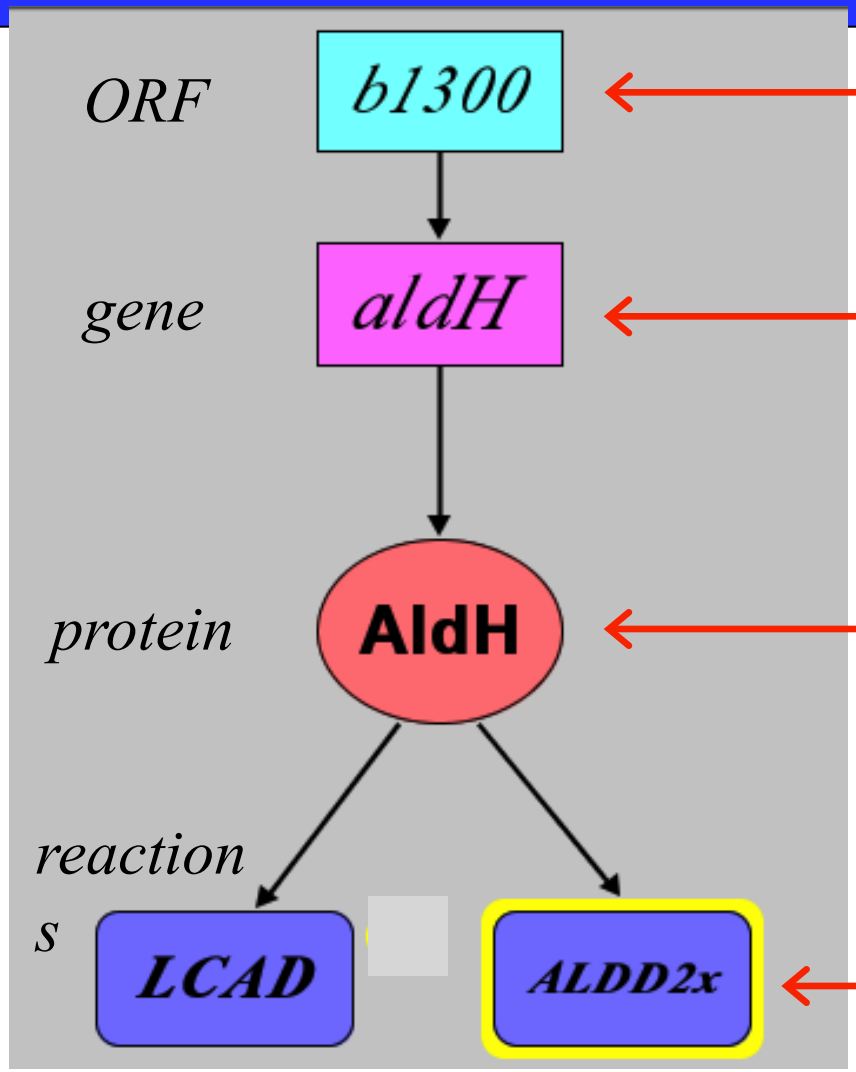
Constraint-Based Workshops:

Methods for Integrating Gene Expression Data

Integrating "-omics" Data



Genomics
ORF annotation

Transcriptomics
mRNA levels

Proteomics
protein levels

"Fluxomics"
flux measurements

Gene Expression: Method by Shlomi et al.

- Genes that are highly expressed are likely associated with reactions carrying non-zero flux
- Genes that are lowly expressed are likely associated with reactions carrying little or no flux.
- APPROACH: Find flux distributions where patterns of reaction usage match experimental data.

Shlomi, et al. Nature Biotechnology, **26(9)**: 1003-10 (2008).

