Analysis of Flux Distributions of *Clostridium thermocellum* Mutants Through Physiological Characterization and Proteomics

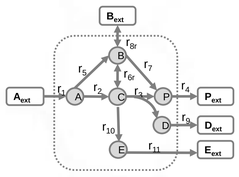
Sergio Garcia[1,2], R. Adam Thompson[2,3], Richard Giannone[2,4], Robert L. Hettich[2,4], and Cong T. Trinh[1,2,3]

[1] Department of Chemical and Biomolecular Engineering, University of Tennessee; Knoxville, Tennessee; [2]BioEnergy Science Center, Oak Ridge, Tennessee; [3]Bredesen Center for Interdisciplinary Research and Education [4] Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge,Tennessee.

Abstract: Novel high-throughput techniques such as transcriptomics, proteomics, and metabolomics, serve to produce massive amounts of data from biological systems. This explosion in our ability to produce data has not been matched by our ability to analyze it. We propose a method that serves to refine the prediction capability of genome scale models of metabolism from proteomics data. Our method uses protein abundances to effectively reduce the solution space of the model, under the sole assumption that reaction fluxes are to be minimal while satisfying the cellular objective (e.g. growth rate maximization). We validated our method by examining randomly sampled reaction flux distributions of the genome scale model of *Clostrium thermocellum* iAT601, constrained with proteomic data, and measured reaction fluxes for the wild type and **Δ**hydG-**Δ**ech strains. The uncertainty of the model was effectively reduced in both parent and mutant strains. Additionally, many of the central metabolism estimated flux changes between strains are in good agreement with previous studies[3]. We expect the enhanced predictive accuracy of the model to drive *in silico* metabolic engineering for the production of biofuels and chemicals.

Basics of metabolic network modeling

How is a metabolic network model built?



$S =$

	r_1	r_2	r_3	r_4	r_5	r_{6r}	r_7	r_{8r}	r_9
A	1	-1	0	0	-1	0	0	0	0
B	0	0	0	0	1	-1	-1	-1	0
C	0	1	-1	0	0	1	0	0	0
D	0	0	1	0	0	0	0	0	-1
P	0	0	1	-1	0	0	2	0	0

$\boldsymbol{r} = [r_1\ r_2\ r_3\ r_4\ r_5\ r_{6r}\ r_7\ r_{8r}\ r_9\ r_{10}\ r_{11}]^T$

At quasi-steady state:

$\mathbf{S \cdot r = 0}$
$\mathbf{lb \leq r \leq ub}$

lb and ub are the upper and lower bound for the reaction fluxes.

Problem:

Many reactions bounds (**lb**,**ub**) are not known, therefore there are many flux distributions(**r**) that can satisfy the model equations. We need to constraint the reaction bounds using experimental data.

How can the model predict reaction fluxes?

- Loopless Flux balance analysis[1]:

  Maximize Biomass flux

  Subject to:

  $\mathbf{S \cdot r = 0}$

  $\mathbf{lb \leq r \leq ub}$

  loop law constraints

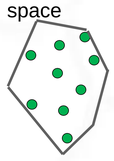
Flux balance analysis yields a flux value for each reaction in the network

- Sampling of the solution space[2]:

  Obtain a set of flux distributions {r}, homogenously distributed in the space enclosed by:

  $\mathbf{S \cdot r = 0}$

  $\mathbf{lb \leq r \leq ub}$



The solution space is a convex polyhedron

Random sampling yields a set of possible flux values for each reaction in the network.

Theory of proteomics data integration

Relation between protein abundance and reaction flux

The reaction rate is described by:

$$r_i = \alpha_i \cdot E_i$$

- E_i is the enzyme concentration
- α_i is a condition specific function, which depends on metabolite concentrations, temperature, pH, post-translational regulation, etc.

The enzyme concentration is directly proportional by some constant β_i to the net protein abundance for each reaction:

$$E_i \cdot \beta_i = A_i = \sum_{j=1}^{\text{genes coding reaction i}} (\text{protein abundance})_j$$

