Introduction to systems biology: metabolic modeling

Esa Pitkänen

Elements of bioinformatics

7 December 2010
Lecture outline

- Biological networks
- Metabolic networks
- Metabolic reconstruction
- Flux balance analysis

A part of this lecture’s material is by Juho Rousu.
Biological networks

Many biological systems can be modeled as graphs or networks:

- Signaling networks
- Gene regulatory networks
- Metabolic networks
- Protein-protein interaction networks

This lecture concentrates on metabolic networks.
Biological networks

Biological systems of networks

Systems biology
- Biological networks
- Signaling networks
- Gene regulatory networks
- Metabolic networks
- Course: Computational Methods of Systems Biology

Metabolism
- Metabolic networks
- Metabolic reconstruction

Flux balance analysis

References
Signal transduction

- signal molecule & receptor
- activated relay molecule
- inactive signaling protein
- active signaling protein
- end product of the signaling cascade (activated enzyme)
Gene regulatory networks

Transcriptional regulation

regulatory region

gene

co-operative regulation

transcription factor

microarray experiments
Course: Computational Methods of Systems Biology

- III period, 2011
- Lecturer: Juho Rousu
- Course book: Analysis of Biological Networks (Junker, Schreiber; editors), 2008
What is metabolism?

- Metabolism (from Greek "Metabolismos" for "change", or "overthrow") is the set of chemical reactions that happen in living organisms to maintain life (Wikipedia)
What is metabolism?

- Metabolism (from Greek "Metabolismos" for "change", or "overthrow") is the set of chemical reactions that happen in living organisms to maintain life (Wikipedia).

- Metabolism relates to various processes within the body that convert food and other substances into energy and other metabolic byproducts used by the body.
What is metabolism?

- Metabolism (from Greek "Metabolismos" for "change", or "overthrow") is the set of chemical reactions that happen in living organisms to maintain life (Wikipedia).
- Metabolism relates to various processes within the body that convert food and other substances into energy and other metabolic byproducts used by the body.
- Cellular subsystem that processes small molecules or metabolites to generate energy and building blocks for larger molecules.
Why should we study metabolism?

- Metabolism is the “ultimate phenotype”
- Metabolic diseases (such as diabetes)
- Applications in bioengineering

Diabetes II pathway in KEGG

Lactose → Ethanol pathway, 2009.igem.org
Cellular space

- Density of biomolecules in the cell is high: plenty of interactions!
- Figure: *Escherichia coli* cross-section
  - Green: cell wall
  - Blue, purple: cytoplasmic area
  - Yellow: nucleoid region
  - White: mRNA

Image: David S. Goodsell
Enzymes

- Reactions catalyzed by *enzymes*
  - Example: Fructose biphosphate aldolase enzyme catalyzes reaction
    Fructose 1,6-biphosphate $\rightarrow$ D-glyceraldehyde 3-phosphate + dihydroxyacetone phosphate
- Enzymes are very specific: one enzyme catalyzes typically only one reaction
- Specificity allows *regulation*
Fructose biphosphate aldolase
Metabolism: an overview
KEGG Pathway overview: 8049 reactions (27 Nov 2009)
Metabolism

- What is metabolism?
- Why should we study metabolism?
- Cellular space
- Enzymes
- Metabolism: an overview
- Metabolism in KEGG

Metabolic networks

Metabolic reconstruction

Flux balance analysis

References

KEGG Pathway overview: 8049 reactions (27 Nov 2009)
Metabolism in KEGG

KEGG Pathway overview: 8049 reactions (27 Nov 2009)
Metabolic networks

- Metabolic network is a graph model of metabolism
- Different flavors: bipartite graphs, substrate graphs, enzyme graphs
- Bipartite graphs:
  - Nodes: reactions, metabolites
  - Edges: consumer/producer relationships between reactions and metabolites
  - Edge labels can be used to encode stoichiometry
Metabolic networks

- Metabolic network is a graph model of metabolism
- Different flavors: bipartite graphs, substrate graphs, enzyme graphs
- Bipartite graphs:
  - Nodes: reactions, metabolites
  - Edges: consumer/producer relationships between reactions and metabolites
  - Edge labels can be used to encode stoichiometry