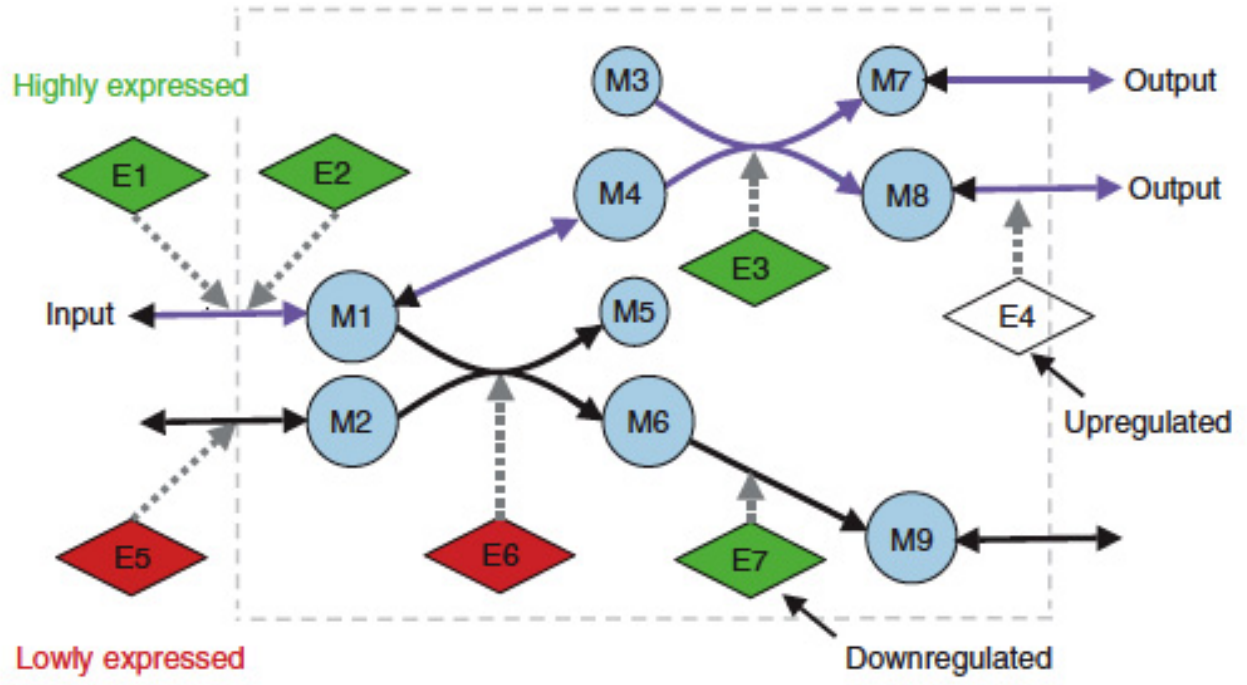


Figure 1 An example of predicting flux-activity states of genes based on a metabolic network model and gene-expression measurements. Circular nodes represent metabolites, whereas diamond nodes represent enzymes. White, red and green represent normal, significantly low and significantly high expression of the enzyme-encoding genes, respectively. Solid edges represent metabolic reactions. Broken edges associate enzymes with the reactions they catalyze. The predicted steady-state flux distribution, involving the activation of reactions, is shown as purple arrows. Enzyme E4 is predicted to be post-transcriptionally upregulated and E7 is predicted to be post-transcriptionally downregulated.

Shlomi, et al. Nature Biotechnology, **26(9)**: 1003-10 (2008).

$$\max_{v,y^+,y^-} \left(\sum_{i \in R_H} (y_i^+ + y_i^-) + \sum_{i \in R_L} y_i^+ \right)$$

Maximize Agreement:
Low Expression → Zero Flux
High Expression → Non-Zero Flux

$$s.t$$

$$S \cdot v = 0$$

Mass balance

$$v_{\min} \leq v \leq v_{\max}$$

Enzyme Capacity, Thermodynamics

$$v_i + y_i^+ (v_{\min,i} - \varepsilon) \geq v_{\min,i}, i \in R_H$$

$$v_i + y_i^- (v_{\max,i} + \varepsilon) \leq v_{\max,i}; i \in R_H$$

$y^+ = 0$, $v \geq vmin$
$y^+ = 1$, $v \geq epsilon$
$y^- = 0$, $v \leq vmax$
$y^- = 1$, $v \leq -epsilon$

$$v_{\min,i}(1 - y_i^+) \leq v_i \leq v_{\max,i}(1 - y_i^+), i \in R_L$$

$y^+ = 1$, $0 \leq v \leq 0$
$y^+ = 0$, $vmin \leq v \leq vmax$

$$v \in R^m$$

$$y_i^+, y_i^- \in [0,1]$$

Mapping Expression via GPR Rules in GAMS:

***Here X,Y,Z are binary (0/1), and represent genes and reactions**

- Z active if X
- Z active if X AND Y
- Z active if X OR Y

- Z = X
- Z = min(X,Y)
- Z = max(X,Y)

- E.g. ACONT can occur if acnA or acnB is present
- E.g. NADTRHD can occur if pntA and pntB
- E.g. PFL can occur if (pflA and pflB) or (pflC and pflD)

- ACONT=max (acnA,acnB)
- NADTRHD=min (pntA,pntB)
- PFL = max { min(pflA,pflB), min(pflC,pflD) }

Gene to Reaction Associations

- The reactionstatus represents the mRNA expression levels associated with a reaction and is used to determine which reactions likely take place.
- The reactionstatus for subunits is the minimum expression level of all subunits.
- The reactionstatus for isozymes is the maximum expression level of the isozymes.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%Change Reaction Status to 0 When Genes Needed Aren't Expressed%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
reactionstatus('ICL')=genestatus('aceA');
reactionstatus('MALS')=genestatus('aceB');
reactionstatus('ACKr')=genestatus('ackA');
reactionstatus('ADHEr')=genestatus('adhE');
reactionstatus('ADK1')=genestatus('adk');
reactionstatus('FUMt2_2')=genestatus('dctA');
reactionstatus('SUCCt2_2')=genestatus('dctA');
reactionstatus('SUCCt2b')=genestatus('dcuC');
```

Only One Gene

```
%These reactions have isozymes so if at least one is present the reaction can occur
reactionstatus('ACONT')=max(genestatus('acnA'),genestatus('acnB'));
reactionstatus('FUM')=max(genestatus('fumA'),genestatus('fumB'),genestatus('fumC'));
reactionstatus('PFK')=max(genestatus('pfkA'),genestatus('pfkB'));
reactionstatus('PFL')=max(min(genestatus('pflA'),genestatus('pflB'),genestatus('pflD')));
reactionstatus('PYK')=max(genestatus('pykA'),genestatus('pykF'));
reactionstatus('TKT1')=max(genestatus('tktA'),genestatus('tktB'));
reactionstatus('TKT2')=max(genestatus('tktA'),genestatus('tktB'));

