Flux Balance Analysis
Gapless metabolic reconstruction

Esa Pitkänen

27.3.2009

Metabolic Modeling, spring 2009

MBI Programme

Department of Computer Science

University of Helsinki

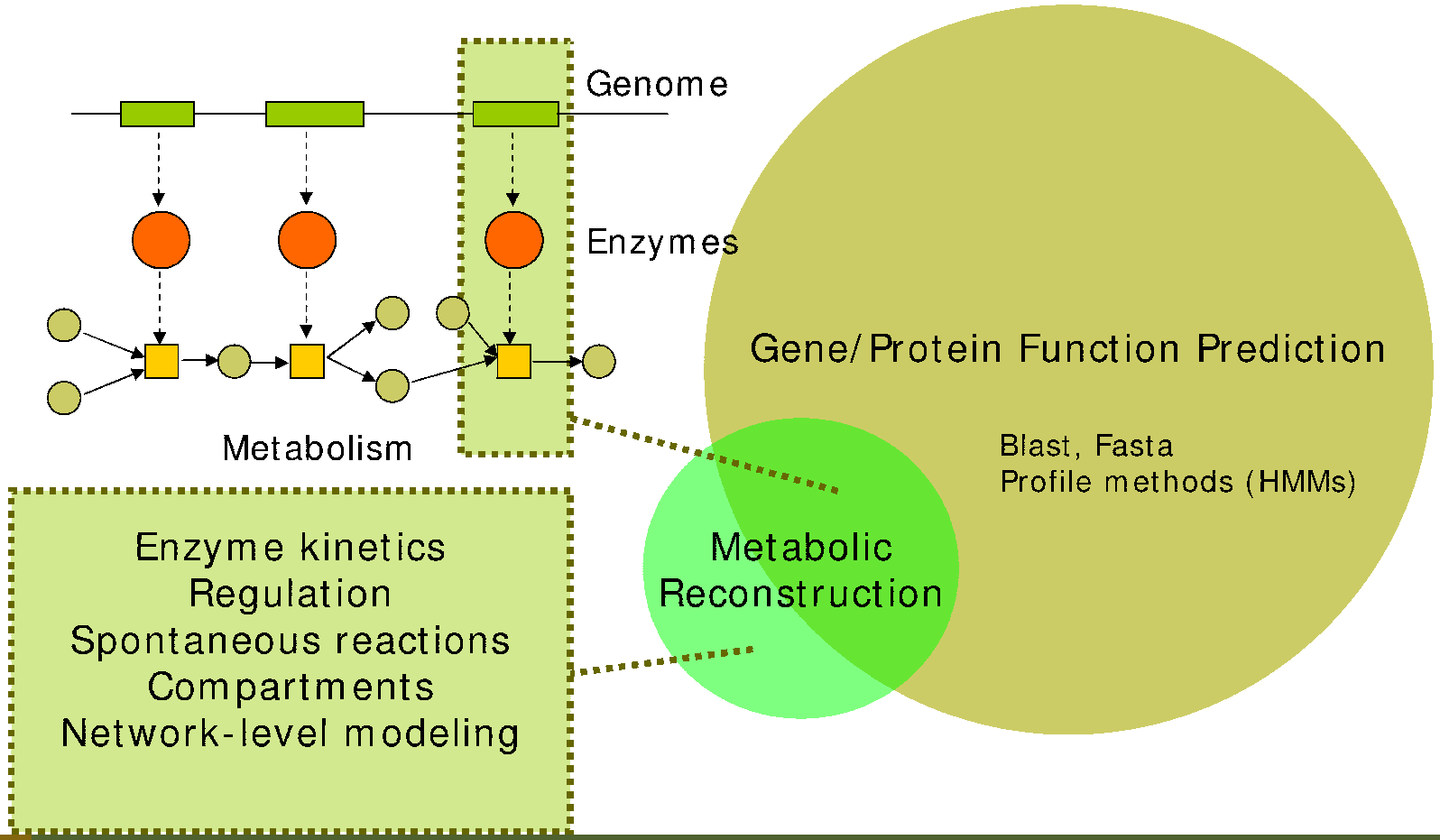
Topics today

- Metabolic reconstruction (revisited)
- In silico validation of reconstructed models
  - Flux Balance Analysis (FBA)
- Gapless metabolic reconstruction

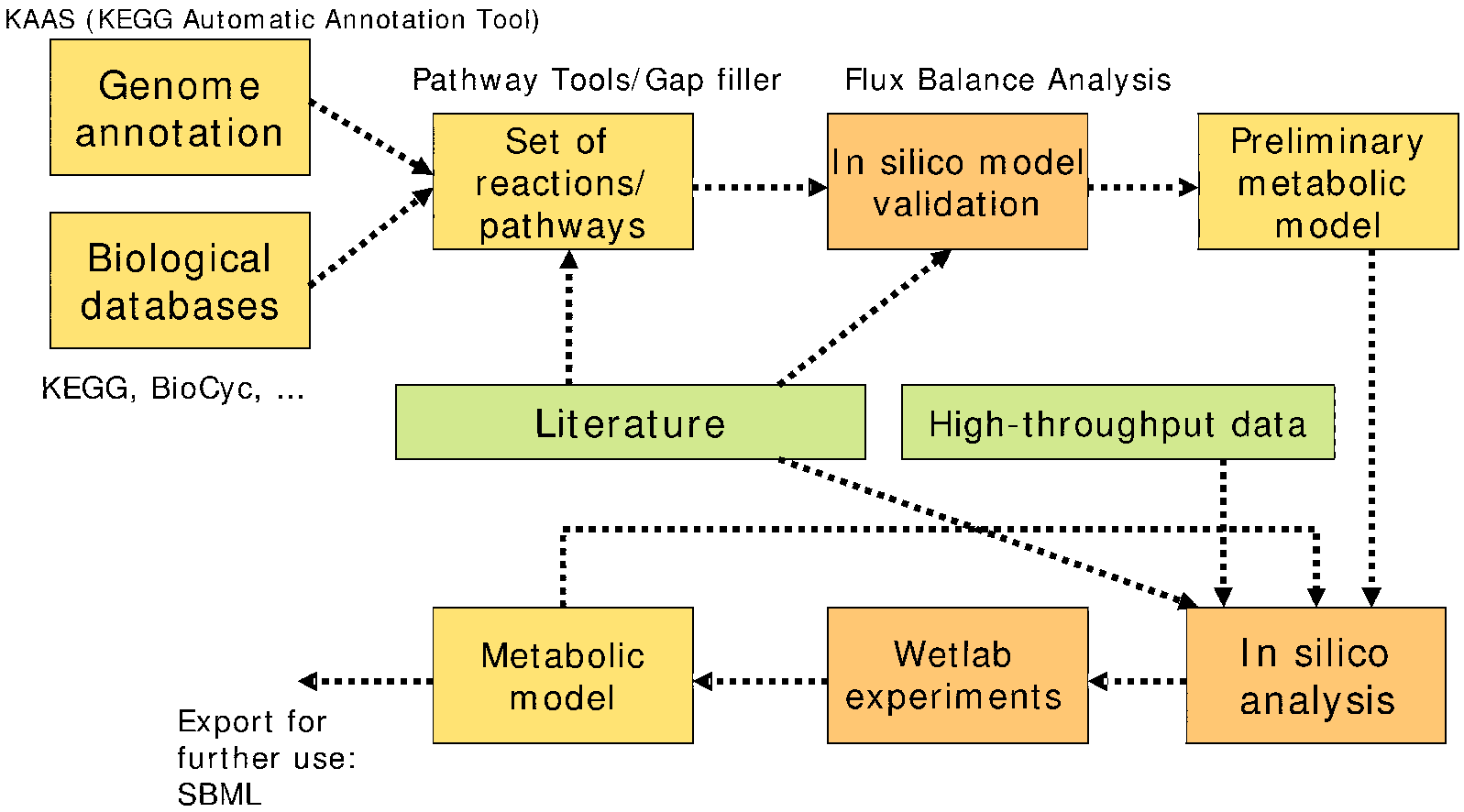
Goals of this lecture

- Introduce two methods for metabolic network analysis
  - FBA (established)
  - Gapless reconstruction (recent work)
- Discuss (integer) linear programming (on a brief level only)
- Discuss some of the many challenges of metabolic modeling
  - Is it possible to achieve useful results with simple models such as stoichiometric models

Metabolic reconstruction (revisited)



Reconstruction process



Read more: Puchałka et al., Genome-Scale Reconstruction and Analysis of the Pseudomonas putida KT2440 Metabolic Network Facilitates Applications in Biotechnology. PLoS Computational Biology 2008.

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,

$$\frac{dX_i}{dt} = \sum_{j=1}^{r} s_{ij} v_j = 0, \ \texttt{for}\ i = 1, \ldots, n,$$

