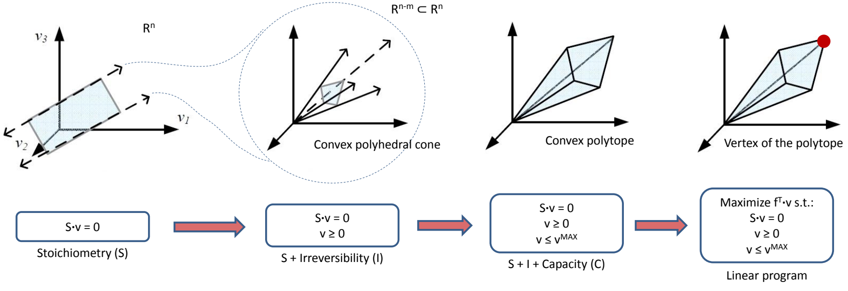
$$\text{LB} = \begin{bmatrix} -\infty & 0 & -\infty & 0 & 0 & -\infty & -\infty & -2 \end{bmatrix}$$

For every metabolite X_i, $i = 1, \ldots, m$ a material balance is $\dfrac{dX_i}{dt} = \displaystyle\sum_{j=1}^{n} S_{ij} v_j$.

FBA: Steady state conditions $dX/dt = 0 \Rightarrow Sv = 0$

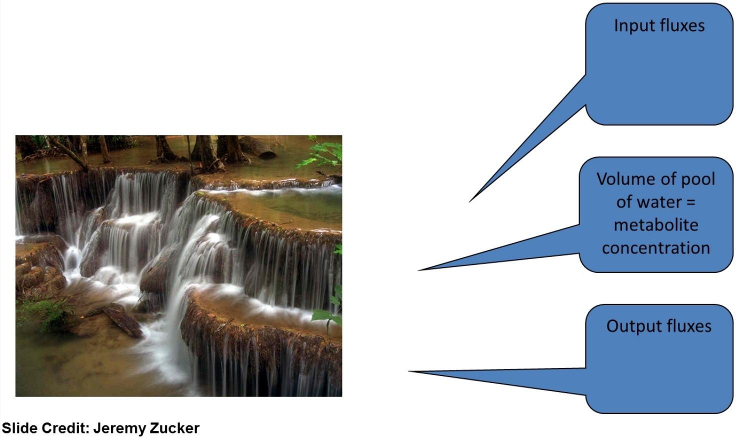
$$\begin{aligned} \text{maximise (or minimise)} \quad & f^{\mathsf{T}} v \\ & \text{such that } Sv = 0 \\ & V_j^{min} \leq v_j \leq V_j^{max}, \quad j = 1, \ldots, n \end{aligned} \qquad (1)$$

Geometric interpretation of FBA [Orth et al., 2010]



Mathematical formulation of flux balance analysis. The stoichiometric matrix S of n reactions and m metabolites restricts the search of possible flux distributions to the hyperplane $\mathbb{R}^{n-m}$. Thermodynamic constraints (irreversibility of reactions) limit the space of feasible solutions, which becomes a polyhedral cone. Capacity constraints (enzyme or transport capacities) constitute an upperbound for the flux rates; if this is available for every flux in the network (more specifically, it is sufficient that the upper bound is available for the edges of the cone), the feasible space reduces to a convex polytope. Once an objective function has been defined, the final linear program finds the final flux distribution as a vertex of the convex polytope, and then reconstructs the solution in the initial space $\mathbb{R}^n$.