%These reactions are carried out by multiple gene products, all have to be
reactionstatus('SUCOAS')=min(genestatus('sucC'),genestatus('sucD'));
reactionstatus('AKGDH')=min(genestatus('sucA'),genestatus('sucB'),genestatus
reactionstatus('PDH')=min(genestatus('aceE'),genestatus('aceF'),genestatus
reactionstatus('PIt')=min(genestatus('pitA'),genestatus('pitB'));
reactionstatus('NADTRHD')=min(genestatus('pntA'),genestatus('pntB'));
```

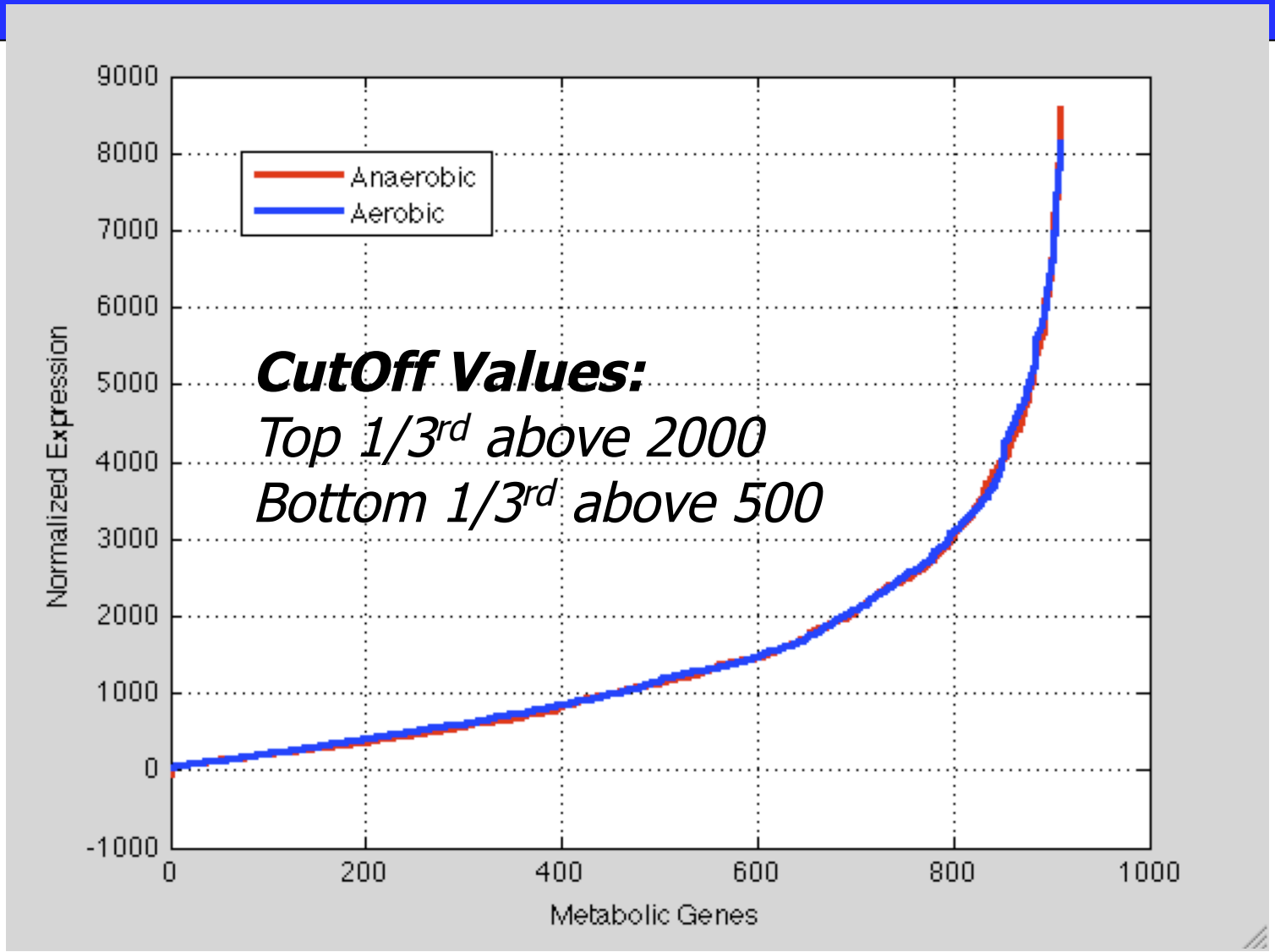
Multiple Isozymes: "max"

Multiple Subunits "min"

E. coli: Aerobic vs. Anaerobic

- Growth in Glucose minimal media
- Measured uptake and secretion rates of substrates (glucose and oxygen) and by-products (acetate, ethanol, succinate, etc.).
- Measured gene expression data on AffyMetrix arrays.

E. coli Gene Expression Data (Metabolic Genes) Covert et al. Nature 2004



C:\Documents and Settings\Jennie Reed\Desktop\GeneExpressionData.gms

GeneExpressionData.gms