and that a stoichiometric model is given by

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

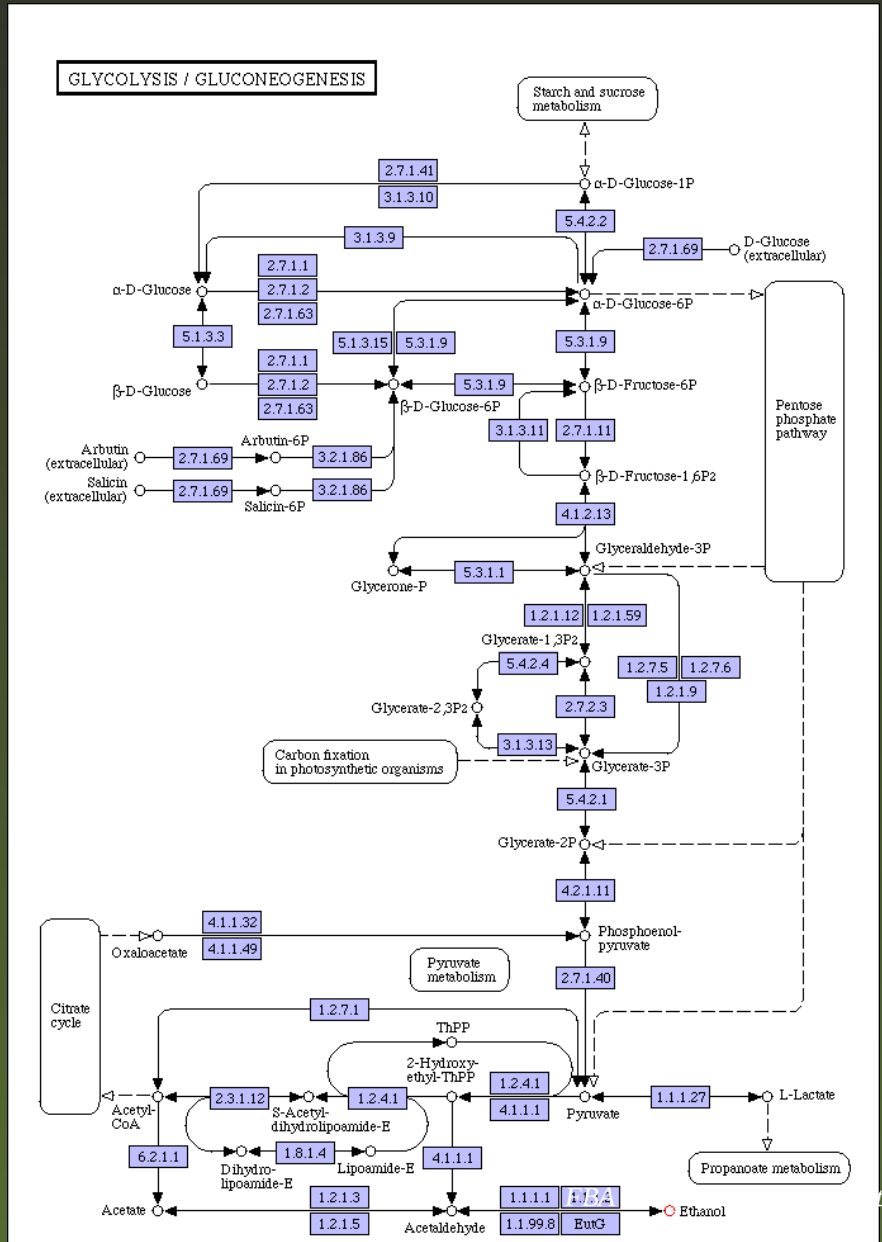
where S_{II} describes internal metabolites - internal reactions, and S_{IE} internal metabolites - exchange reactions.

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 $\mathbf{S}v = 0$
  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}.

Flux Balance Analysis (FBA)

- In FBA, we are interested in determining the theoretical maximum (minimum) *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,

$$\text{(1)} \qquad \max_v Z = \max_v \sum_{i=1}^{r} c_i v_i \quad \texttt{such that}$$

$$\text{(2)} \qquad \mathbf{S}v = 0$$

$$\text{(3)} \qquad V_{min} \leq v \leq V_{max}$$

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

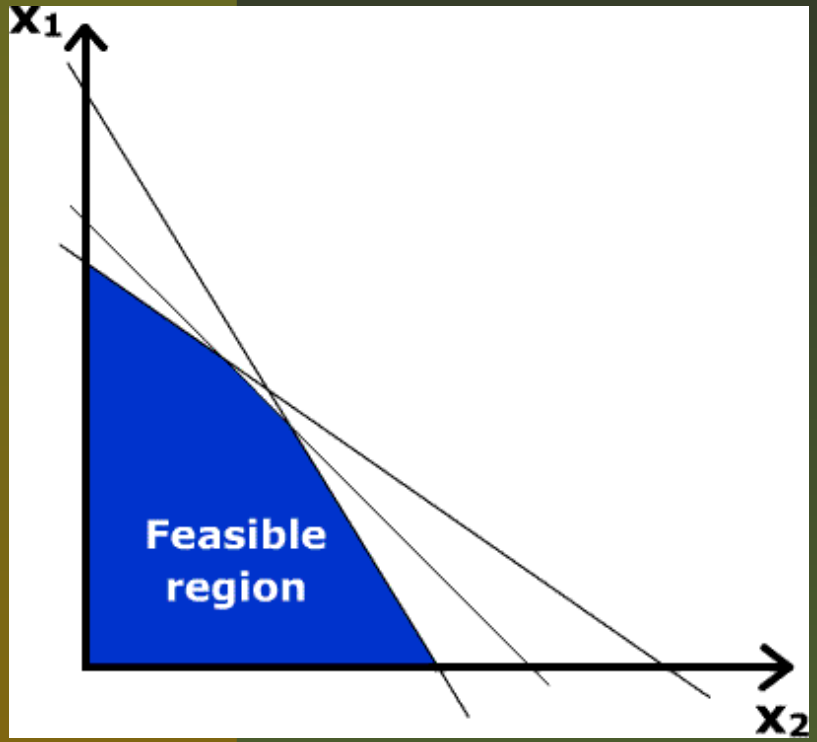
Solving a linear program

- General linear program formulation:

$$\max_{x_i} \sum_i c_i x_i \quad \texttt{such that}$$
$$Ax \leq b$$

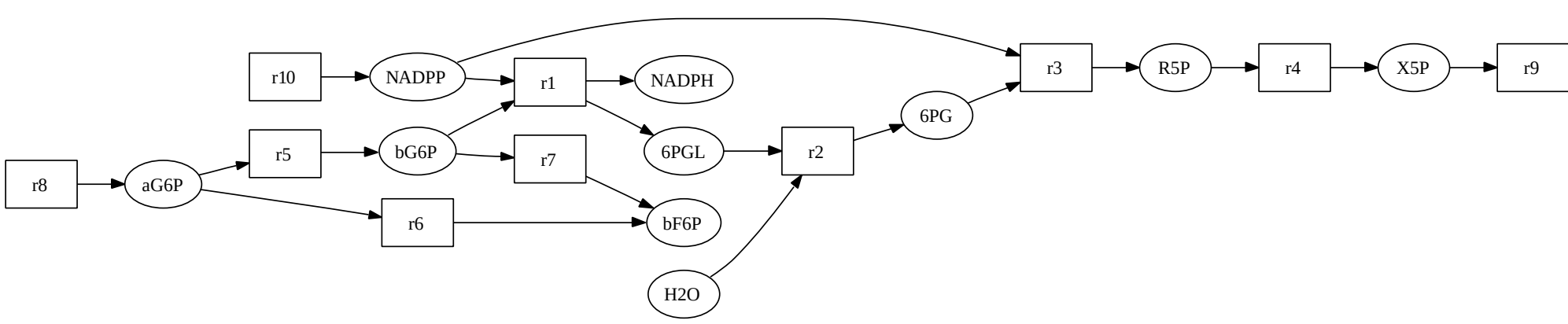
- Algorithms: simplex (worst-case exponential time), interior point methods (polynomial)
- Matlab solver: linprog (Statistical Toolbox)
- Many solvers around, efficiency with (very) large models varies

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?

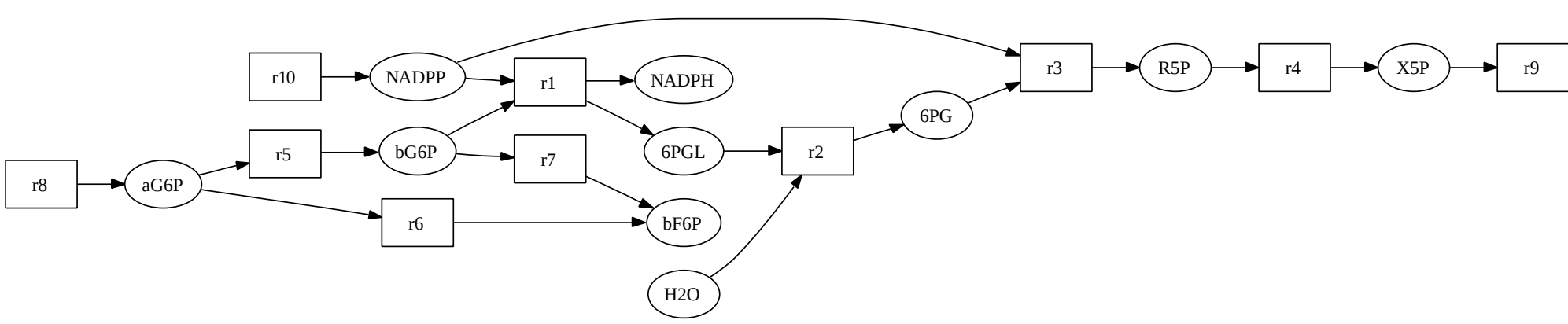
Flux Balance Analysis: example



- Let's take the course's running example...
- Unconstrained uptake (exchange) reactions for $NADP^+$ (r_{10}), NADPH and H_2O (not drawn)
- Constrained uptake for αG6P, $0 \leq v_8 \leq 1$
- Objective: production of X5P (v_9)

$$c = (0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0)$$

Flux Balance Analysis: example



	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}	r_{11}	r_{12}
βG6P	-1	0	0	0	1	0	-1	0	0	0	0	0
αG6P	0	0	0	0	-1	-1	0	1	0	0	0	0
βF6P	0	0	0	0	0	1	1	0	0	0	0	0
6PGL	1	-1	0	0	0	0	0	0	0	0	0	0
6PG	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
$NADP^+$	-1	0	-1	0	0	0	0	0	0	1	0	0
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

Flux Balance Analysis: example

- Solve the linear program

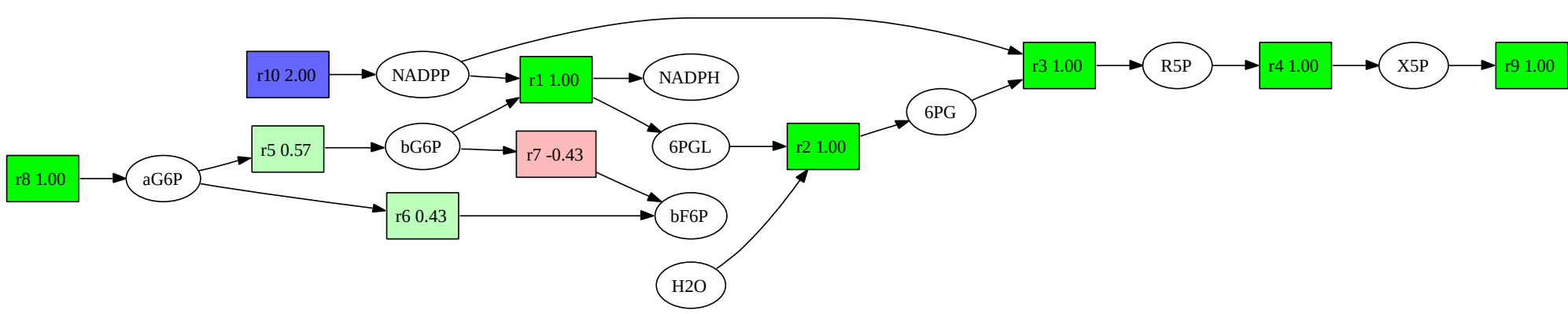
$$\max_v \sum_i^r c_i v_i = \max v_9 \quad \text{subject to}$$