- Absorb β_i in α_i such that $r_i = \alpha_i \cdot A_i$

Finding unknown parameters

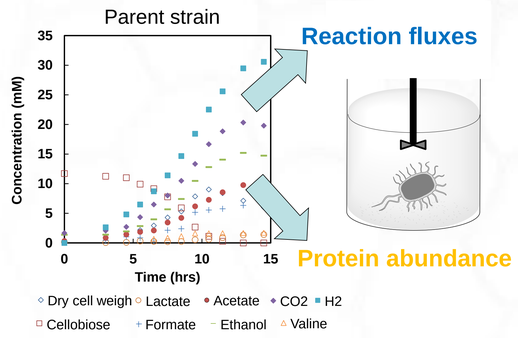
- Define an upper bound $\gamma \geq \alpha_i\ \forall\ i_{unconstrained}$ such that $r_i \leq \gamma \cdot A_i$.
- In order to find γ, invoke a parsimony assumption, whereby the cell satisfies its biological objective (e.g. maximization of growth rate) with the smallest fluxes possible:

$$\min \gamma$$
$$\text{s.t.}\ \max \text{biomass flux}$$
$$\text{s.t.}$$
$$\mathbf{S \cdot r = 0}$$
$$-\gamma \cdot A_i \leq r_i \leq \gamma \cdot A_i\ \forall\ i_{reversible}$$
$$0 \leq r_i \leq \gamma \cdot A_i\ \forall\ i_{forward}$$
$$-\gamma \cdot A_i \leq r_i \leq 0\ \forall\ i_{backward}$$
$$lb_i \leq r_i \leq ub_i\ \forall\ i_{experimental}, i_{zeroAbundance}$$

- Experimentally constrained[3] bounds are not modified.
- Bounds of reactions that have a net abundance of zero are not changed, since a net abundance of zero can be due to a failure in detecting the corresponding proteins.

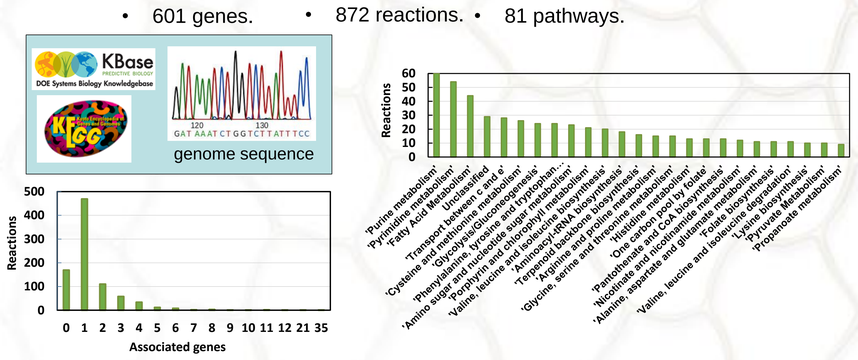
How does the proteomic data relate to the model?

Experimental data for parent and ΔhydG-Δech strains[3]

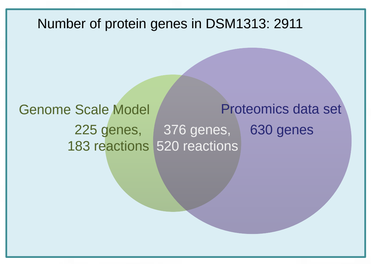


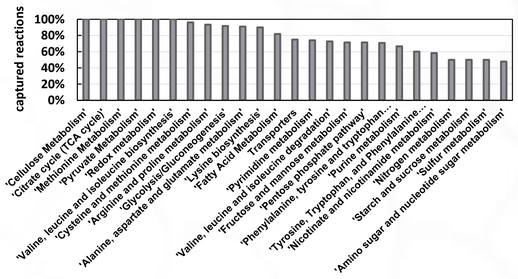
iAT601 genome scale model[4] of C. thermocellum DSM 1313

- 601 genes.
- 872 reactions.
- 81 pathways.



Mapping of proteomics data to the model





Protein abundance and randomly sampled flux distributions indicate potential regulation mechanism[5].

Enzymes Showing Transcriptional Regulation (P > 0.6)

Fructose bisphosphate aldolase
ATP phosphoribosyltransferase
UDP-glucuronate 5-epimerase
ADP-ribose ribophosphohydrolase

Enzymes Showing Metabolic Regulation (P > 0.9)

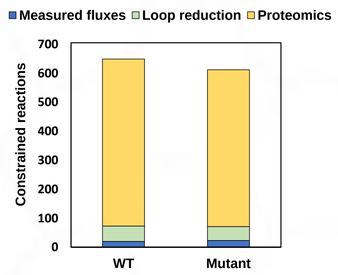
Iron(III) Uptake
hydrogen:ferredoxin oxidoreductase
Lumped Cellulosome Term
ATP:D-ribose-5-phosphate diphosphotransferase
Fatty acid synthase
ATP:D-glucose 6-phosphotransferase
Glutamate racemase
Isoleucine transport
UDP-N-acetylmuramoyl-L-alanyl-D-glutamate:(L)-meso-2,6-diaminoheptanedioate gamma-ligase (ADP-forming)