```
Table ExpressionData(genes,conditions)
          aerobic           anaerobic
aceA      5745.963333       2280.975
aceB      6467.956667       3797.2925
aceE      5633.123333       4035.13
aceF      4028.54            2885.97
ackA      2143.503333       6021.365
acnA      1052.466333       427.129
acnB      3099.566667       499.66575
adhE      2285.843333       6199.2725
adk       3439.543333       3936.2275
atpA      5102.27            4914.475
atpB      4277.09            4414.9
atpC      2137.526667       2122.3175
atpD      3249.136667       3225.165
atpE      6293.003333       6372.2875
atpF      6194.773333       6097.8775
atpG      4937.786667       4621.525
atpH      7723.083333       7441.3075
atpI      3674.956667       4026.5125
crr       4626.223333       5005.2225
cydA      1744.373333       3014.0975
cydB      2154.123333       3408.4275
dctA      345.8756667       262.2005
dcuC      908.7576667       1088.8065
```

MapData2Rxns.gms

```
$offlisting

Parameters genestatus(genes),reactionstatus(j);
reactionstatus(j)=0;
genestatus(genes)=0;

loop(select_condition, genestatus(genes)=E

*******************************************************************************
**********Change Reaction Status to 0 When G                    ******
*******************************************************************************
reactionstatus('ICL')=genestatus('aceA');
reactionstatus('MALS')=genestatus('aceB');
reactionstatus('ACKr')=genestatus('ackA');
reactionstatus('ADHEr')=genestatus('adhE');
```

Picks the dataset to use based on the set select_condition

```
reactionstatus('FRD')=min(genestatus('frdA'),genestatus('frdB'),genestatus('frdC'),genestatus('frdD'));
reactionstatus('NADH11')=min(genestatus('nuoA'),genestatus('nuoB'),genestatus('nuoE'),genestatus('nuoF'),g
reactionstatus('GLCpts')=min(genestatus('ptsG'),genestatus('ptsH'),genestatus('ptsI'),genestatus('crr'));
reactionstatus('ATPS4r')=min(genestatus('atpA'),genestatus('atpB'),genestatus('atpC'),genestatus('atpD'),g
reactionstatus('CYTBD')=min(genestatus('cydA'),genestatus('cydB'));

*These reactions have no genes associated with them
set NoGPR(j) /EX_ac_e,EX_akg_e,EX_co2_e,EX_etoh_e,EX_for_e,EX_fum_e,EX_glc_e,EX_h_e,EX_h2o_e,EX_lacD_e,EX_
ACt2r,AKGt2r,ATPM,Biomass,CO2t,DLACt2,ETOHt2r,H2Ot,O2t,PYRt2r/;
reactionstatus(NoGPR)=(high_cutoff+low_cutoff)/2;

sets Rhigh(j),Rmed(j),Rlow(j);
Rhigh(j)=no;Rmed(j)=no;Rlow(j)=no;
loop(j, if ( reactionstatus(j)>high_cutoff,
                Rhigh(j)=yes;
            elseif (reactionstatus(j)<low_cutoff),
                Rlow(j)=yes;
            else
                 Rmed(j)=yes; );
       );

display Rhigh, Rmed, Rlow;
```

Determines which reactions belong to the different subsets (high, med, low)

Main File:

GeneExpression_CoreTextbookModel_Shlomi.gms

```
$offlisting
Sets conditions conditions for gene expression datasets /aerobic,anaerobic/
select_condition(conditions) condition to analyze /anaerobic/
genes list of genes in the model /aceA,aceB,ackA,acnA,acnB,adhE,adk,aceE,aceF
atpA,atpB,atpC,atpD,atpE,atpF,atpG,atpH,atpI
crr,cydA,cydB,dctA,dcuC,eno,fba,fbp,focA,frdA,frdB,frdC,frdD
fumA,fumB,fumC,gapA,gltA,gnd,icdA,ldhA,lpdA,maeB,mdh
nuoA,nuoB,nuoE,nuoF,nuoG,nuoH,nuoI,nuoJ,nuoK,nuoL,nuoM,nuoN
pckA,pfkA,pfkB,pflA,pflB,pflC,pflD,pgi,pgk,pgl,pgm,pitA,pitB
pntA,pntB,ppc,ppsA,pta,ptsG,ptsH,ptsI,pykA,pykF,rpe,rpi
sdhA,sdhB,sdhC,sdhD,sfcA,sucA,sucB,sucC,sucD,talA,tktA,tktB,tpi,zwf/;

Parameter high_cutoff /2000/, low_cutoff /500/;

*Read in the appropriate S matrix
$include CoreTextbookModel.gms
*Read in gene expression data
$include GeneExpressionData.gms
*Determine which reactions belong to set Rhigh and Rlow
$include MapData2Rxns.gms
```

Pick Dataset

Cutoff Expression Values

To run this program you will need a full license to GAMS, since the number of binary variables exceeds the maximum available in the demo license.

Main File (cont.):

GeneExpression_CoreTextbookModel_Shlomi.gms

```
UpperLimits(j)=Vmax;
*CARBON SOURCE: select upper and lower limits for exchange flux
LowerLimits('EX_glc_e')=-17.3;
UpperLimits('EX_glc_e')=-17.3;
LowerLimits('EX_o2_e')=0;
UpperLimits('EX_o2_e')=0;
LowerLimits('EX_ac_e')=8;
UpperLimits('EX_ac_e')=8;