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

$$0 \leq v_8 \leq 1$$

- Hint: Matlab's linprog offers nice convenience functions for specifying equality constraints and bounds

Flux Balance Analysis: example



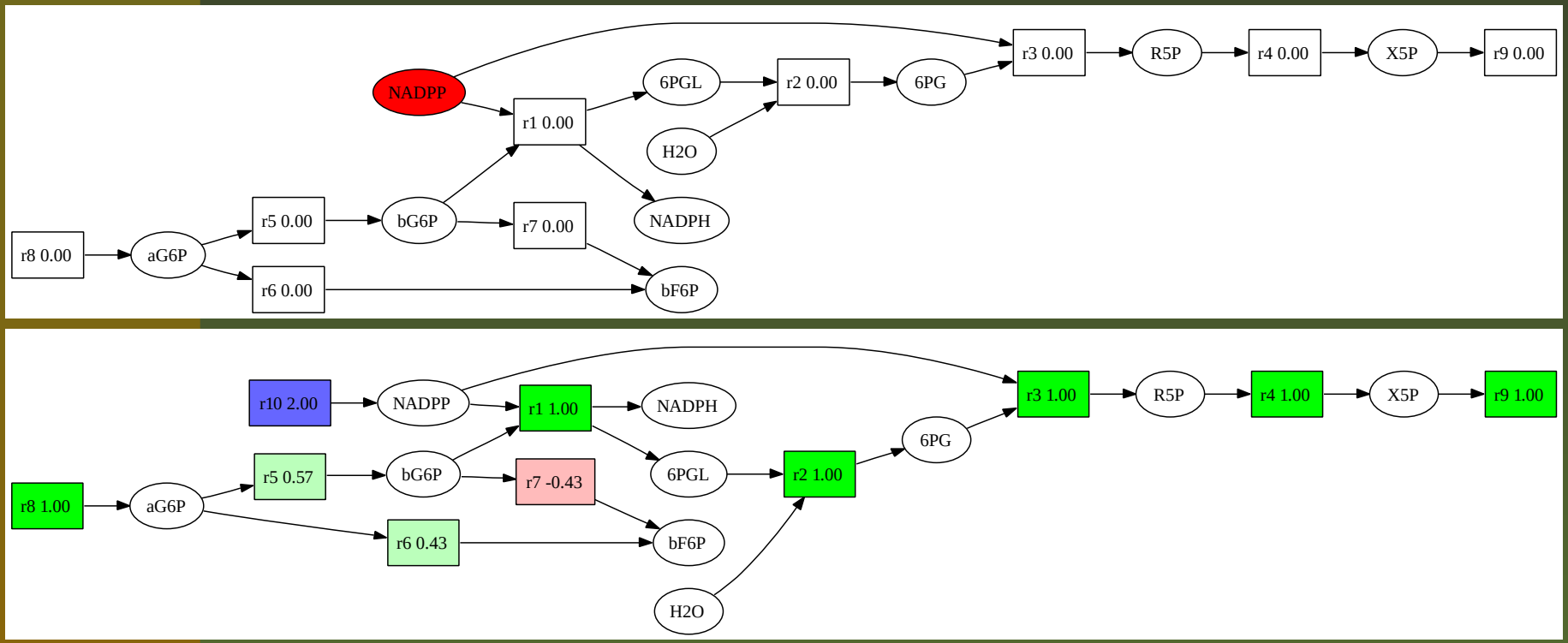
- Figure gives one possible solution (flux assignment v)
- Reaction r_7 (red) operates in backward direction
- Uptake of NADP$^+$ $v_{10} = 2v_8 = 2$
- How many solutions (different flux assignments) are there for this problem?

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

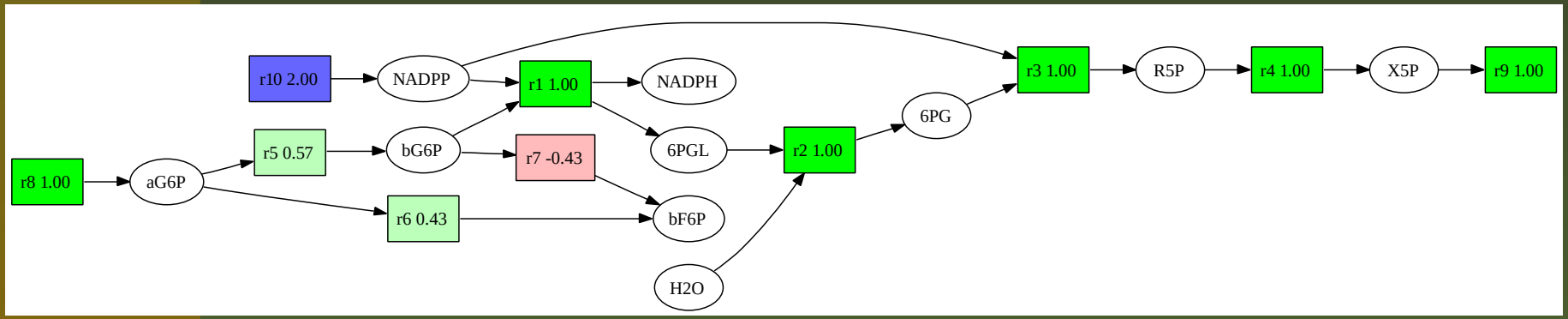
FBA validation of a reconstruction

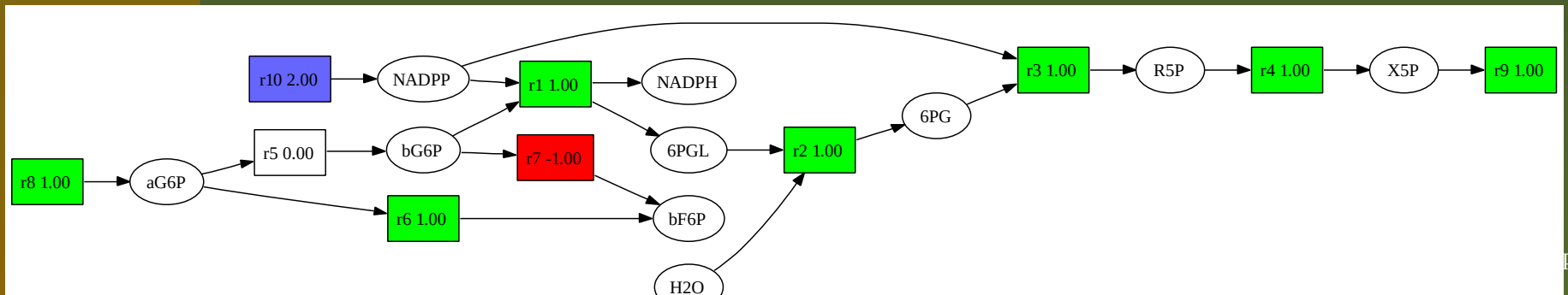
- If a maximum yield of some metabolite is lower than measured $\longrightarrow$ missing pathway
- Iterative process: find metabolite that cannot be produced, fix the problem by changing the model, try again



FBA validation of a reconstruction

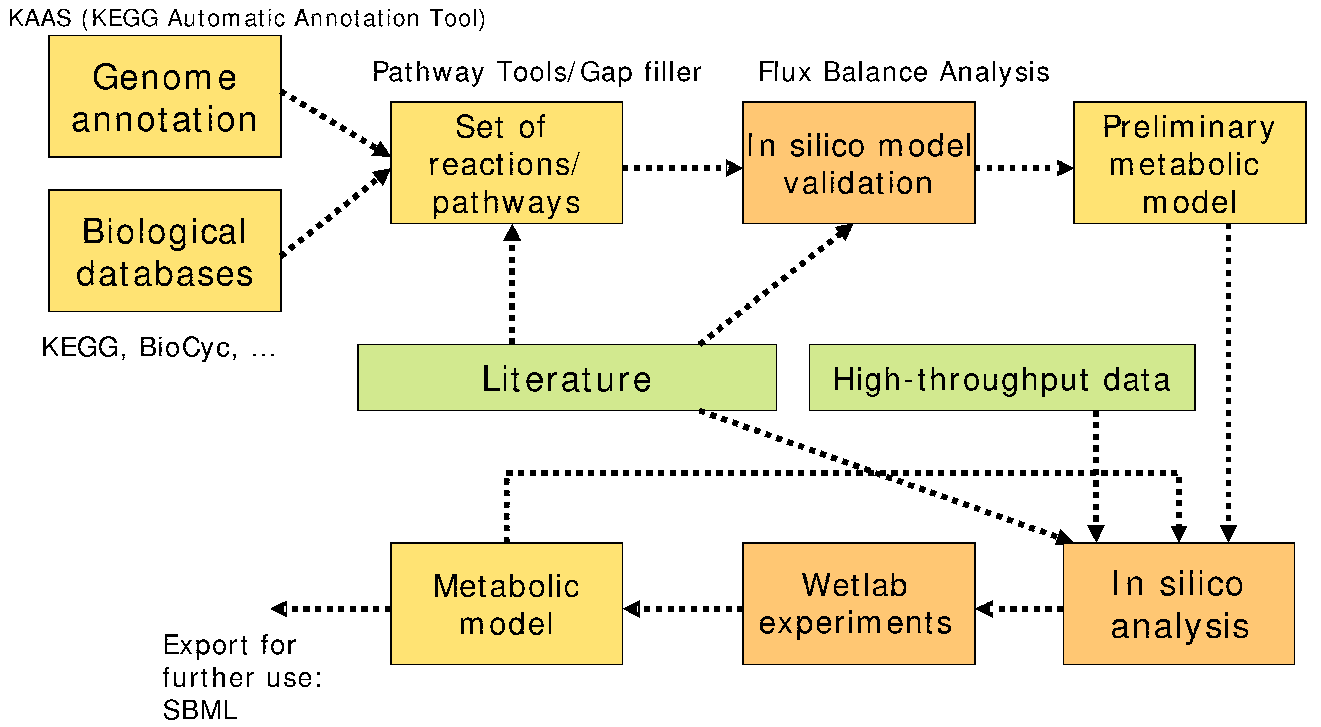
- FBA gives the maximum flux given stoichiometry only, i.e., not constrained by regulation or kinetics
- In particular, assignment of internal fluxes on alternative pathways can be arbitrary (of course subject to problem constraints)