Enzymes Showing Post-Translational Regulation (P > 0.99)

Nicotinate-nucleotide:dimethylbenzimidazole phospho-D-ribosyltransferase
UDP-glucuronate 5-epimerase
Thiamine biosynthesis protein
D-Glucitol:NAD+ 2-oxidoreductase
reduced ferredoxin:dinitrogen oxidoreductase (ATP-hydrolysing)
GTP:GTP guanylyltransferase
propane-1,3-diol:NAD+ 1-oxidoreductase
ATPase, proton translocating
AMP-dependent synthetase and ligase
L-Alanine:tRNA(Ala) ligase (AMP-forming)
Cellodextrin Uptake
Glycogen Cycling 1
glycine:tRNA(Gly) ligase (AMP-forming)
AMP-dependent synthetase and ligase
propane-1,3-diol:NAD+ 1-oxidoreductase
(R)-4-Phosphopantothenate:L-cysteine ligase
L-Aspartate:tRNA(Asp) ligase (AMP-forming)
GTP:GTP guanylyltransferase
Polysaccharide biosynthesis protein
GTP 7,8-8,9-dihydrolase (diphosphate-forming)
UDP-glucose:NAD+ 6-oxidoreductase
L-cysteine:[Thil] sulfurtransferase

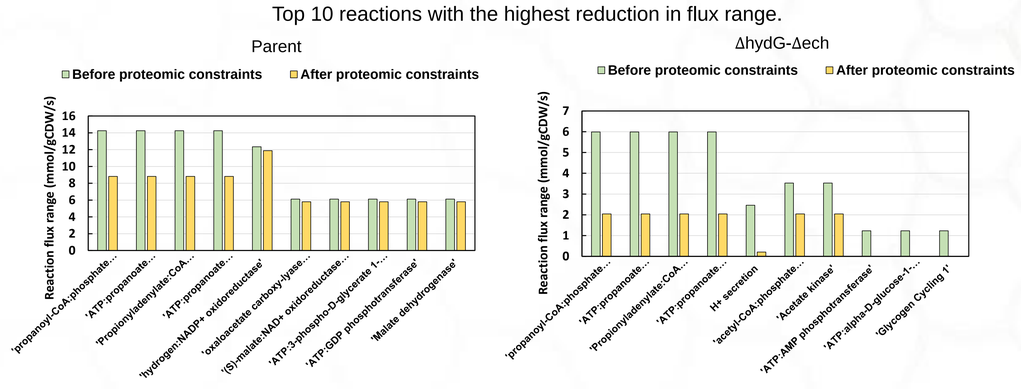
Theory validation

Effects of the proteomics constraints in the solution space



The proteomics constraints effectively reduce the flux variability ranges of 24 reactions in the parent and 22 reactions in the mutant.

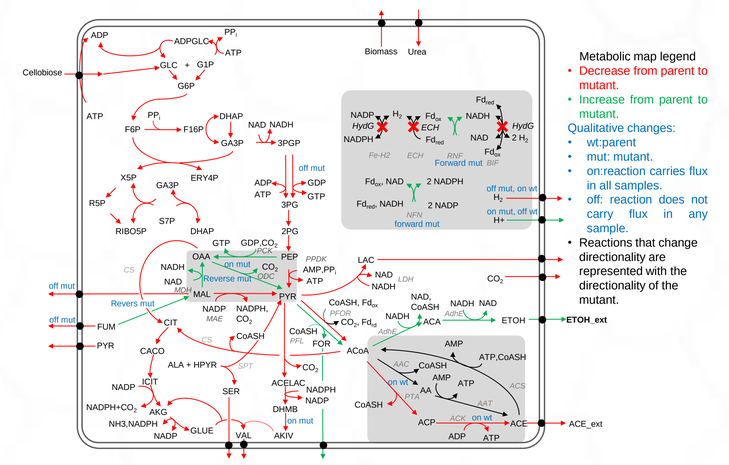
Top 10 reactions with the highest reduction in flux range.