*allow co2,pi,o2,h,h2o to be taken up by the cell
LowerLimits('EX_co2_e')=-Vmax;
LowerLimits('EX_h2o_e')=-Vmax;
LowerLimits('EX_h_e')=-Vmax;
LowerLimits('EX_pi_e')=-Vmax;

LowerLimits('ATPM')=7.6;
S(i,'Biomass')=1.3*S(i,'Biomass');


Variables
v(j) flux values through reacti
Obj  this is the value of the ob
growth growth rate;

Binary Variables x(j),y(j);
Parameter epsilon /0.5/;
```

Fix Uptake and Secretion Rates to Measured Values

Define what is considered non-zero using epsilon
e.g. if flux is between +/-epsilon than it is considered to be inactive

Main File (cont.):

```
Solve Shlomi_GeneExp using mip maximizing Obj;
Obj.fx=Obj.l;

*find the alternate solution with the maximum growth rate
Solve Shlomi_GeneExp using mip maximizing growth;
maxgrowth(j)=v.l(j);

*find the alternate solution with the maximum growth rate
Solve Shlomi_GeneExp using mip minimizing growth;
mingrowth(j)=v.l(j);
```

Solves the problem three times.

1. *Find the maximum agreement between fluxes and expression*
2. *Maximize growth rate keeping agreement the same.*
3. *Minimize growth rate keeping agreement the same.*

Aerobic Questions

To run answer these questions you will need a full license to GAMS, since the number of binary variables exceeds the maximum available in the demo license.

- Use FBA to determine what the maximum and minimum growth rate is when you constrain the glucose uptake (9.02 mmol/gDW/h), oxygen uptake (14.93 mmol/gDW/h) and acetate secretion (4.15 mmol/gDW/h). Make sure you include the extra lines listed below in the FBA code.