Gapless metabolic reconstruction

- Motivation: Current workflows choose “good” reactions by sequence evidence, fix problems later manually or automatically



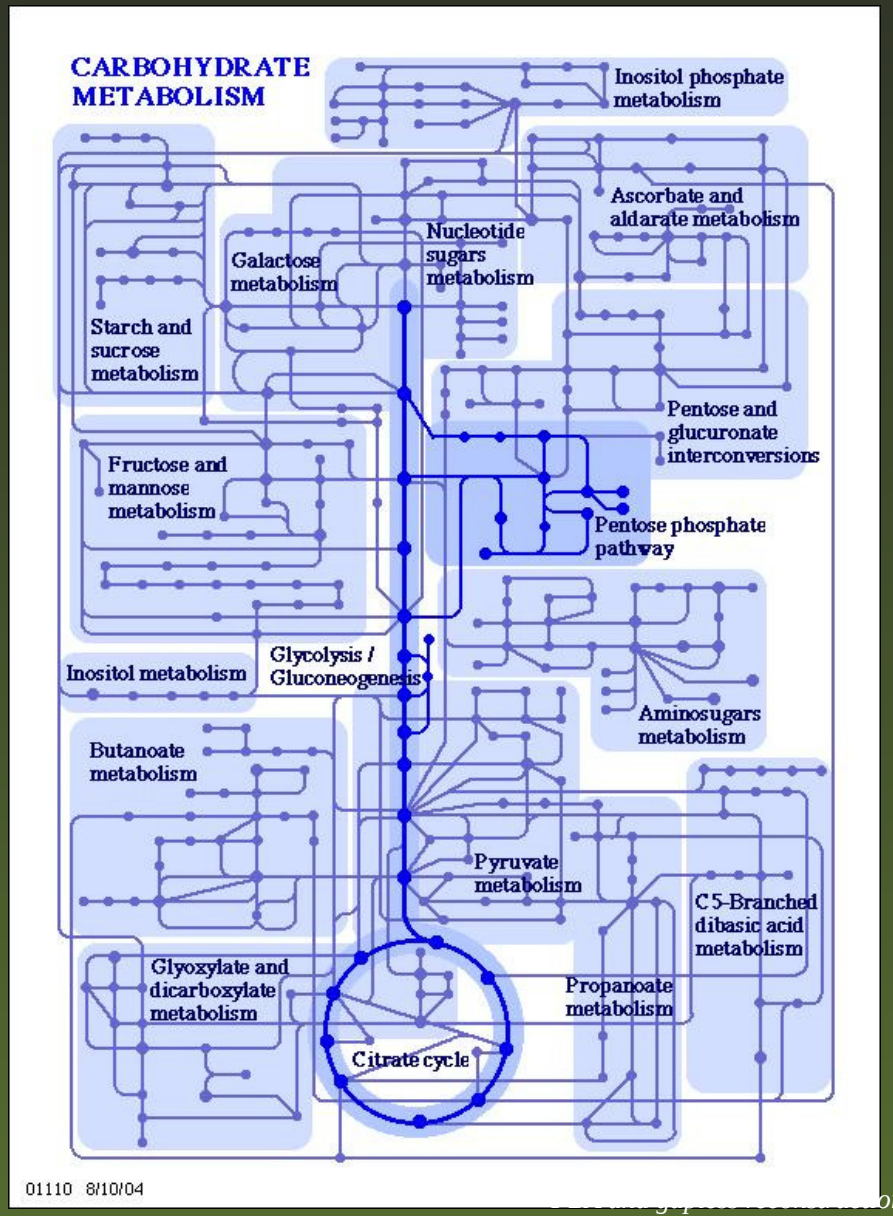
Read more: Puchałka et al., Genome-Scale Reconstruction and Analysis of the Pseudomonas putida KT2440 Metabolic Network Facilitates Applications in Biotechnology. PLoS Computational Biology 2008.

What is a (reaction) gap?

- A reaction in the metabolic network that “should be there” but is not
  - Sequencing failure
  - Correct ortholog not found
  - Correct ortholog found but misannotated
  - Correct reaction not in metabolic database(s) (previously unknown function)
- In the prediction context, a gap is a *false negative* prediction

Gaps in metabolic models

- Central metabolism usually well covered (well conserved!)
  - Glycolysis
  - TCA cycle
  - Pentose phosphate pathway
  - Amino acid pathways
- Lots of problems with other parts even in commonly used models

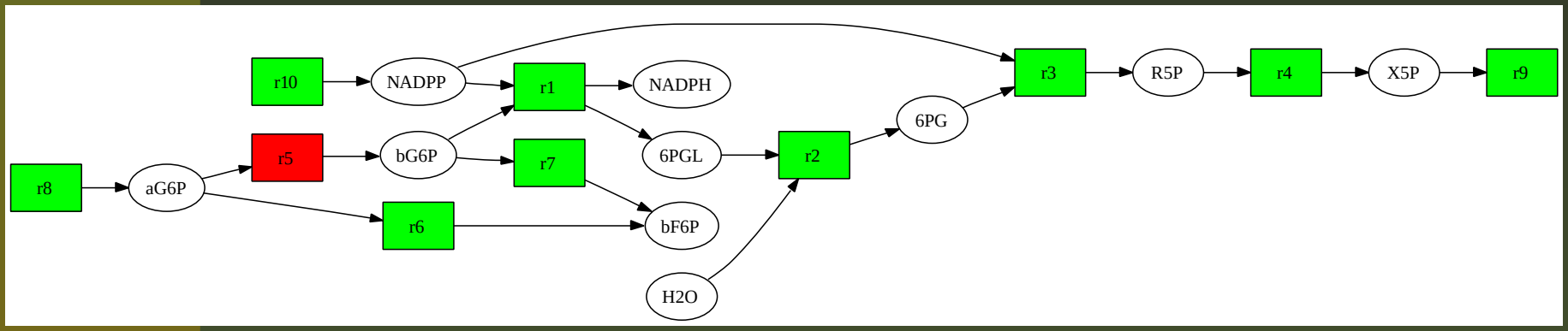


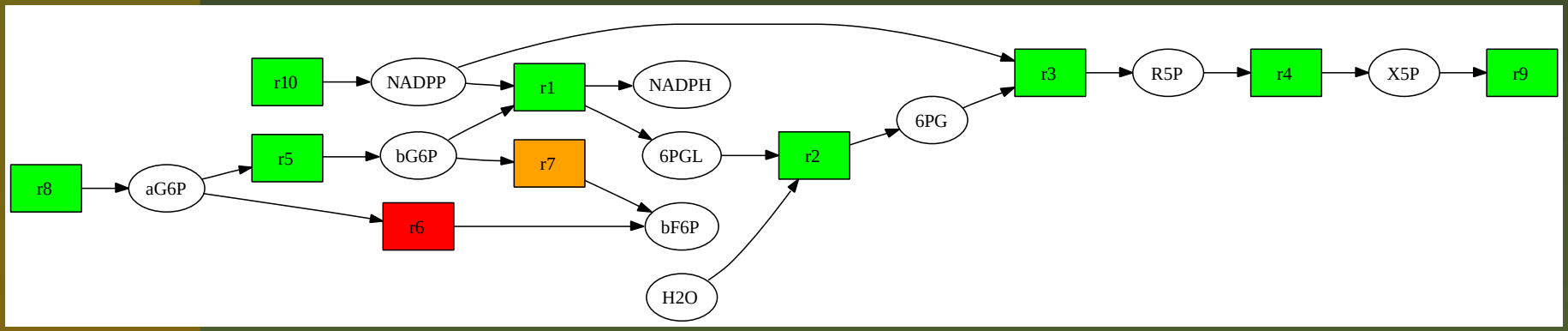
Why bother with gaps?

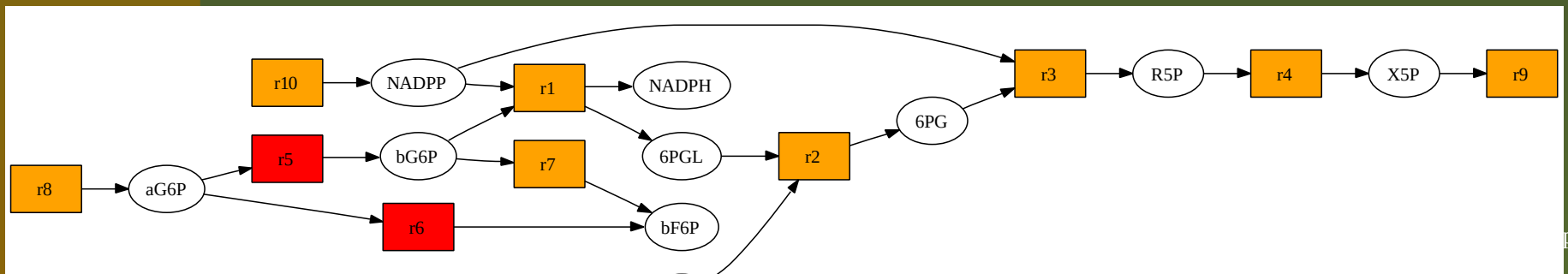
- Gaps cause problems with both qualitative and quantitive analyses
- Consider FBA for example
  - A single reaction gap can block flux through multiple reactions
  - Particularly problematic with branching pathways
- Ultimately, gaps can lead into false predictions, leading in the worst case to unnecessary experiments
  - (Same applies to *false positives*, i.e., extra reactions)

Effect of gaps

Red: gap, Orange: cannot carry flux, Green: can carry flux

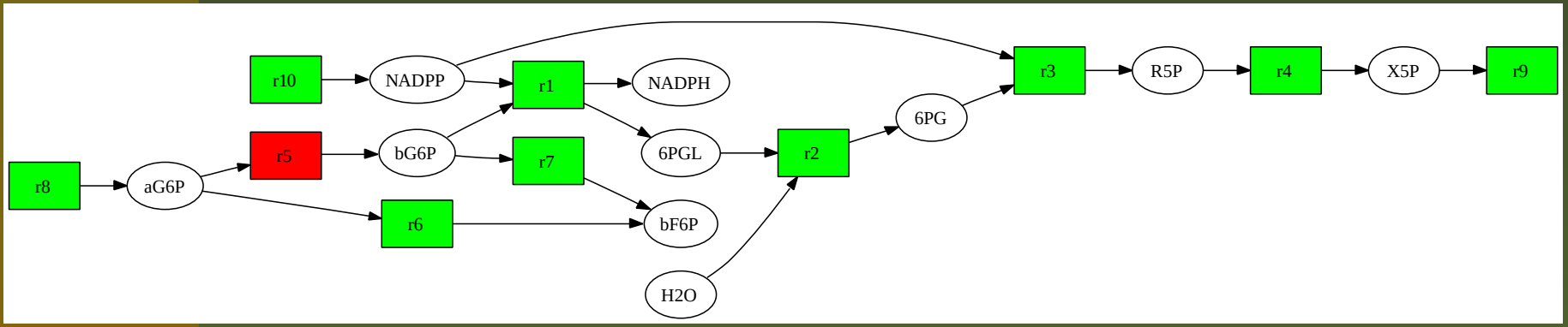






Modeling gaps with AND-OR graphs

- Let A be the set of *input* metabolites and reactions
- Reaction r is *reachable*, iff all its substrates are reachable, or $r \in A$ (AND node)
- Metabolite m is reachable, iff at least one of its producing reactions is reachable, or $m \in A$ (OR node)