![Metabolic network diagram](image-url)
Stoichiometric matrix

- The stoichiometric coefficient $s_{ij}$ of metabolite $i$ in reaction $j$ specifies the number of metabolites produced or consumed in a single reaction step
  - $s_{ij} > 0$: reaction produces metabolite
  - $s_{ij} < 0$: reaction consumes metabolite
  - $s_{ij} = 0$: metabolite does not participate in reaction

- Example reaction: $2 \, m_1 \rightarrow m_2 + m_3$
  - Coefficients: $s_{1,1} = -2$, $s_{2,1} = s_{3,1} = 1$

- Coefficients comprise a stoichiometric matrix $S = (s_{ij})$. 
Systems equations

- Rate of concentration changes determined by the set of *systems equations*:

\[
\frac{dx_i}{dt} = \sum_j s_{ij}v_j,
\]

- \(x_i\): concentration of metabolite \(i\)
- \(s_{ij}\): stoichiometric coefficient
- \(v_j\): rate of reaction \(j\)
Stoichiometric matrix: example

\[
\begin{array}{cccccccccccc}
\beta \text{G6P} & -1 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\
\alpha \text{G6P} & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\
\beta \text{F6P} & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
6\text{PGL} & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
6\text{PG} & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
R5P & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
X5P & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
\text{NADP}^+ & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
\text{NADPH} & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
H_2O & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\end{array}
\]
Kinetic models

- Dynamic behaviour: how metabolite and enzyme concentrations change over time → Kinetic models
  - Detailed models for individual enzymes
- For simple enzymes, the Michaelis-Menten equation describes the reaction rate $v$ adequately:

$$v = \frac{v_{max} [S]}{K_M + [S]},$$

where $v_{max}$ is the maximum reaction rate, $[S]$ is the substrate concentration and $K_M$ is the Michaelis constant.
Kinetic models

• Require a lot of data to specify
  ◦ 10-20 parameter models for more complex enzymes
• Limited to small to medium-scale models
Spatial modelling

- “Bag-of-enzymes”
  - all molecules (metabolites and enzymes) in one “bag”
  - all interactions potentially allowed
- Compartmentalized models
- Models of spatial molecule distributions
Compartments

- Metabolic models of eukaryotic cells are divided into *compartments*
  - Cytosol
  - Mitochondria
  - Nucleus
  - ...and others
- Extracellular space can be thought as a “compartment” too
- Metabolites carried across compartment borders by *transport reactions*
Modelling metabolism: steady-state models

- **Steady-state assumption:** internal metabolite concentrations are constant over time

\[ \frac{dx}{dt} = 0 \]

- External (exchange) metabolites not constrained
Modelling metabolism: steady-state models

- **Steady-state assumption:** internal metabolite concentrations are constant over time
  \[
  \frac{dx}{dt} = 0
  \]
  
  - External (exchange) metabolites not constrained
  - Net production of each internal metabolite \(i\) is zero:
    \[
    \sum_j s_{ij}v_j = Sv = 0
    \]

- *Is this assumption meaningful? Think of questions we can ask under the assumption!*
Modelling metabolism: steady-state models

- **Steady-state assumption**: internal metabolite concentrations are constant over time

\[
\frac{dx}{dt} = 0
\]

- External (exchange) metabolites not constrained
- Net production of each internal metabolite \( i \) is zero:

\[
\sum_j s_{ij}v_j = Sv = 0
\]

- *Is this assumption meaningful? Think of questions we can ask under the assumption!*
- Steady-state reaction rate (*flux*) \( v_i \)
- Holds in certain conditions, for example in chemostat cultivations
Metabolic reconstruction

- Reconstruction problem: *infer the metabolic network from sequenced genome*
- Determine genes coding for enzymes and assemble metabolic network?
  - *Subproblem of genome annotation?*
Metabolic reconstruction

- Reconstruction process
- Data sources for reconstruction
- Annotating sequences
- Assembling the metabolic network
- Gaps in metabolic networks
- In silico validation of metabolic models

Enzyme kinetics
Regulation
Spontaneous reactions
Compartments
Network-level modeling

Gene/Protein Function Prediction
Blast, Fasta
Profile methods (HMMs)
Reconstruction process

KAAS (KEGG Automatic Annotation Tool)

Genome annotation

Set of reactions/pathways

In silico model validation

Preliminary metabolic model

Pathway Tools/Gap filler

Flux Balance Analysis

Biological databases

Literature

High-throughput data

Metabolic model

Wetlab experiments

In silico analysis

Export for further use: SBML

References

Elements of bioinformatics: Introduction to systems biology
Data sources for reconstruction