  - LowerLimits('ATPM')=7.6;
  - S(i,'Biomass')=1.3*S(i,'Biomass')

- Using the same flux measurements and gene expression data determine what the maximum and minimum growth rate could be when you constrain fluxes using the Shlomi method.

- Based on this result does the Shlomi method have multiple flux distributions that are optimal (i.e. match expression patterns)?

Aerobic Answers

- Use FBA to determine what the maximum and minimum growth rate is when you constrain the glucose uptake (9.02 mmol/gDW/h), oxygen uptake (14.93 mmol/gDW/h) and acetate secretion (4.15 mmol/gDW/h). Make sure you include the extra lines listed below in the FBA code.
  - Max Growth=0.544
  - Min Growth=0
- Using the same flux measurements and gene expression data determine what the maximum and minimum growth rate could be when you constrain fluxes using the Shlomi method.
  - Max Growth=0.517
  - Min Growth=0
- Based on this result does the Shlomi method have multiple flux distributions that are optimal (i.e. match expression patterns)? Yes

More Aerobic Questions

- When biomass is minimized what by-products are produced (note that if biomass is not made the equivalent carbon must be secreted as by-products)?
- How many reactions are in the three different sets: Rhigh, Rlow, and Rmed? How many of these can you match the flux patterns to (i.e. what is the optimal objective value)
- What happens to the number of Rhigh, Rlow, and Rmed reactions and the when you make the low_cutoff value large (e.g. 1000)?