FBA: pool in a waterfall



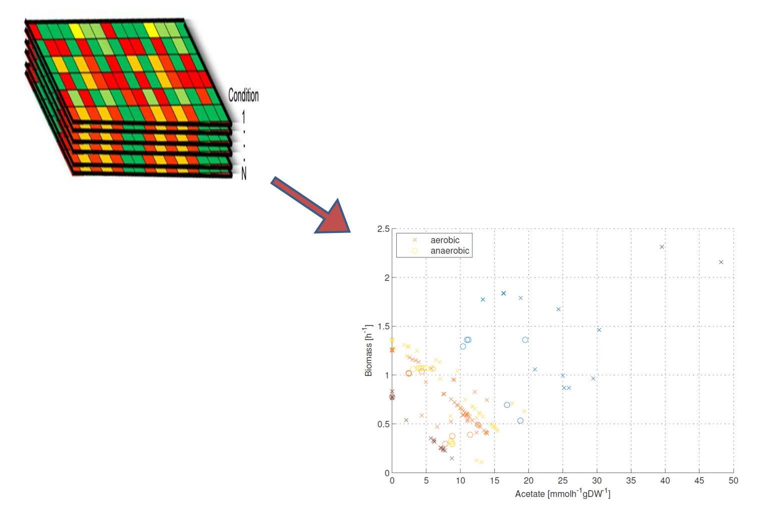
Input fluxes

Volume of pool of water = metabolite concentration

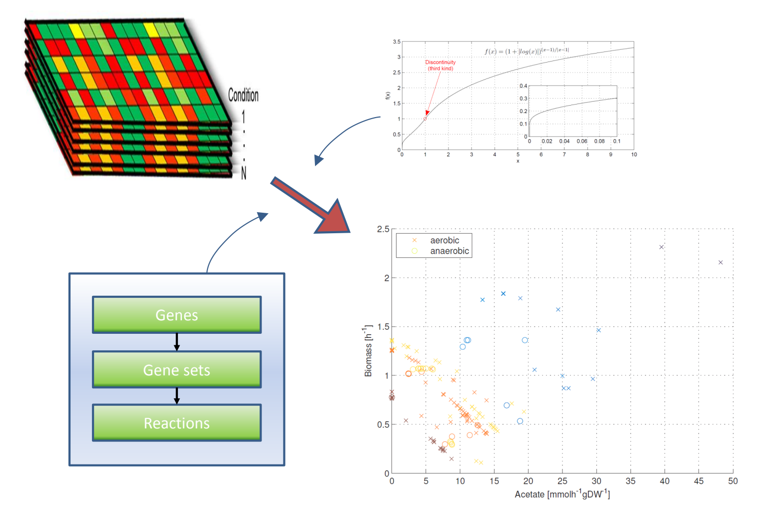
Output fluxes

Slide Credit: Jeremy Zucker

Gene expression profiles to Pareto optimality



Gene expression profiles to Pareto optimality


Condition
1
N
$f(x) = (1+|log(x)|)^{(x-1)/|x-1|}$
Discontinuity (third kind)
f(x)
x
aerobic
anaerobic
Biomass [h^{-1}]
Acetate [$mmolh^{-1}gDW^{-1}$]

Genes

Gene sets

Reactions

Condition = gene expression profile

Redefining Constraints for Linear Programming

The bounds of the fluxes depend on the gene expression:

$$\begin{array}{lll} \max & g^{\intercal} v & \\ \text{such that} & \max & f^{\intercal} v \\ & \text{such that} & Sv = 0 \\ & & v_i \geq V_i^{min} h(y_i) \\ & & v_i \leq V_i^{max} h(y_i) \end{array} \tag{2}$$

$$h(y_i) = \begin{cases} (1 + |log(y_i)|)^{\operatorname{sgn}(y_i - 1)} & \text{if } y_i \in \mathbb{R}^+ \setminus \{1\} \\ 1 & \text{if } y_i = 1 \end{cases} \tag{3}$$

where $\operatorname{sgn}(y_i - 1) = (y_i - 1)/\,|y_i - 1|$. y_i is the gene set expression of the ith gene set, responsible for the ith reaction of the model.

The bilevel problem can be converted to a single-level problem using linear programming duality theory: for every linear programming problem (primal) there exists a unique optimization problem (dual) whose optimal objective value is equal to that of the primal problem.

Redefining Constraints for Linear Programming

The bounds of the fluxes depend on the gene expression:

$$\begin{array}{lllr} \max & g^\intercal v & & \\ \text{such that} & \max & f^\intercal v & \\ & \text{such that} & Sv = 0 & (2) \\ & & v_i \geq V_i^{min} h(y_i) & \\ & & v_i \leq V_i^{max} h(y_i) & \end{array}$$

$$h(y_i) = \begin{cases} (1 + |log(y_i)|)^{\mathrm{sgn}(y_i - 1)} & \text{if } y_i \in \mathbb{R}^+ \setminus \{1\} \\ 1 & \text{if } y_i = 1 \end{cases} \qquad (3)$$

where $\mathrm{sgn}(y_i - 1) = (y_i - 1)/\,|y_i - 1|$. y_i is the gene set expression of the ith gene set, responsible for the ith reaction of the model.