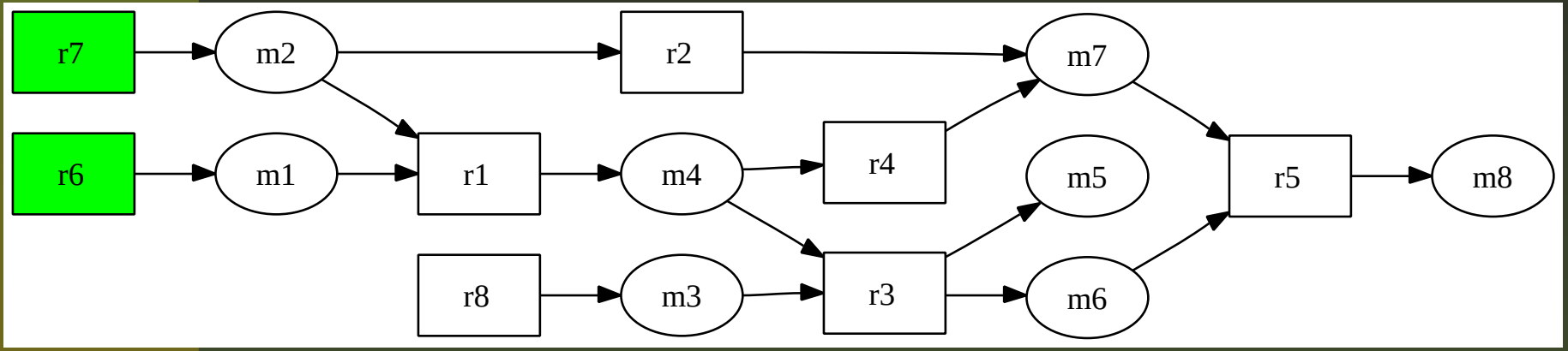
- For example, reaction r_1 is reachable only if both NADPP and βG6P are reachable.

Modeling gaps with AND-OR graph

- Reachable reaction: has substrates that the network is able to produce (under this simple model)
- Reachable metabolite: can be produced by one or more reactions that are able to operate
- A gap in the model: reaction or metabolite not reachable given inputs A.

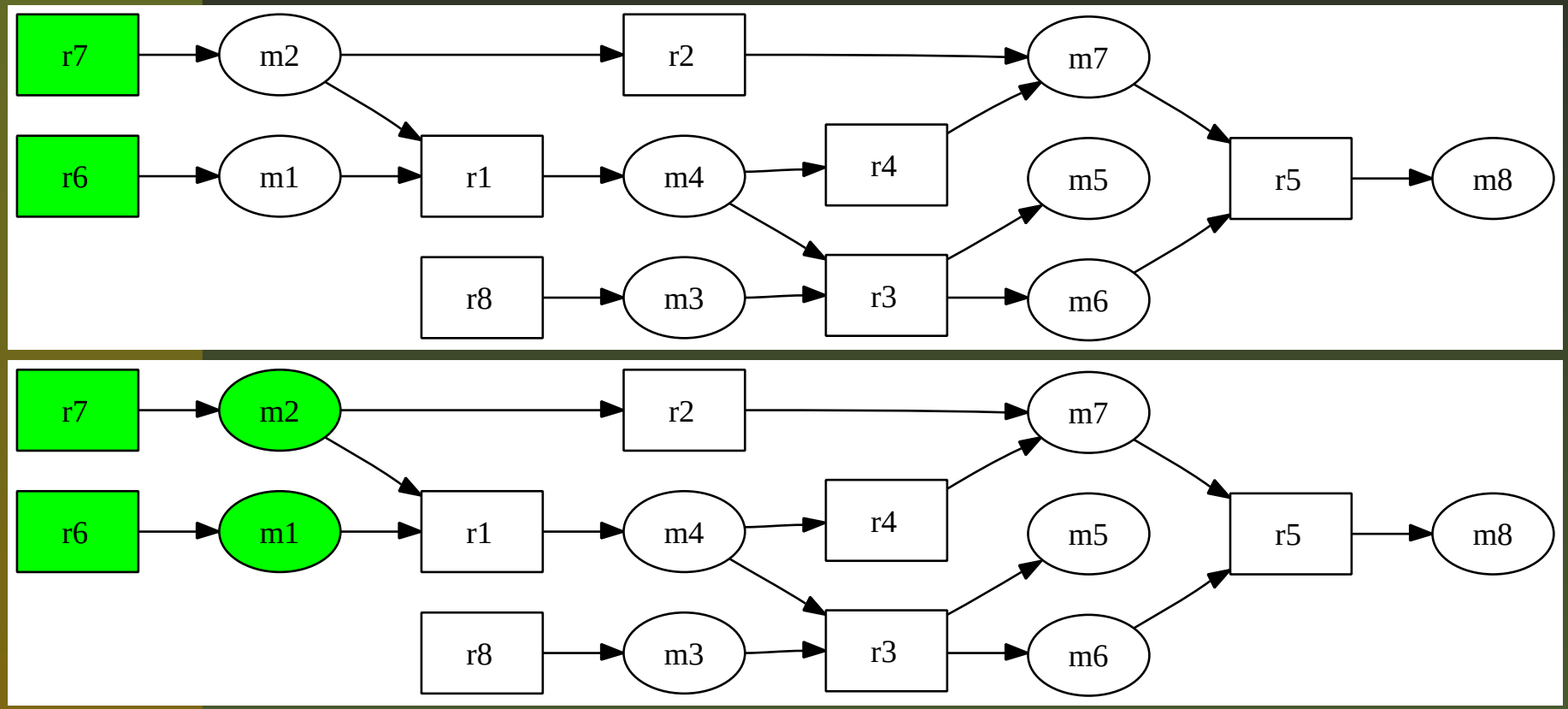
- Let’s take an example...

Reachability in AND-OR graphs 1/7



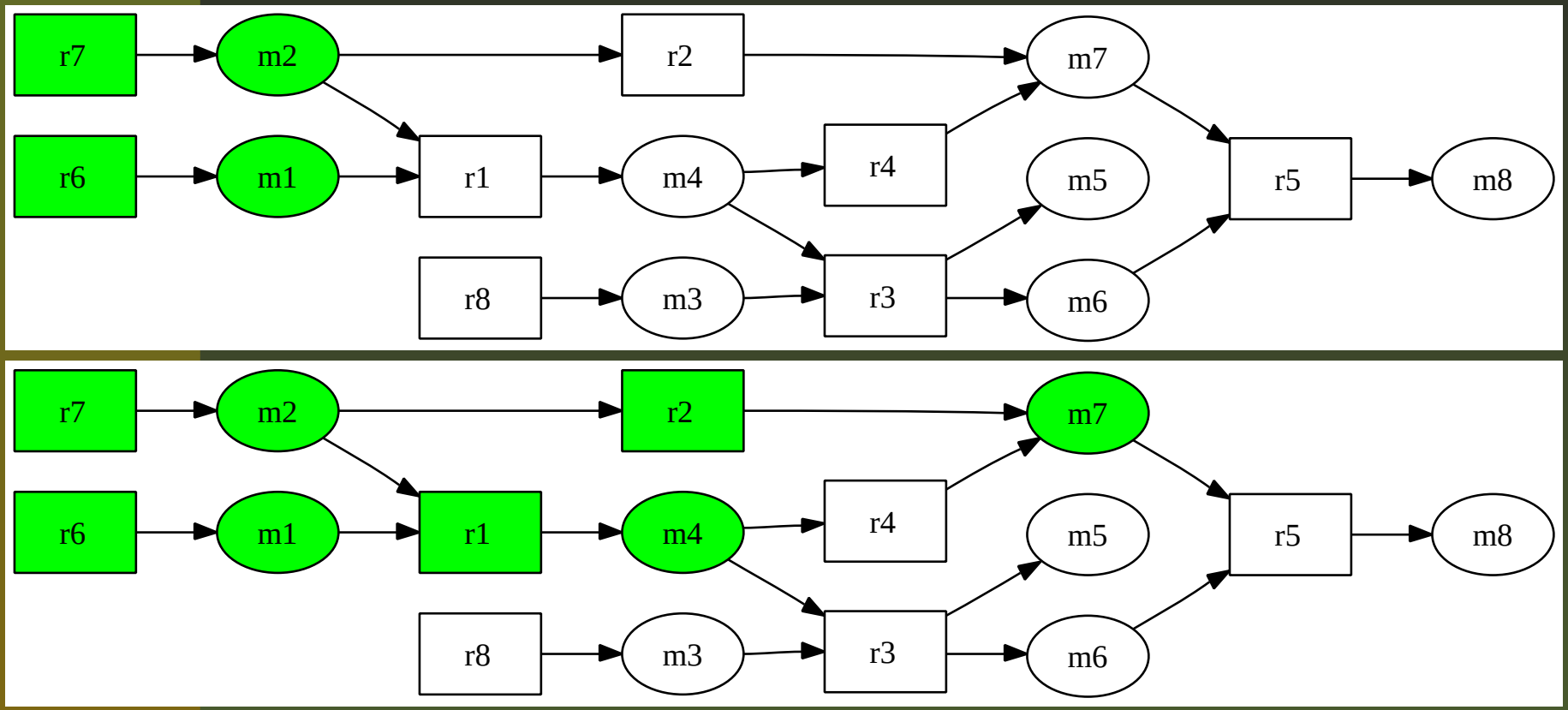
Set $A = \{r_6, r_7\}$. Inputs A used to model system boundaries.