- What happens to the max and min biomass when you make epsilon bigger (e.g. 1)? Set the low_cutoff back to 500.

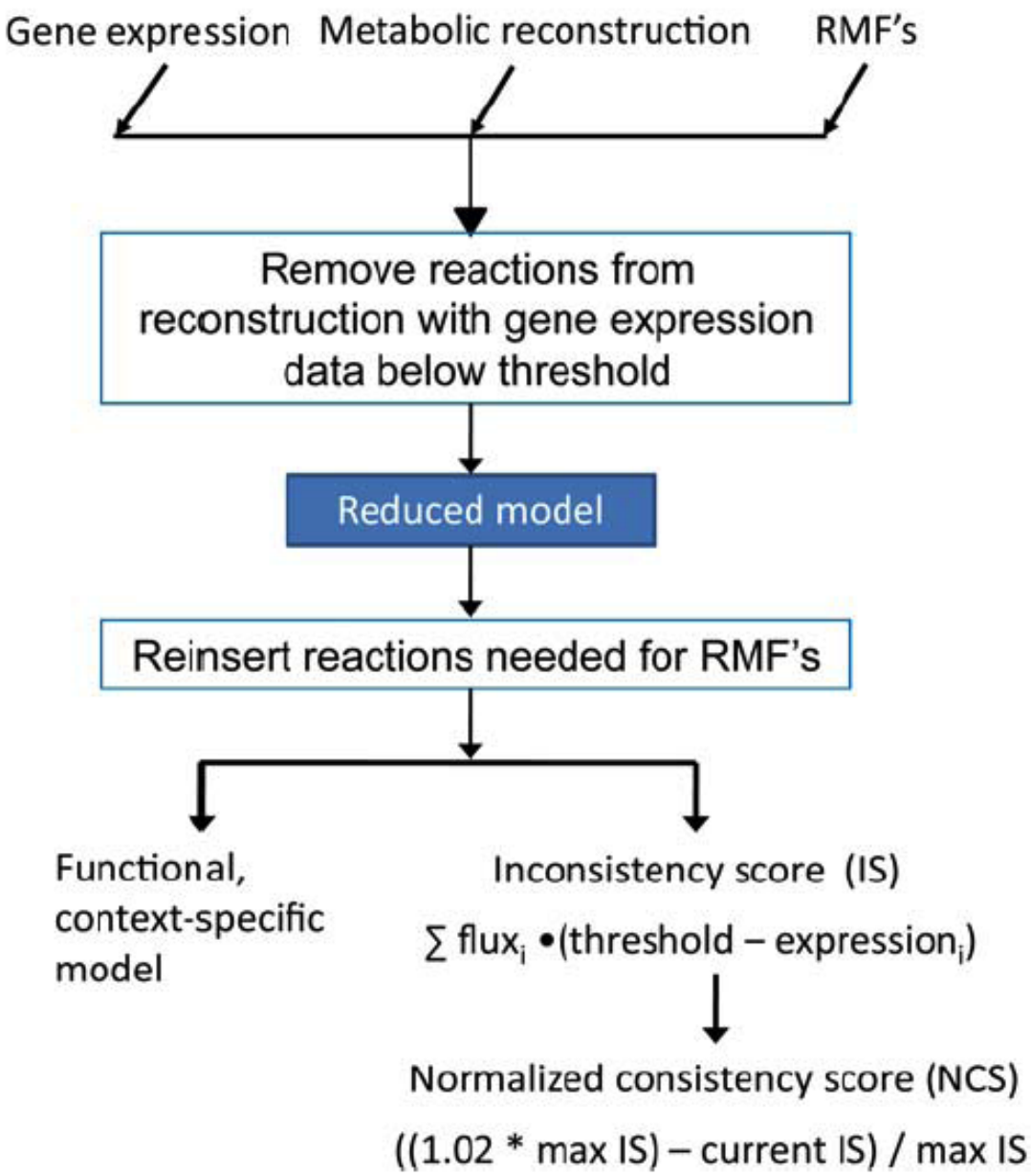
More Aerobic Questions

- When biomass is minimized what by-products are produced (note that if biomass is not made the equivalent carbon must be secreted as by-products)?
  - Lacate, succinate, formate are now produced and not produced at maximal growth rate
- How many reactions are in the three different sets: Rhigh, Rlow, and Rmed? How many of these can you match the flux patterns to (i.e. what is the optimal objective value)?
  - Rhigh=30, Rlow=3 and Rmed=44
  - Max possible score is 30+3 (Rhigh+Rlow).
  - Optimal value is 33
- What happens to the number of Rhigh, Rlow, and Rmed reactions and the when you make the low_cutoff value large (e.g. 1000)?
  - Rhigh=30, Rlow=8 and Rmed=39
  - Max possible score is now 38 (Rhigh+Rlow) and optimal value is 37
  - PTAr (used to make acetate) has to have flux but it is now in the Rlow set
- What happens to the max and min biomass when you make epsilon bigger (e.g. 1)? Set the low_cutoff back to 500
  - Max growth rate=0.483 and Min growth rate =0

Another Approach

Eliminate reactions below threshold first.

Add them back in if they are needed to meet some known function (e.g. growth, product formation, nutrient degradation)



Becker and Palsson. PLoS Comp Bio. 4(5):e1000092 (2008)

Minimize Inconsistency Score

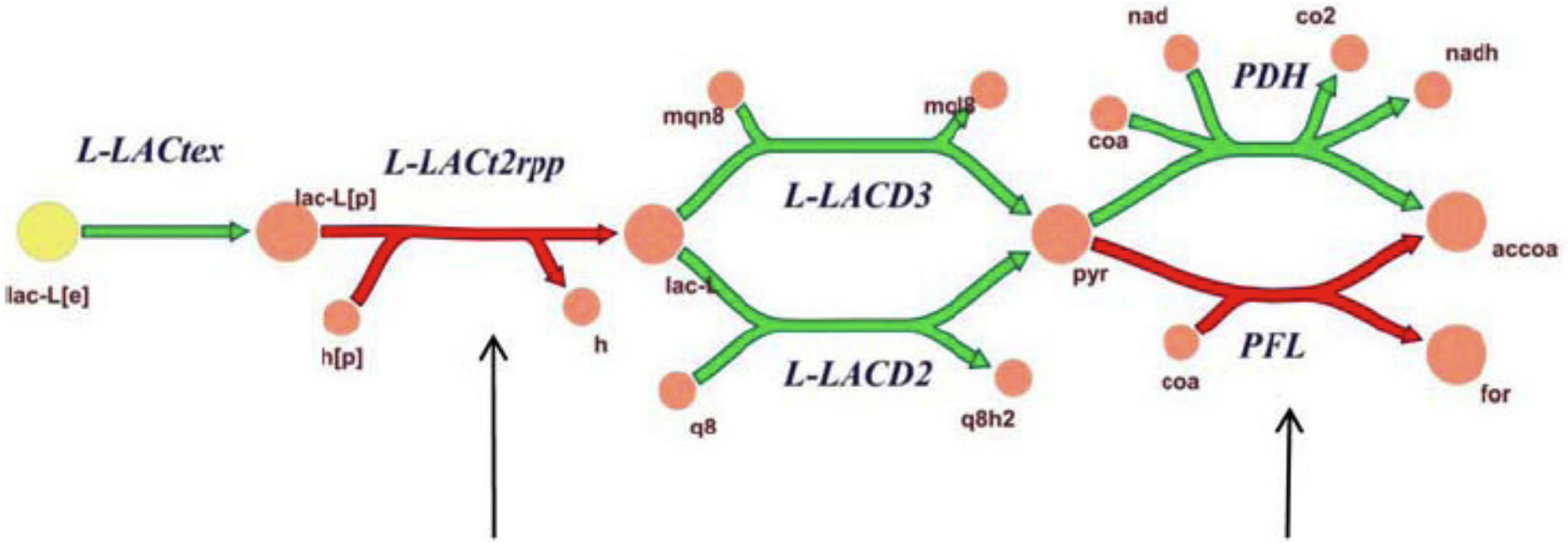
$$
\begin{aligned}
&\text{minimize :} && \textstyle\sum c_i \bullet |v_i| \\