The bilevel problem can be converted to a single-level problem using linear programming duality theory: for every linear programming problem (primal) there exists a unique optimization problem (dual) whose optimal objective value is equal to that of the primal problem.

Redefining Constraints for Linear Programming

The bounds of the fluxes depend on the gene expression:

$$\begin{aligned} \max \quad & g^{\mathsf{T}} v \\ \text{such that} \quad & \max \quad f^{\mathsf{T}} v \\ & \text{such that} \quad Sv = 0 \\ & \qquad v_i \geq V_i^{min} h(y_i) \\ & \qquad v_i \leq V_i^{max} h(y_i) \end{aligned} \tag{2}$$

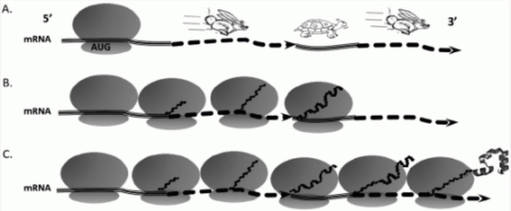
$$h(y_i) = \begin{cases} (1 + |log(y_i)|)^{\text{sgn}(y_i - 1)} & \text{if } y_i \in \mathbb{R}^+ \setminus \{1\} \\ 1 & \text{if } y_i = 1 \end{cases} \tag{3}$$

where $\text{sgn}(y_i - 1) = (y_i - 1)/\,|y_i - 1|$. y_i is the gene set expression of the ith gene set, responsible for the ith reaction of the model.

The bilevel problem can be converted to a single-level problem using linear programming duality theory: for every linear programming problem (primal) there exists a unique optimization problem (dual) whose optimal objective value is equal to that of the primal problem.

Codon usage bias affects protein production

- Codon = mRNA nucleotide triplet
- In a protein, each amino acid is encoded by up to six synonymous codons
- 4^3 codons, 61 of which actually encode for amino acids plus 3 stop codons
- but only 20 different translated amino acids
- some codons are slow, some are fast in the translation process
- therefore, the choice between these codons affects the translation rate and therefore the final amount of protein produced



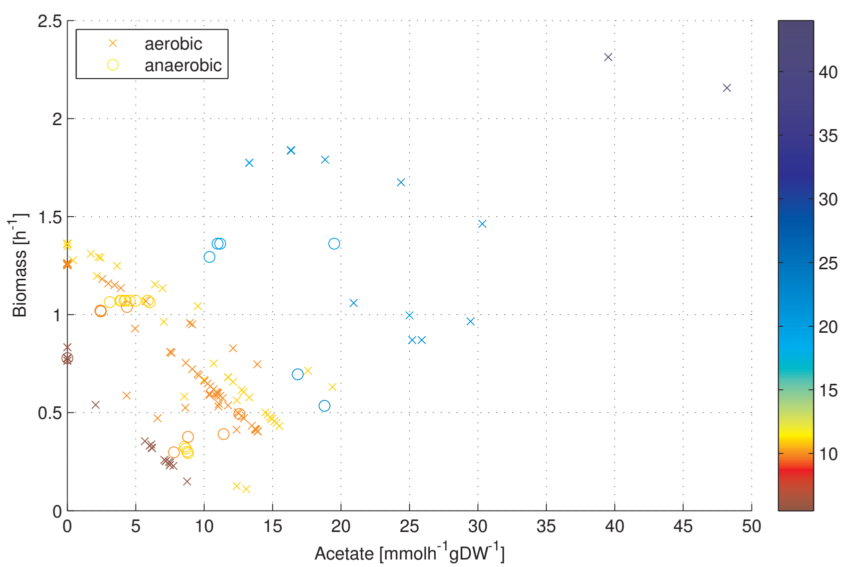
Applications of the multi-omic model

Why mapping gene expression profiles to a metabolic model?

