Phenomics validation of the Escherichia coli underground metabolic reconstruction

Claudio Angione
Genome-scale models

Huge variable space and objective space

Computationally expensive to explore
Phenomics validation of the Escherichia coli underground metabolic reconstruction

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]

For every metabolite $X_i$, $i = 1, \ldots, m$ a material balance is

$$\frac{dX_i}{dt} = \sum_{j=1}^{n} S_{ij} v_j.$$

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

maximise (or minimise) $f^T v$

such that $Sv = 0$

$$V_{j}^{\text{min}} \leq v_j \leq V_{j}^{\text{max}}, \quad j = 1, \ldots, n$$

(1)
Mathematical interpretation of FBA [Orth et al., 2010]

For every metabolite $X_i, i = 1, \ldots, m$ a material balance is

$$\frac{dX_i}{dt} = \sum_{j=1}^{n} S_{ij} v_j. \tag{1}$$

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

maximise (or minimise) $f^T v$

such that $Sv = 0$

$V_{j}^{\text{min}} \leq v_j \leq V_{j}^{\text{max}}, \quad j = 1, \ldots, n$
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 $= \text{gene expression profile}$
Redefining Constraints for Linear Programming

The bounds of the fluxes depend on the gene expression:

\[
\begin{align*}
\text{max} & \quad g^T v \\
\text{such that} & \quad \text{max} \quad f^T v \\
\text{such that} & \quad Sv = 0 \\
\text{such that} & \quad v_i \geq V_i^{\text{min}} h(y_i) \\
\text{such that} & \quad v_i \leq V_i^{\text{max}} h(y_i)
\end{align*}
\]  

(2)

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

(3)

where \(\text{sgn}(y_i - 1) = (y_i - 1)/|y_i - 1|\). \(y_i\) is the gene set expression of the \(i\)th gene set, responsible for the \(i\)th 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:

\[
\begin{align*}
\max & \quad g^T v \\
\text{such that} & \quad \max & \quad f^T v \\
\text{such that} & \quad Sv = 0 \\
& \quad v_i \geq V_{im} \cdot h(y_i) \\
& \quad v_i \leq V_{im} \cdot h(y_i)
\end{align*}
\]  \quad (2)

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

where \(\text{sgn}(y_i - 1) = (y_i - 1)/|y_i - 1|\). \(y_i\) is the gene set expression of the \(i\)th gene set, responsible for the \(i\)th 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:

\[
\begin{align*}
\text{max } & \quad g^T v \\
\text{such that } & \quad \text{max } f^T v \\
\text{such that } & \quad Sv = 0 \\
& \quad v_i \geq V_i^{\min} h(y_i) \\
& \quad v_i \leq V_i^{\max} h(y_i)
\end{align*}
\]

\[\tag{2} \]

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

\[\tag{3} \]

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

![Graph showing acetate concentration and biomass production under aerobic and anaerobic conditions.](graph.png)
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. $
exists 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.
  $$
  \nexists y \text{ s.t. } f_i(y) > f_i(y^*), \forall i = 1, ..., r
  $$

[Angione et al. - Theoretical Computer Science, 2015]
METRADE: the common framework

- 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)

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

Phenomics validation of the Escherichia coli underground metabolic reconstruction

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

[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).

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

<table>
<thead>
<tr>
<th>PID</th>
<th>Pathway</th>
<th>High glucose</th>
<th>Low glucose</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Aerobic</td>
<td>Anaerobic</td>
</tr>
<tr>
<td>5</td>
<td>Nucleotide salvage</td>
<td>0.0865</td>
<td>0.0965</td>
</tr>
<tr>
<td>17</td>
<td>Valine, leucine, and isoleucine metabolism</td>
<td>0.2219</td>
<td>0.2147</td>
</tr>
<tr>
<td>25</td>
<td>Alanine and aspartate metabolism</td>
<td>0.1544</td>
<td>0.1487</td>
</tr>
</tbody>
</table>
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

[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

<table>
<thead>
<tr>
<th>1841.1</th>
<th>60386.1</th>
<th>Biomass</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon$</td>
<td>1</td>
<td>1.68886396876317</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1.68886396876317</td>
</tr>
<tr>
<td>1</td>
<td>$\epsilon$</td>
<td>1.68886396876280</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1.68886396876280</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>$\epsilon$</td>
<td>0.61110686732399</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1.43E-27</td>
</tr>
</tbody>
</table>

- Deoxynucleotides synthesis
- Mitochondrial transport reactions

Claudio Angione
University of Cambridge (UK)
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

claudio.angione@cl.cam.ac.uk

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