&\text{subject to :} && \boldsymbol{S} \bullet \boldsymbol{v} = 0 \\
& && a_i < v_i < b_i \\
&\text{where} && c_i = \{x_{cutoff} - x_i \text{ where } x_{cutoff} > x_i \\
& && \quad 0 \text{ otherwise} \\
& && \text{for all i.}
\end{aligned}
$$

Need to map gene expression values to reactions using GPRs for cases where multiple genes are associated with a reaction

Implemented in MATLAB (see TIGER toolbox)

Becker and Palsson. PLoS Comp Bio. 4(5):e1000092 (2008)

Example:



14.6 * (12- 11.4) = 8.8

flux * (threshold – data)

0.005 * (12 - 8.2) = 0.02

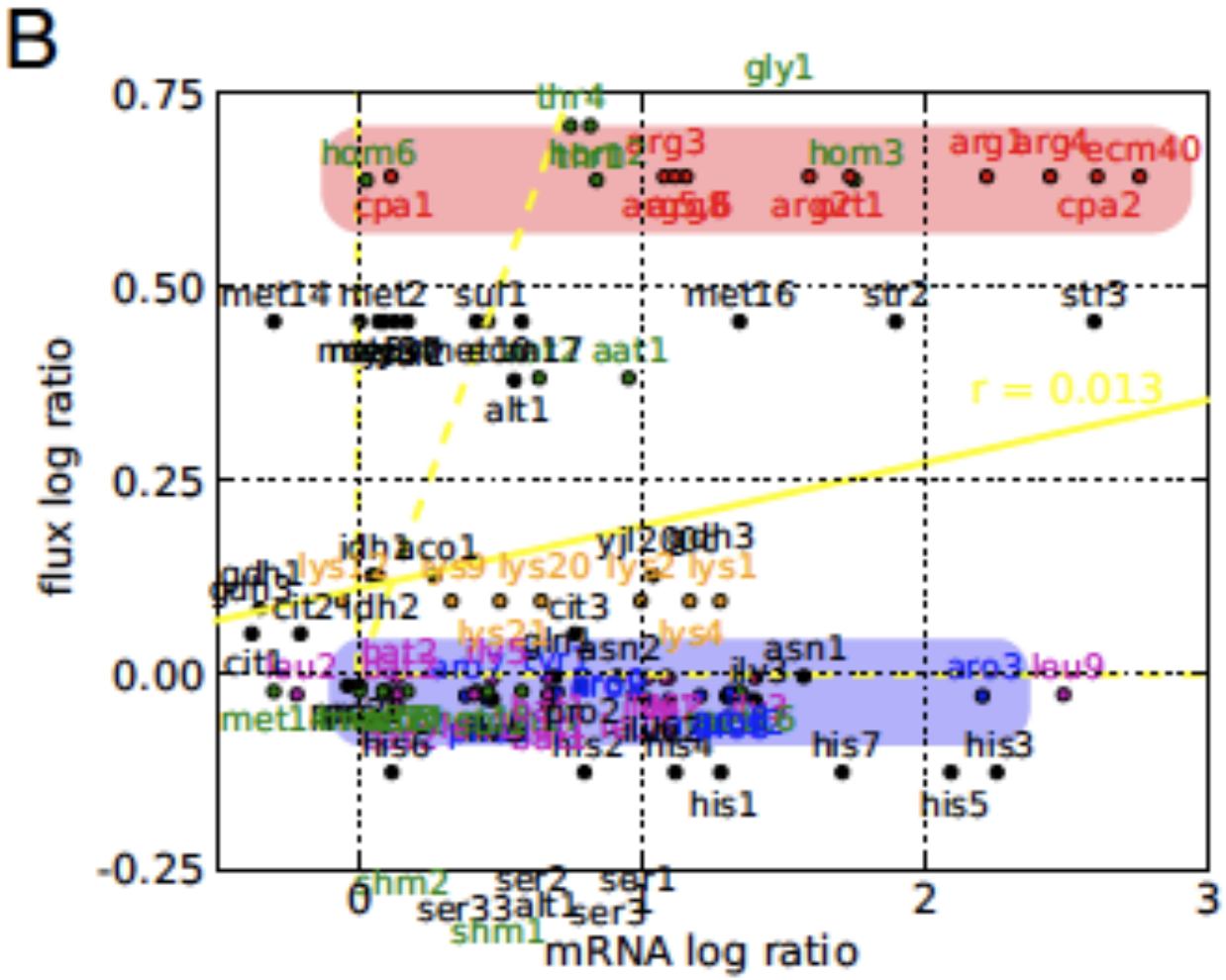
flux * (threshold – data)

Becker and Palsson. PLoS Comp Bio. 4(5):e1000092 (2008)

Recommended Additional Reading

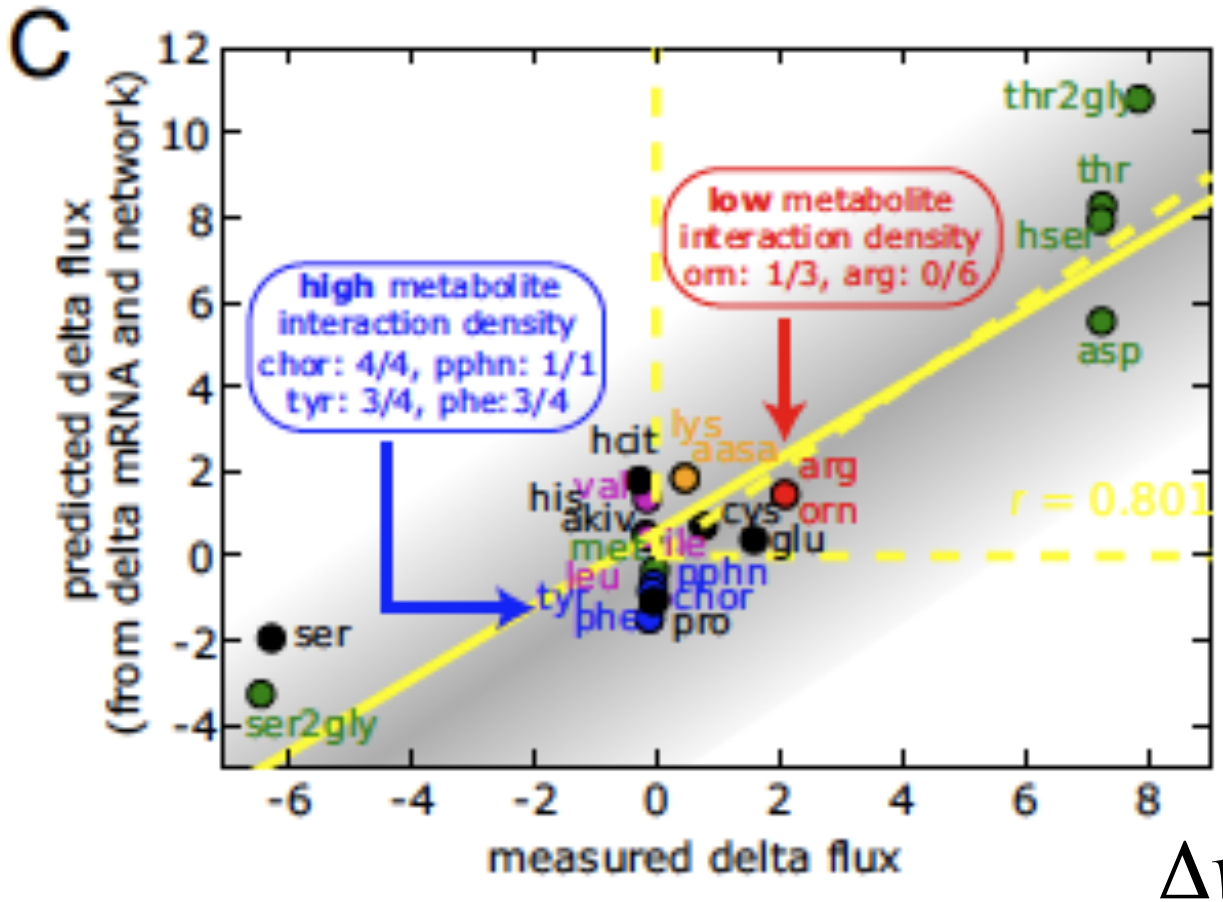
- Gene Expression:
  - Colijn et al. PLOS Computational Biology. 5(8), (2009).
    - Uses relative expression values to place upper limits on flux values.
    - GeneX has higher expression than GeneY so upperlimit of RxnX is higher than upperlimit of RxnY.
  - Moxley et al. PNAS. 106(16): 6477-6482 (2009).
    - Uses changes in gene expression to predict changes in flux values.
    - Flux change depends on expression change and additional model parameters.

No Direct Correlation Between Flux Values and mRNA Expression



Moxley et al. PNAS. 106(16): 6477-6482 (2009)

Another Approach: Quantitative Prediction of Flux Changes



Requires estimates for parameters p1 and p2

$$\Delta v = \frac{e^{-(d*p1)} \cdot \Delta mRNA}{p2}$$

Moxley et al. PNAS. 106(16): 6477-6482 (2009)