Goal: analyzing (e.g. clustering) gene expression profiles using metabolism, rather than performing the analysis directly on gene expression data.

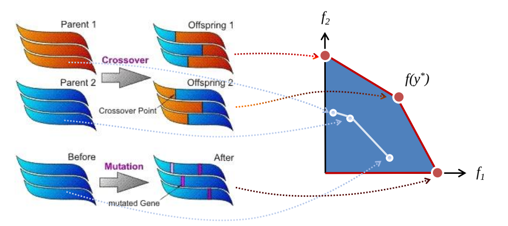
1. Multi-omic models cluster genes in their relative pathway; pathways can then be clustered and ranked through an effect-based approach (i.e., looking at the output outcomes in the phenotypic space)
2. The model acts as a ranking and noise-reduction tool: effect of low-importance genes is filtered out even if their expression is highly variable across conditions; without the multi-omic model, these genes would be incorrectly regarded as key genes to differentiate conditions
3. Performing inference directly on gene expression values may lead to incorrect prediction of the centrality of a gene whose level seems to be highly correlated with many other genes, but with only a marginal role in the metabolism (e.g., no impact on the biomass and on key metabolites)

466 *E. coli* experimental conditions



Multi-objective optimization

Let f be the vector of r objective functions to optimize in the objective space

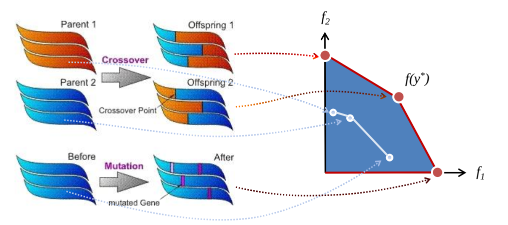


- Solution of a multi-objective problem: set of points called Pareto front
- Represents the best trade-off between two or more requirements
- A point y^* in the solution space is Pareto optimal if there does not exist a point y such that $f(y)$ dominates $f(y^*)$, i.e. $\nexists\ y$ s.t. $f_i(y) > f_i(y^*), \forall i = 1, ..., r$

[Angione et al. - Theoretical Computer Science, 2015]

Multi-objective optimization

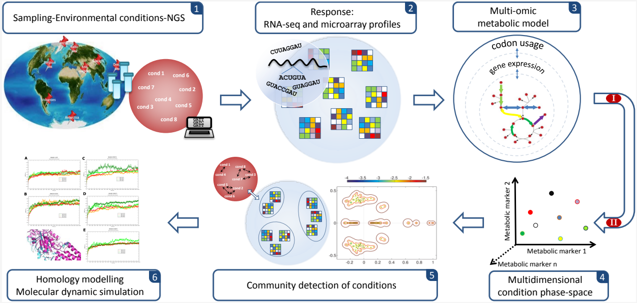
Let f be the vector of r objective functions to optimize in the objective space



- Solution of a multi-objective problem: set of points called Pareto front
- Represents the best trade-off between two or more requirements
- A point y^* in the solution space is Pareto optimal if there does not exist a point y such that $f(y)$ dominates $f(y^*)$, i.e.
  $\nexists\, y$ s.t. $f_i(y) > f_i(y^*), \forall i = 1, ..., r$

[Angione et al. - Theoretical Computer Science, 2015]

METRADE: the common framework



1 Sampling-Environmental conditions-NGS

2 Response:
RNA-seq and microarray profiles

3 Multi-omic
metabolic model

4 Multidimensional
condition phase-space

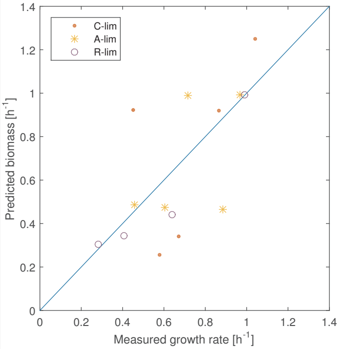
5 Community detection of conditions

6 Homology modelling
Molecular dynamic simulation

- Multi-omic model to associate conditions to a phenotypic outcome in a set of objective spaces
- Microarray expression profiles (environmental conditions) are mapped to a multi-omic model of metabolism
- A spectral method for community detection infers condition similarities according to the metabolic response in the multi-objective space.