Major flux differences of parent and ΔhydG-Δech strains after proteomics integration

Important reactions

Pathway	reaction name	Parent flux (mmol/gCDW/s) mean	stdev	min	max	Protein abundance (A.U.)	ΔhydG-Δech flux (mmol/gCDW/s) mean	stdev	min	max	Protein abundance (A.U.)
Redox	RNF-Ferredoxin:NAD oxidoreductase	0.21	0.59	-1.13	2.47	53.4	3.89	0.01	3.80	3.93	48.9
	NfnAB	0.25	0.53	-1.46	2.68	26.8	1.60	0.01	1.57	1.64	27.2
	hydrogen:NADP+ oxidoreductase	3.86	1.03	-0.27	7.60	11.1	0.00	0.00	0.00	0.00	10.5
Pyruvate	Malate dehydrogenase	-0.73	0.56	-3.78	0.13	14.8	0.08	0.01	0.02	0.08	15.8
	Pyruvate formate lyase	1.78	0.00	1.78	1.78	14.6	2.69	0.01	2.68	2.75	15.4
	(S)-malate hydro-lyase (fumarate-forming)	-0.09	0.03	-0.13	0.03	23.3	-0.08	0.00	-0.08	-0.08	23.6
	oxaloacetate carboxy-lyase (pyruvate-forming)	0.78	0.58	0.00	3.88	11.7	6.68	0.01	6.63	6.70	12.4
	'L-Serine:pyruvate aminotransferase'	-1.46	0.98	-3.60	0.00	0.0	-1.95	1.36	-5.10	0.00	0.0
	'Malate transport'	0.04	0.03	0.00	0.16	0.0	0.00	0.00	0.00	0.00	0.0
	'Fumarate transport'	0.04	0.03	0.00	0.16	0.0	0.00	0.00	0.00	0.00	0.0
	'Pyruvate transport'	0.18	0.02	0.13	0.20	0.0	0.00	0.00	0.00	0.00	0.0
Acetate	'Acetate:CoA ligase (AMP-forming)'	2.28	0.91	0.00	3.60	14.2	2.86	1.21	0.00	5.10	14.0
	'acetyl adenylate:CoA acetyltransferase'	1.27	0.89	0.00	3.60	14.2	2.05	1.25	0.00	5.10	14.0
	'ATP:acetate adenylyltransferase'	1.27	0.89	0.00	3.60	14.2	2.05	1.25	0.00	5.10	14.0
	'acetyl-CoA:phosphate transacetylase'	4.49	1.08	0.56	7.08	13.1	0.96	0.54	0.00	1.96	13.1
	'Acetate kinase'	4.49	1.08	0.56	7.08	14.1	0.96	0.54	0.00	1.96	13.4



- pyruvate phosphate dikinase (PPDK)
- oxaloacetate decarboxylase (ODC)
- malic enzyme (MAE)
- pyruvate formate lyase (PFL)
- pyruvate ferredoxin oxidoreductase (PFOR)
- lactate dehydrogenase (LDH)
- alcohol dehydrogenase (AdhE)
- phosphotransacetylase (PTA)
- citrate synthase (TCA1)
- Ni-Fe energy conserving hydrogenase (ECH)
- Rnf-type NADH:Fd oxidoreductase (RNF)
- NADH-Fdrd:NADP+ oxidoreductase (NFN)
- NADH-Fdrd bifurcating hydrogenase (BIF)
- **Fe-Fe NADPH-depenent hydrogenase (Fe-H2)**
- acetyl adenylate, acetyltransferase(AAC)
- acetate adenylyltransferase(AAT)
- Acetyl-Coa Synthase(ACS)
- Citrate synthase(CS)
- Serine:pyruvate aminotransferase(SPT)
- HPYR: Hydroxypyruvate
- ALA:Alanine

The flux redistribution between parent and mutant strain at key metabolic nodes is in good agreement with the results from reference 3, which used a different analysis method and a different model.

Summary

- We developed a theory to enhance the predictive capability of metabolic models by integration of proteomics into a genome scale model.
- The theory was validated using experimental data for *C. thermocellum* mutants.
- We expect this method to drive metabolic engineering efforts in general, and for the production of next generation biofuels and chemicals in *Clostridium thermocellum*

References

1-Schellenberger et. al (2011). *Biophysical journal*, *100*(3), 544-553.
2-Megchelenbrink et. al (2014). *PloS one*, *9*(2), e86587.
3-**Thompson,** et. al (2015). *Metabolic engineering*, *32*, 207-219.
4-**Thompson,** et. al (2016) under review.
5-Bordel et. al (2010). *PLoS Comput Biol*, *6*(7), e1000859.

The BioEnergy Science Center (BESC) is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.