Reachability in AND-OR graphs 2/7



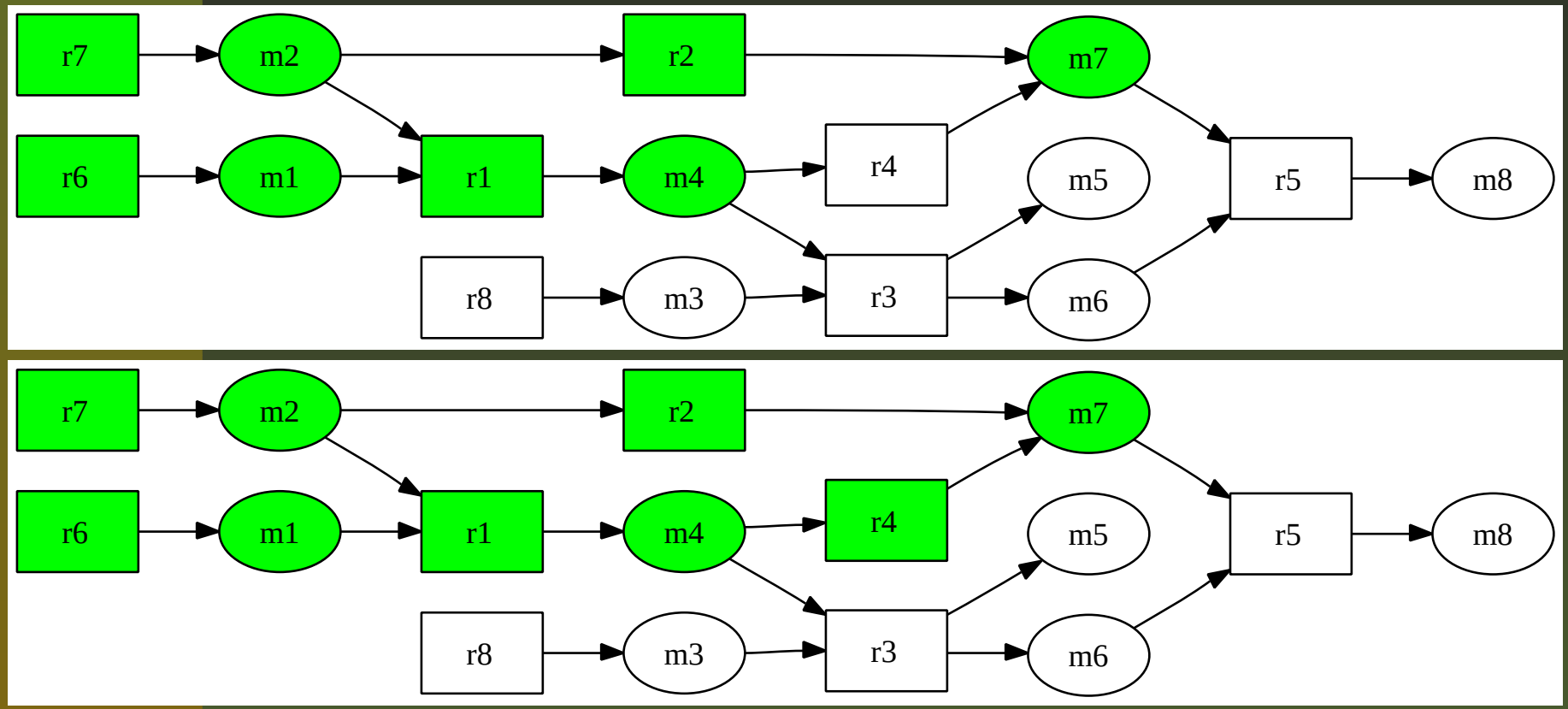
r_6 produces m_1, r_7 produces m_2.

Reachability in AND-OR graphs 3/7



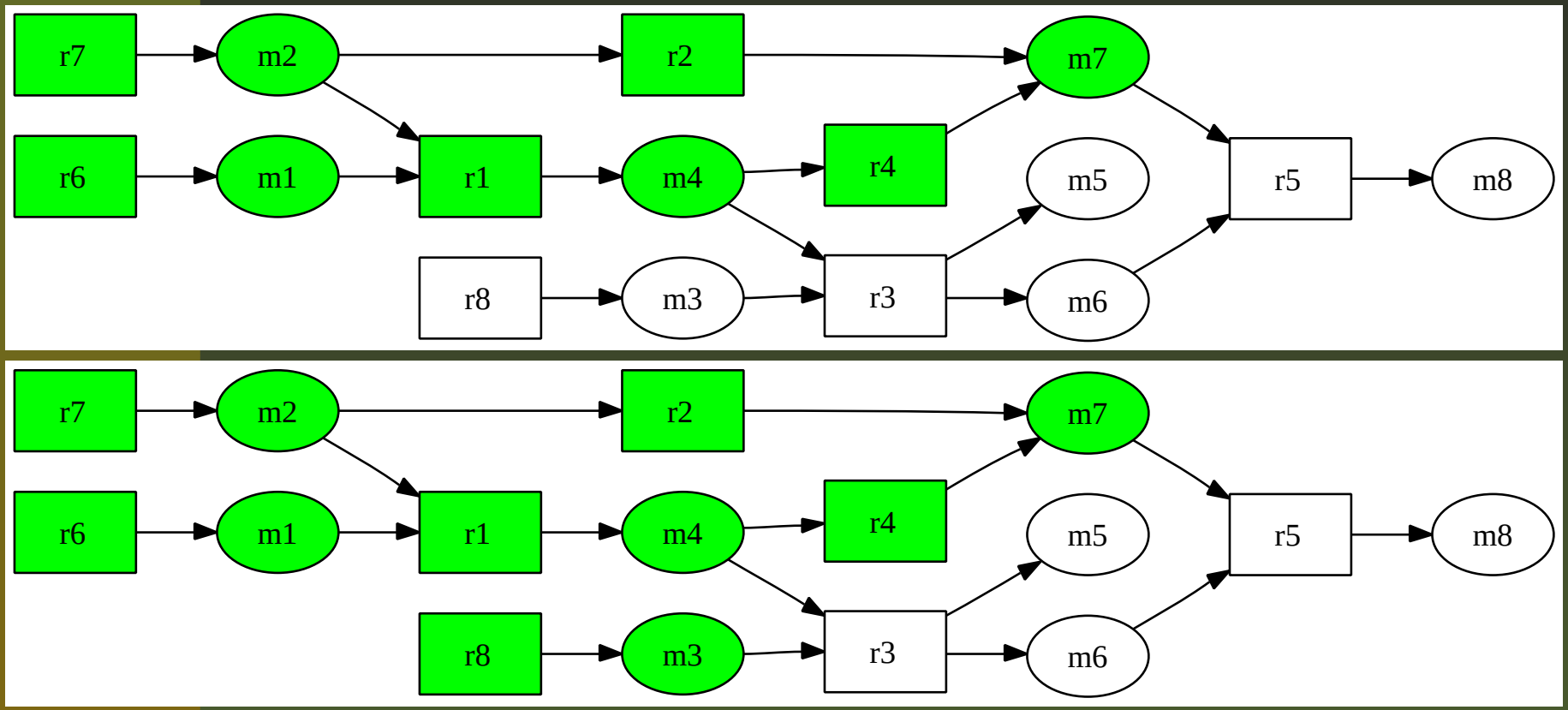
Both r_1 and r_2 have all substrates reachable.

Reachability in AND-OR graphs 4/7



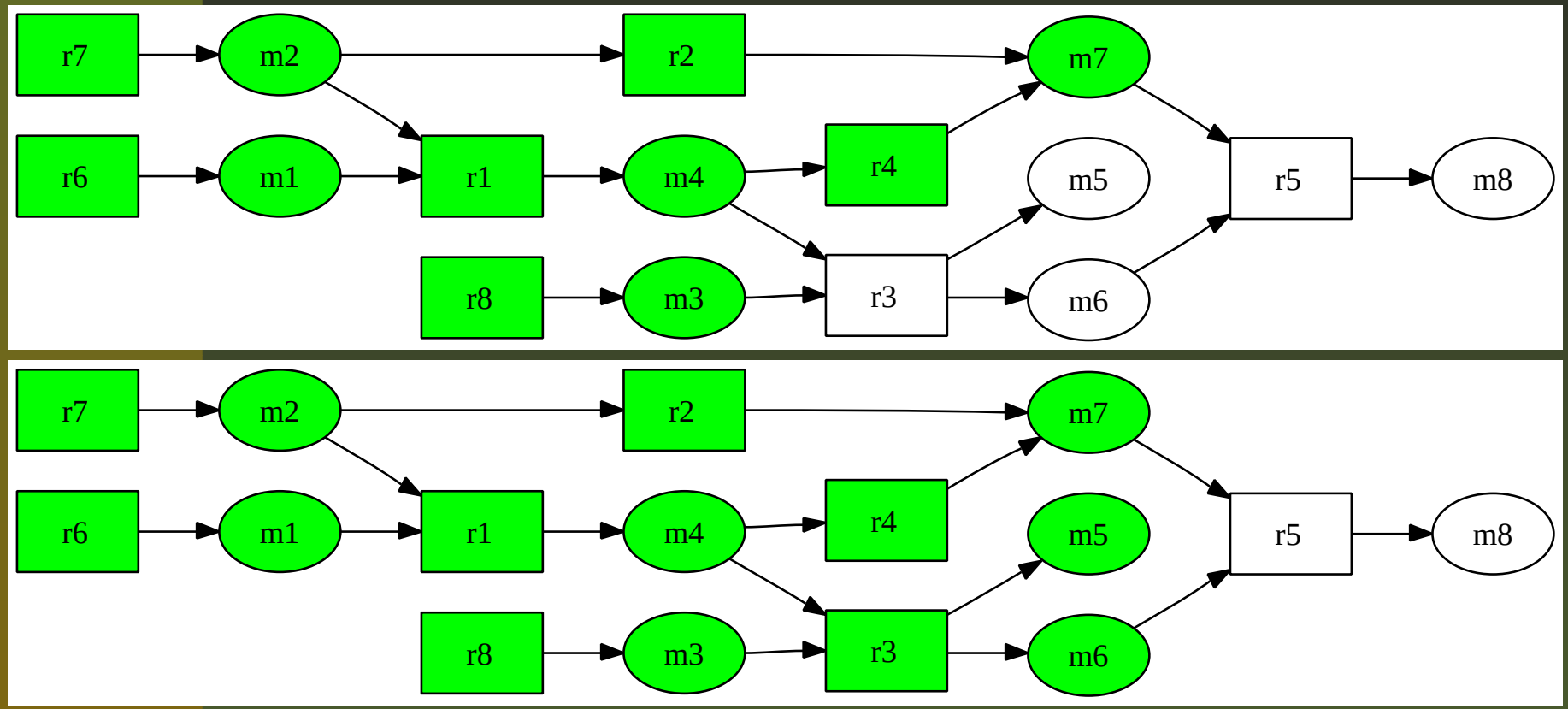
r_4 is reachable. r_3 and r_5 remain unreachable.