[Angione and Lió - Nature Scientific Reports, 2015 (in press)]

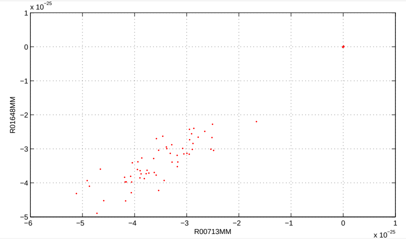
Validation of METRADE in *E. coli* underground metabolism



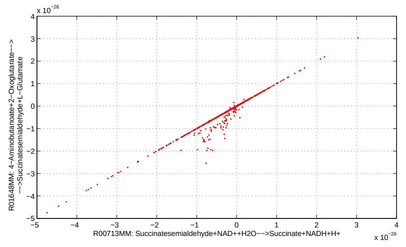
- Pearson's $r = 0.680$ (p-value $= 0.007$)
- Spearman's $\rho = 0.678$ (p-value $= 0.008$)

Predictions of biological processes through multi-omic models

Mitochondrial diseases (identifiability analysis)



(a) Fumarase deficiency

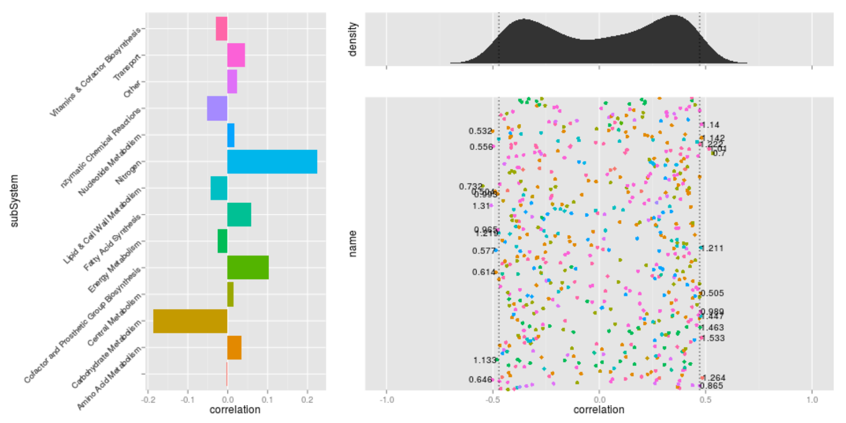


(b) Succinate dehydrogenase deficiency

- Monogenic diseases
- Inferring functional relations between flux rates
- Predicting the type of monogenic disorder through the shape of the functional relation

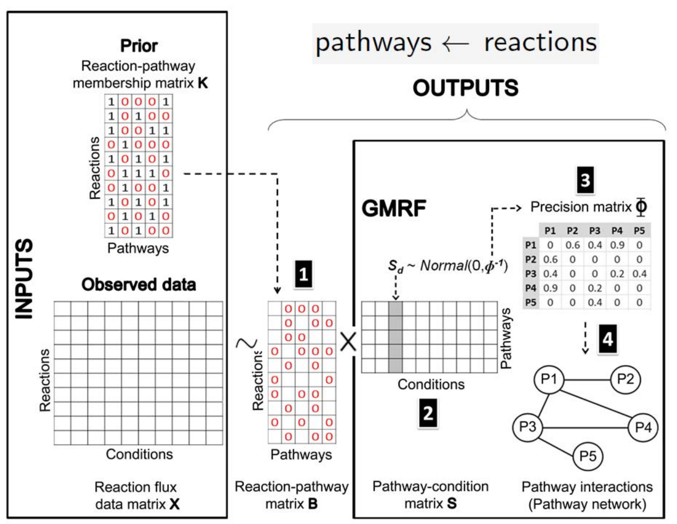
[Angione et al. - PLoS One, 2015 (in press)]

Geobacter sulfurreducens - correlation gene expression-position on the Pareto front



[Conway, Angione, Lió - Current Bioinformatics, 2015 (in press)]