- Biochemistry
  - Enzyme assays: measure enzymatic activity
- Genomics
  - Annotation of open reading frames
- Physiology
  - Measure cellular inputs (growth media) and outputs
  - Biomass composition
Resources

- Databases
  - KEGG
  - BioCyc

- Ontologies
  - Enzyme Classification (EC)
  - Gene Ontology

- Software
  - Pathway Tools
  - KEGG Automatic Annotation Server (KAAS)
  - MetaSHARK, MetaTIGER
  - IdentiCS
  - RAST
Annotating sequences

1. Find genes in sequenced genome (available software)
   - GLIMMER (microbes)
   - GlimmerM (eukaryotes, considers intron/exon structure)
   - GENSCAN (human)

2. Assign a function to each gene
   - BLAST, FASTA against a database of annotated sequences (e.g., UniProt)
   - Profile-based methods (HMMs, see InterProScan for a unified interface for different methods)
   - Protein complexes, isozymes
Assembling the metabolic network

- In principle: for each gene with annotated enzymatic function(s), add reaction(s) to network (gene-protein-reaction associations)
Assembling the metabolic network

- In principle: for each gene with annotated enzymatic function(s), add reaction(s) to network (gene-protein-reaction associations)
- Multiple peptides may form a single protein (top)
- Proteins may form complexes (middle)
- Different genes may encode isozymes (bottom)

Reed et al., Genome Biology, 2003.
Gaps in metabolic networks

- Assembled network often contains so-called *gaps*
- Informally: gap is a reaction
  - “missing” from the network...
  - ...required to perform some function.
- A large amount of manual work is required to fix networks
- Recently, computational methods have been developed to fix network consistency problems
Gaps in metabolic networks

May carry steady-state flux – Blocked – Gap

Gaps in metabolic networks
- May carry steady-state flux
- Blocked – Gap

Flux balance analysis

References
Gaps in metabolic networks

May carry steady-state flux – Blocked – Gap

Systems biology
Metabolism
Metabolic networks
Metabolic reconstruction
- Metabolic reconstruction
- Reconstruction process
- Data sources for reconstruction
- Annotating sequences
- Assembling the metabolic network
- Gaps in metabolic networks
- Gaps in metabolic networks
- In silico validation of metabolic models
Flux balance analysis
References

Elements of bioinformatics: Introduction to systems biology
Gaps in metabolic networks

May carry steady-state flux – Blocked – Gap

Elements of bioinformatics: Introduction to systems biology
In silico validation of metabolic models

- Reconstructed genome-scale metabolic networks are very large: hundreds or thousands of reactions and metabolites
- Manual curation is often necessary
- Amount of manual work needed can be reduced with computational methods
- Aims to provide a good basis for further analysis and experiments
- Does not remove the need for experimental verification
Flux Balance Analysis: preliminaries

- Recall that in a steady state, metabolite concentrations are constant over time,

\[ \frac{dx_i}{dt} = \sum_{j=1}^{r} s_{ij}v_j = 0, \text{ for } i = 1, \ldots, n. \]

Stoichiometric model can be given as

\[ S = [S_{II} \ S_{IE}] \]

where \( S_{II} \) describes internal metabolites - internal reactions, and \( S_{IE} \) internal metabolites - exchange reactions.
Flux Balance Analysis (FBA)

- FBA is a framework for investigating the theoretical capabilities of a stoichiometric metabolic model $S$

- Analysis is constrained by
  1. Steady state assumption $Sv = 0$
  2. Thermodynamic constraints: (ir)reversibility of reactions
  3. Limited reaction rates of enzymes: $V_{min} \leq v \leq V_{max}$

- Note that constraints (2) can be included in $V_{min}$ and $V_{max}$. 
• In FBA, we are interested in determining the theoretical maximum (minimum) \textit{yield} of some metabolite, given model

• For instance, we may be interested in finding how efficiently yeast is able to convert sugar into ethanol

Figure: glycolysis in KEGG
Flux Balance Analysis (FBA)

- FBA has applications both in metabolic engineering and metabolic reconstruction
- Metabolic engineering: find out possible reactions (pathways) to insert or delete
- Metabolic reconstruction: validate the reconstruction given observed metabolic phenotype
Formulating an FBA problem