Reachability in AND-OR graphs 5/7



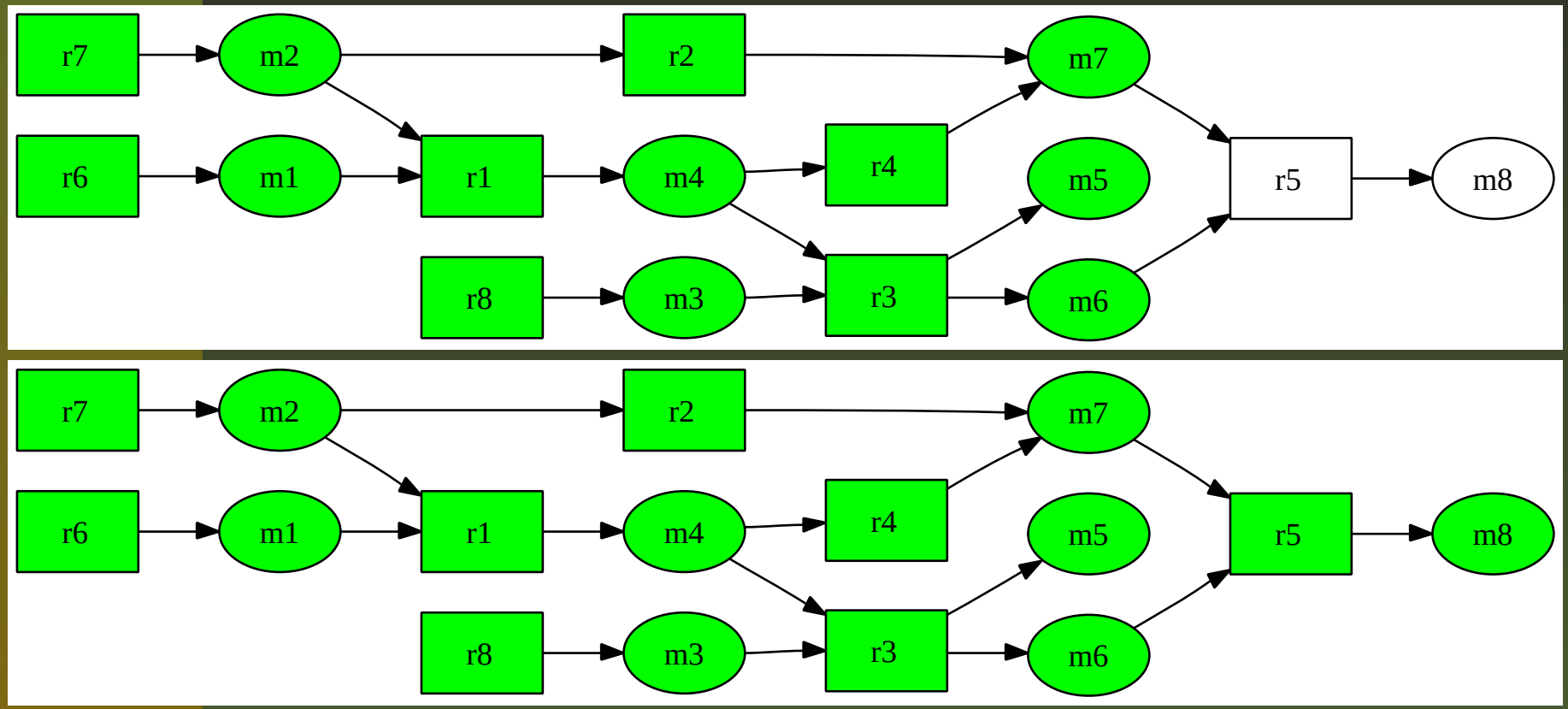
If r_8 is added to the initially reachable reactions, $A = \{r_6, r_7, r_8\}$, m_3 becomes reachable.

Reachability in AND-OR graphs 6/7



r_3, m_5 and m_6 are reached...

Reachability in AND-OR graphs 7/7



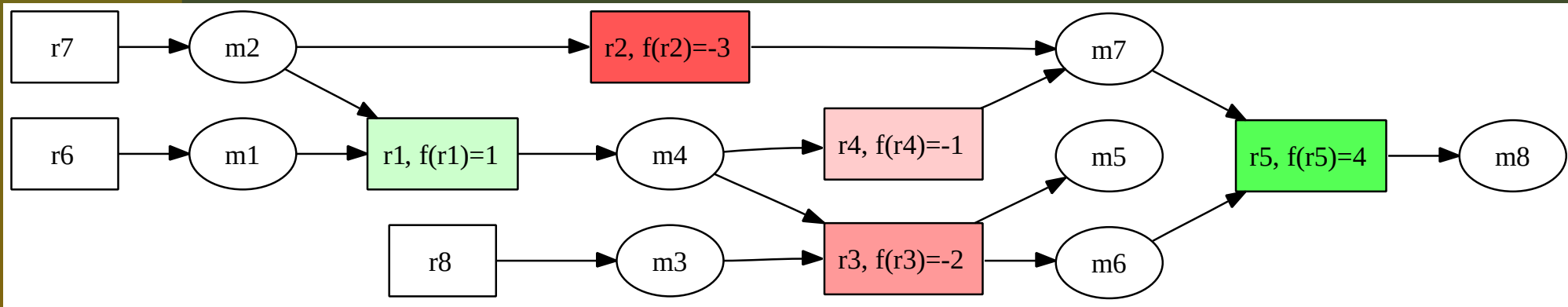
...and finally r_5 and m_8.

Gapless metabolic reconstruction

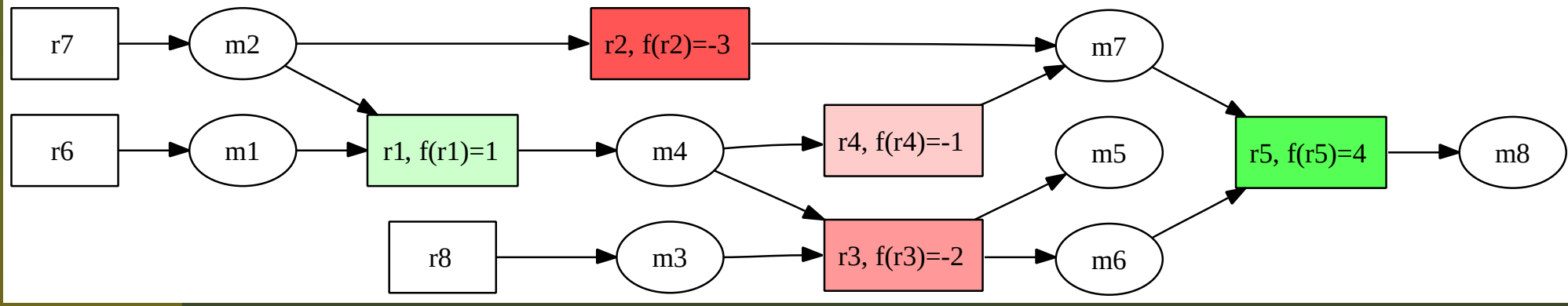
- Input
  - A set of reactions $\mathcal{R}$ (e.g., all reactions in KEGG)
  - Inputs A
  - Score function $f : \mathcal{R} \to \mathbb{R}$.
- Task: find a subset $R \subseteq \mathcal{R}$ such that
  - $F(R) = \sum_{r \in R} f(r)$ is maximized
  - All reactions $r \in R$ are reachable given A in the AND-OR graph induced by R

Score function f

- If $f > 0$, the solution contains all reachable reactions
- If $f < 0$, the solution is empty
- Interesting case: reactions have both positive and negative scores



Example sets R



- $R_1 = \{r_1, r_2, r_3, r_4, r_5\}$, $F(R_1) = 1 - 3 - 2 - 1 + 4 = -1$
- $R_2 = \{r_1, r_2, r_3, r_5\}$, $F(R_2) = 1 - 3 - 2 + 4 = 0$
- $R_3 = \{r_1, r_3, r_4, r_5\}$, $F(R_3) = 1 - 2 - 1 + 4 = 2$
- $R_4 = \{r_1\}$, $F(R_4) = 1$

Reactions in $R_1, \ldots, R_4$ are reachable in the induced graphs.

Establishing connection between genome and reaction scoring

- We would like scores $f(r)$ to reflect the degree of confidence to that the genome codes for an enzyme catalyzing reaction r
- Assume that we have
  - A set of protein sequences from the genome under study
  - An annotated protein sequence database (such as UniProt)
  - A reaction database (such as KEGG or BioCyc)