Probabilistic Matrix Factorization with Gaussian Markov Random Fields



pathways ← reactions

INPUTS
Prior
Reaction-pathway membership matrix **K**
Reactions
Pathways
Observed data
Reactions
Conditions
Reaction flux data matrix **X**
OUTPUTS
1
~
Reactions
Pathways
Reaction-pathway matrix **B**
×
GMRF
S_d ~ Normal(0,ϕ^{-1})
Pathways
Conditions
2
Pathway-condition matrix **S**
3
Precision matrix Φ
4
P1
P2
P3
P4
P5
Pathway interactions (Pathway network)

1	0	0	0	1
1	0	1	0	0
1	0	0	1	1
0	1	0	0	0
1	0	1	0	1
0	1	1	1	0
0	1	0	0	1
1	0	1	0	1
0	1	0	1	0
1	0	1	0	0

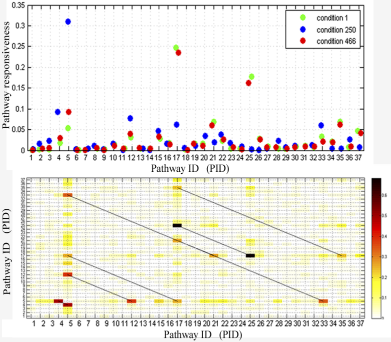
	P1	P2	P3	P4	P5
P1	0	0.6	0.4	0.9	0
P2	0.6	0	0	0	0
P3	0.4	0	0	0.2	0.4
P4	0.9	0	0.2	0	0
P5	0	0	0.4	0	0

[Angione, Pratanwanich, Lió - ACS Synthetic Biology, 2015]

Bayesian pathway analysis in *E. coli*

- Comparison of pathway responsiveness to different conditions (top)

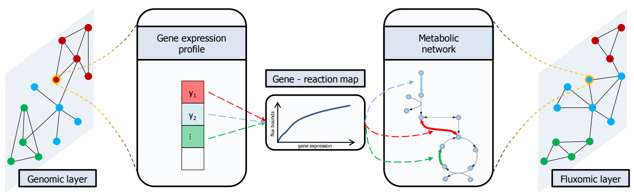
- Crosstalks between pathways can be viewed as a correlation matrix (bottom).



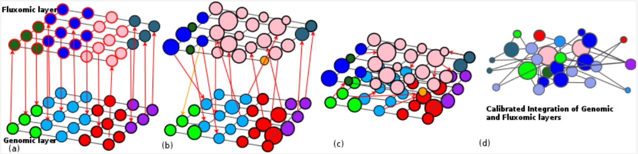
		High glucose		Low glucose	
PID	Pathway	Aerobic	Anaerobic	Aerobic	Anaerobic
5	Nucleotide salvage	0.0865	0.0965	0.1366	0.1714
17	Valine, leucine, and isoleucine metabolism	0.2219	0.2147	0.1974	0.1590
25	Alanine and aspartate metabolism	0.1544	0.1487	0.1285	0.1076

- Average responsiveness of the most responsive pathways across aerobic and anaerobic conditions of high and low glucose in *E. coli*.

Weighted integration of multi-omic layers of conditions

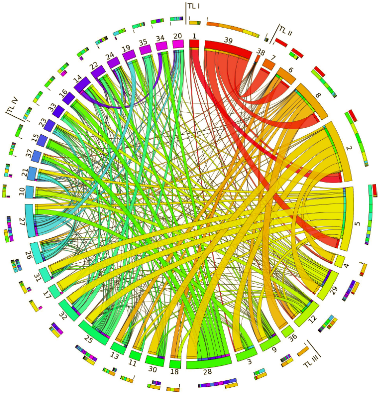


Integration of omic layers (genomics, metabolomics, fluxomics) using multiplex/multilayer network theory



[Angione, Conway, Lió - BMC Bioinformatics, under review]