- We formulate an FBA problem by specifying parameters $c$ in the optimization function $Z$,

$$Z = \sum_{i=1}^{r} c_i v_i.$$

- Examples:
  - Set $c_i = 1$ if reaction $i$ produces “target” metabolite, and $c_i = 0$ otherwise
  - Growth function: maximize production of biomass constituents
  - Energy: maximize ATP (net) production
Solving an FBA problem

- Given a model $S$, we then seek to find the maximum of $Z$ while respecting the FBA constraints,

$$\max_v Z = \max_v \sum_{i=1}^r c_i v_i$$

such that

1. $S v = 0$
2. $V_{min} \leq v \leq V_{max}$

- (We could also replace $\max$ with $\min$.)
- This is a linear program, having a linear objective function and linear constraints
Solving a linear program

- General linear program formulation:

\[
\max_{x_i} \sum_i c_i x_i \quad \text{such that} \quad Ax \leq b
\]

- Algorithms: simplex (worst-case exponential time), interior point methods (polynomial)
- Matlab solver: linprog (Statistical Toolbox)
- Many solvers around, efficiency with (very) large models varies
Linear programs

- Linear constraints define a convex polyhedron (feasible region)
- If the feasible region is empty, the problem is infeasible.
- Unbounded feasible region (in direction of objective function): no optimal solution
- Given a linear objective function, where can you find the maximum value?
Flux Balance Analysis: example

- Let's take our running example...
- Unconstrained uptake (exchange) reactions for NADP$^+$ ($r_{10}$), NADPH and H$_2$O (not drawn)
- Constrained uptake for $\alpha$G6P, $0 \leq v_8 \leq 1$
- Objective: production of X5P ($v_9$)

$$c = (0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0)$$
Flux Balance Analysis: example

- **Flux Balance Analysis: preliminaries**
- **Flux Balance Analysis (FBA)**
- **Formulating an FBA problem**
- **Solving an FBA problem**
- **Linear programs**
- **Flux Balance Analysis: example**
- **FBA validation of a reconstruction**

### Flux balance analysis

<table>
<thead>
<tr>
<th>Metabolite</th>
<th>r1</th>
<th>r2</th>
<th>r3</th>
<th>r4</th>
<th>r5</th>
<th>r6</th>
<th>r7</th>
<th>r8</th>
<th>r9</th>
<th>r10</th>
<th>r11</th>
<th>r12</th>
</tr>
</thead>
<tbody>
<tr>
<td>βG6P</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>αG6P</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>βF6P</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6PGL</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6PG</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R5P</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>X5P</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NADP⁺</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NADPH</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>H₂O</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Flux Balance Analysis: example

- Solve the linear program

$$\max_v \sum_{i}^{r} c_i v_i = \max v_9 \quad \text{subject to}$$

$$\sum_{i=1}^{r} s_{ij} v_i = 0 \quad \text{for all } j = 1, \ldots, 10$$

$$0 \leq v_8 \leq 1$$

- Hint: Matlab’s linprog offers nice convenience functions for specifying equality constraints and bounds
Flux Balance Analysis: example

- Figure gives one possible solution (flux assignment \( v \))
- Reaction \( r_7 \) (red) operates in backward direction
- Uptake of \( \text{NADP}^+ \) \( v_{10} = 2v_8 = 2 \)
- How many solutions (different flux assignments) are there for this problem?
FBA validation of a reconstruction

- Check if it is possible to produce metabolites that the organism is known to produce
  - Maximize production of each such metabolite at time
  - Make sure max. production is above zero

- To check biomass production (growth), add a reaction to the model with stoichiometry corresponding to biomass composition
FBA validation of a reconstruction

- If a maximum yield of some metabolite is lower than measured → missing pathway

- Iterative process: find metabolite that cannot be produced, fix the problem by changing the model, try again
FBA validation of a reconstruction

- FBA gives the maximum flux given stoichiometry only, i.e., not constrained by regulation or kinetics
- In particular, assignment of internal fluxes on alternative pathways can be arbitrary (of course subject to problem constraints)
Further studying

- Computational methods of systems biology course, III period
- M. Durot, P.-Y. Bourguignon, and V. Schachter:  
- E. Pitkänen, A. Rantanen, J. Rousu and E. Ukkonen:  
- E. Pitkänen, J. Rousu and E. Ukkonen:  