Establishing connection between genome and reaction scoring

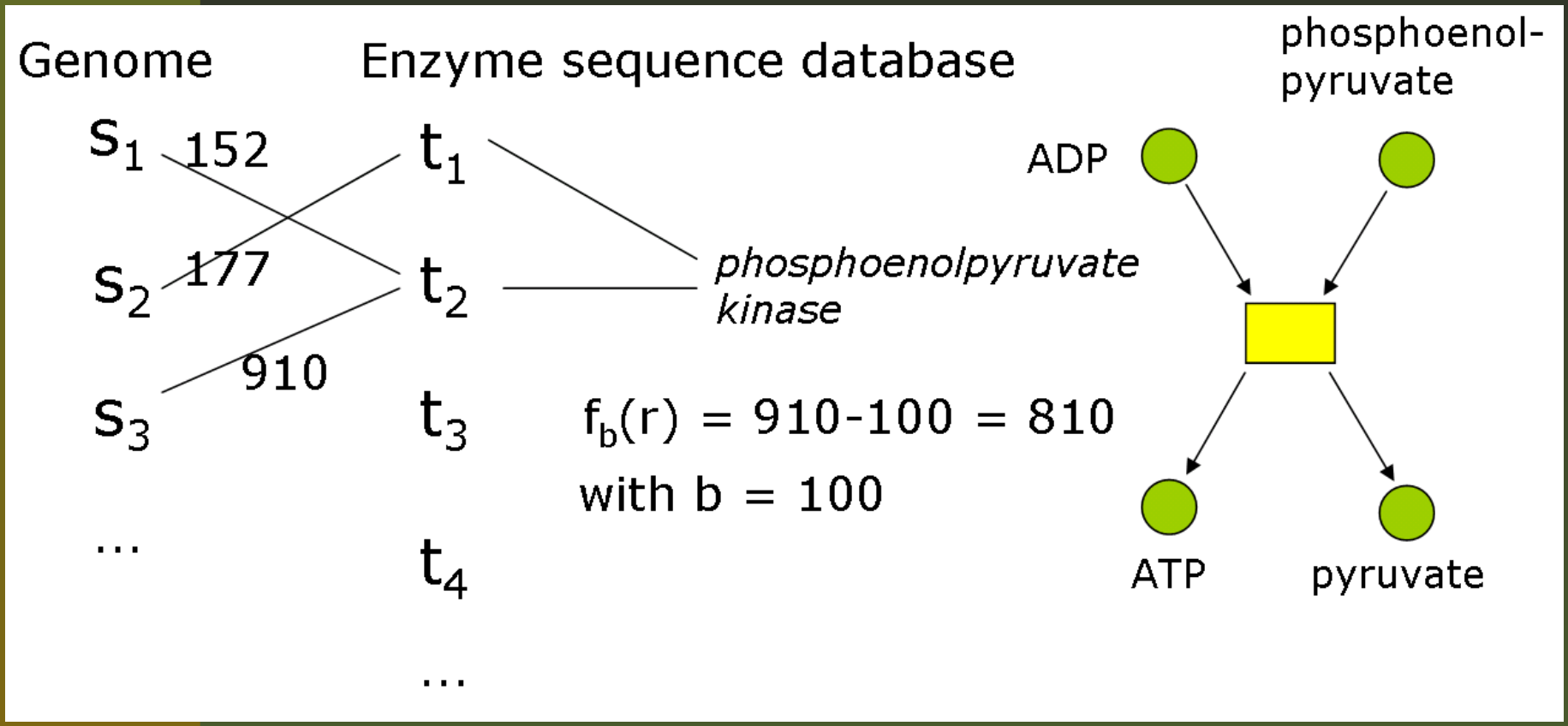
We assign each reaction r score

$$f(r) = \max_{s \in G} \max_{t \in C(r)} B(s,t) - b,$$

where

- G is the set of protein sequences from genome,
- $C(r)$ are the sequences in the database annotated with reaction r,
- $B(s,t)$ is the BLAST score of alignment of s and t and
- $b \in \mathbb{R}$

Establishing connection between genome and reaction scoring



A reaction with a negative score only appear in the solution when it fills a gap!

Solving gapless reconstruction

- Gapless reconstruction can be formulated as a *mixed integer* linear program (MILP)
  - Some variables allowed to only take integer values
- Formulation resembles Flux Balance Analysis
  - We add binary decision variables for each reaction
  - Instead of pure steady state, we allow metabolite net production
  - *Futile cycles* disallowed

Gapless reconstruction as ILP

$$\max_x \sum_{r_i} f(r_i)x_i \text{ such that}$$

$$\begin{aligned}
\frac{1}{N}x_i &\leq v_i \\
v_i &\leq Mx_i \\
\sum_i s_{ij}v_i - t_j &\geq 0 \\
t_j &\geq \alpha \sum_{r_i \in P(m_j)} v_i \\
x_i &\in \{0,1\}
\end{aligned}$$

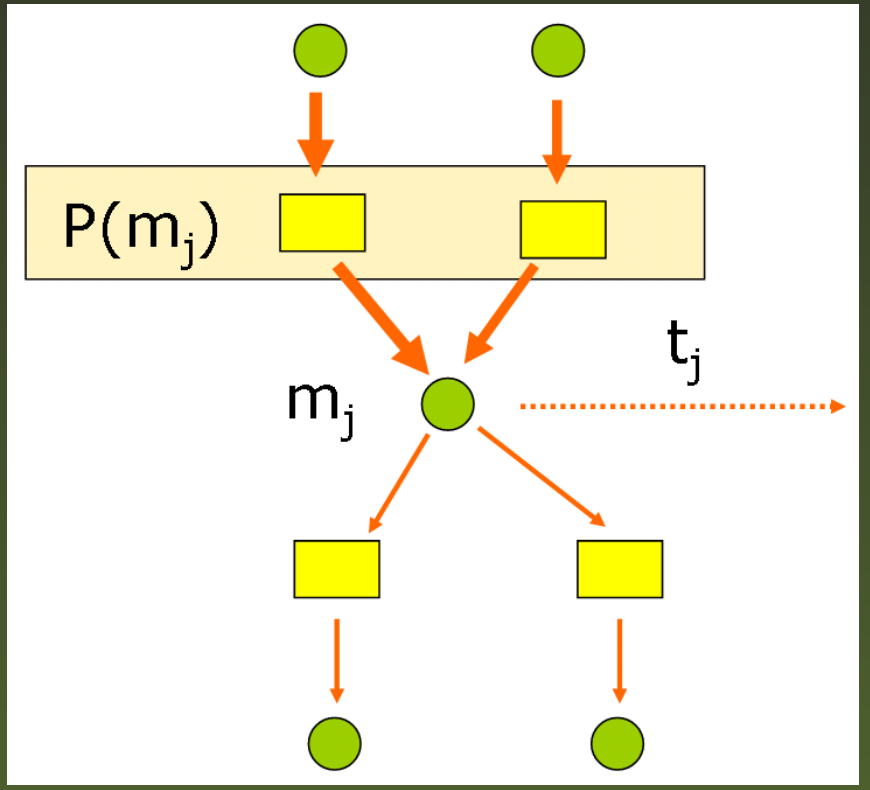
- N, M: large numbers
- First two constraints ensure $x_i = 0 \Leftrightarrow v_i = 0$
- t_j: removes a fraction (α) of flux from the system to disallow futile cycles
- $P(m_j)$: producers of m_j

Gapless reconstruction as an Integer Linear Program

$$\max_x \sum_{r_i} f(r_i)x_i \text{ such that}$$

$$\begin{aligned}
\frac{1}{N}x_i &\leq v_i \\
v_i &\leq Mx_i \\
\sum_i s_{ij}v_i - t_j &\geq 0 \\
t_j &\geq \alpha \sum_{r_i \in P(m_j)} v_i
\end{aligned}$$

$$x_i \in \{0,1\}$$

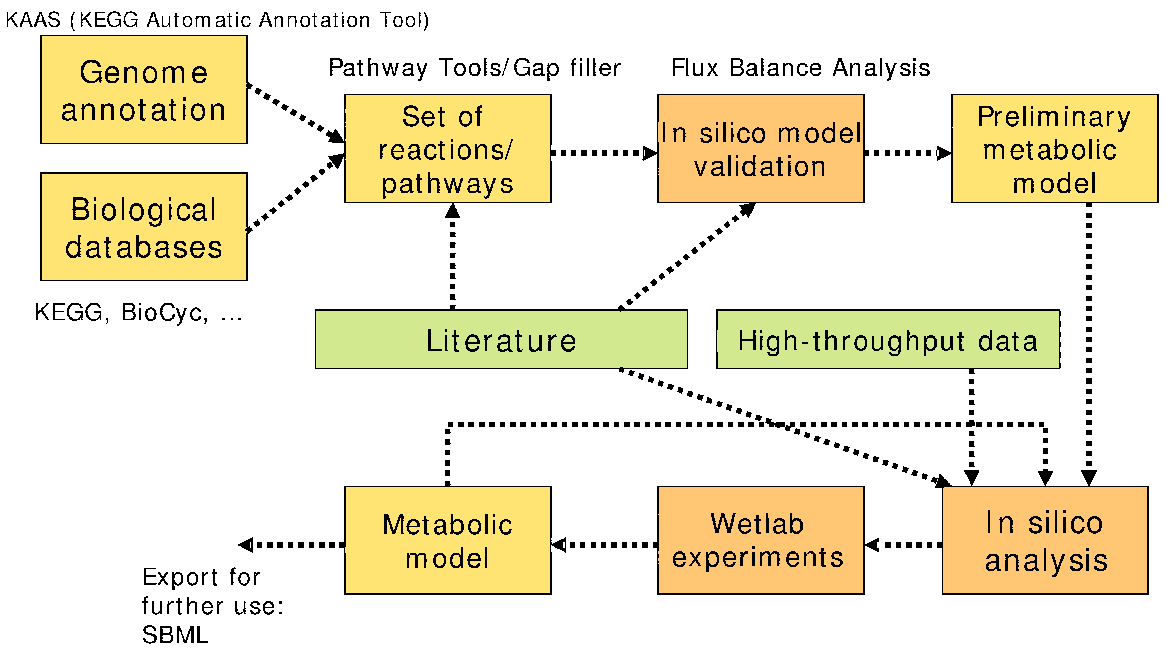


Complexity of ILP

- Unfortunately, integer linear programming is in general NP-hard
- NP-hard even with 0-1 (binary) variables (one of Karp's famous 21 NP-hard problems)
- Solvers typically resort to branch-and-bound or cutting plane methods (such as GLPK or lp_solve)
- We are unable to solve genome-scale gapless reconstruction with previous formulation
- Divide-and-conquer heuristic applied

Gapless reconstruction

- Gapless reconstruction combines two steps in the reconstruction workflow
  - Selection of the initial reaction set
  - Curation of the initial reaction set



Read more: Puchałka et al., Genome-Scale Reconstruction and Analysis of the Pseudomonas putida KT2440 Metabolic Network Facilitates Applications in Biotechnology. PLoS Computational Biology 2008.

Gapless reconstruction

- No previous knowledge on metabolic pathways needed!
  - However, *a priori* knowledge on metabolites and reactions can be plugged in
- Possible to discover pathways that are not previously known
- Article:

  E. Pitkänen, A. Rantanen, J. Rousu, E. Ukkonen: A computational method for reconstructing gapless metabolic networks. *2nd International Conference on Bioinformatics Research and Development (BIRD'08)*, Communications in Computer and Information Science, Vol. 13, Springer, 2008.

Advertisement: Bioinformatics Day

- Bioinformatics Day is the main event of The Finnish Society for Bioinformatics
- Organized in Turku on 13 May 2009
- Keynote lectures and short talks
- Announcement of the prize for the best bioinformatics PhD thesis in Finland in 2008
- `www.helsinki.fi/jarj/bioinfo/`