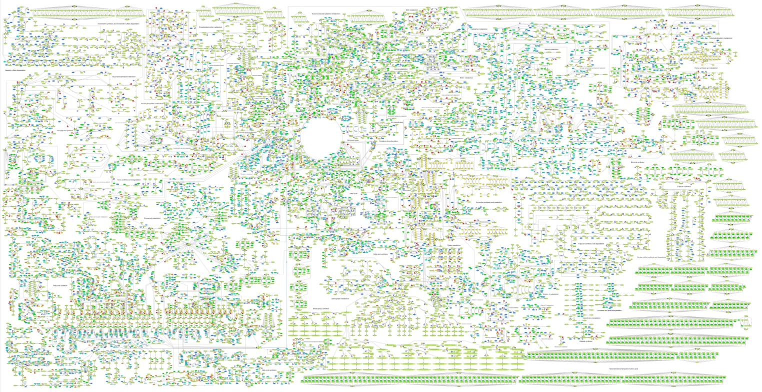
Adriatic Sea food web: PCB bioaccumulation network



1 Phytoplankton
2 Micro and mesozooplankton
3 Macrozooplankton
4 Jellyfish
5 Suprabenthos
6 Polychaetes
7 Commercial bivalves
8 Benthic Invertebrates
9 Shrimps
10 Norway lobster
11 Mantis shrimp
12 Crabs
13 Benthic cephalopods
14 Squids
15 Hake 1
16 Hake 2
17 Other gadiformes
18 Red mullets
19 Conger eel
20 Anglerfish
21 Flatfish
22 Turbot and brill
23 Demersal sharks
24 Demersal skates
25 Demersal fish 1
26 Demersal fish 2
27 Bentopelagic fish
28 European Anchovy
29 European Pilchard
30 Other small Pelagic Fish
31 Horse Mackarel
32 Mackarel
33 Atlantic bonito
34 Large Pelagic Fish
35 Dolphins
36 Loggerhead turtle
37 Sea birds
38 Discard
39 Detritus

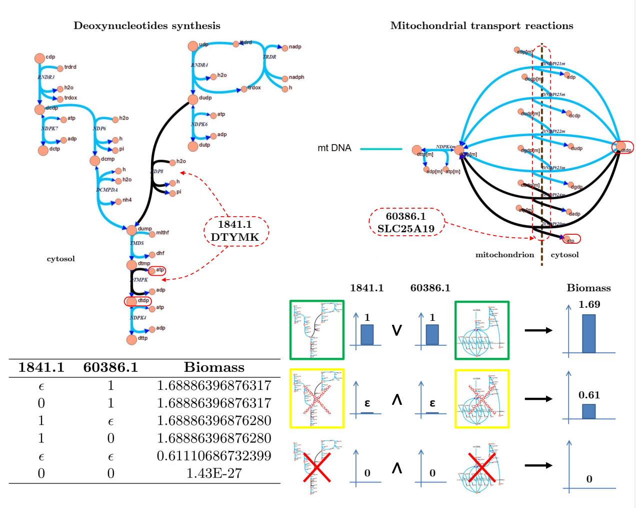
[Taffi et al. - Frontiers in Genetics, 2014]

Human metabolism



- 8 cellular compartments
- 7, 440 reactions
- 1, 789 enzyme-encoding genes
- 2, 626 unique metabolites

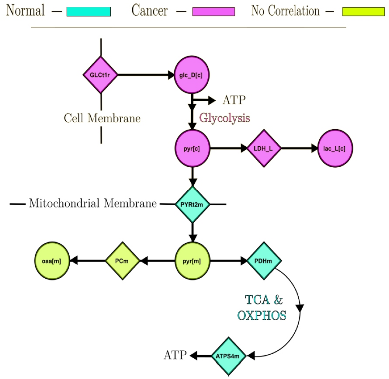
Genome-scale exploration of simultaneous gene effects



1841.1	60386.1	Biomass
ϵ	1	1.68886396876317
0	1	1.68886396876317
1	ϵ	1.68886396876280
1	0	1.68886396876280
ϵ	ϵ	0.61110686732399
0	0	1.43E-27

Human cancer metabolism

- We map METABRIC to the human metabolism using METRADE
- We independently find the Warburg effect
- 8 new biomarkers (reactions) significantly different in cancer/normal conditions
- All these new biomarkers have been previously associated with breast cancer



Contacts

UNIVERSITY OF CAMBRIDGE

claudio.angione@cl.cam.ac.uk

https://www.cl.cam.ac.uk/users/ca394/