VCU
School of Engineering

Integration of RNA-Seq data with an *in silico* metabolic reconstruction of *Clostridium thermocellum* for rational strain design

systems biological engineering laboratory

Control Number: 4306

Poster Presentation Number: O-2282

Chris M Gowen[1], **Stephen S Fong**[1,2]

[1]Department of Chemical and Life Science Engineering, Virginia Commonwealth University, Richmond, VA, USA
[2]Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA, USA

Contact:
gowencm@vcu.edu
(804) 827-7000 ext 414

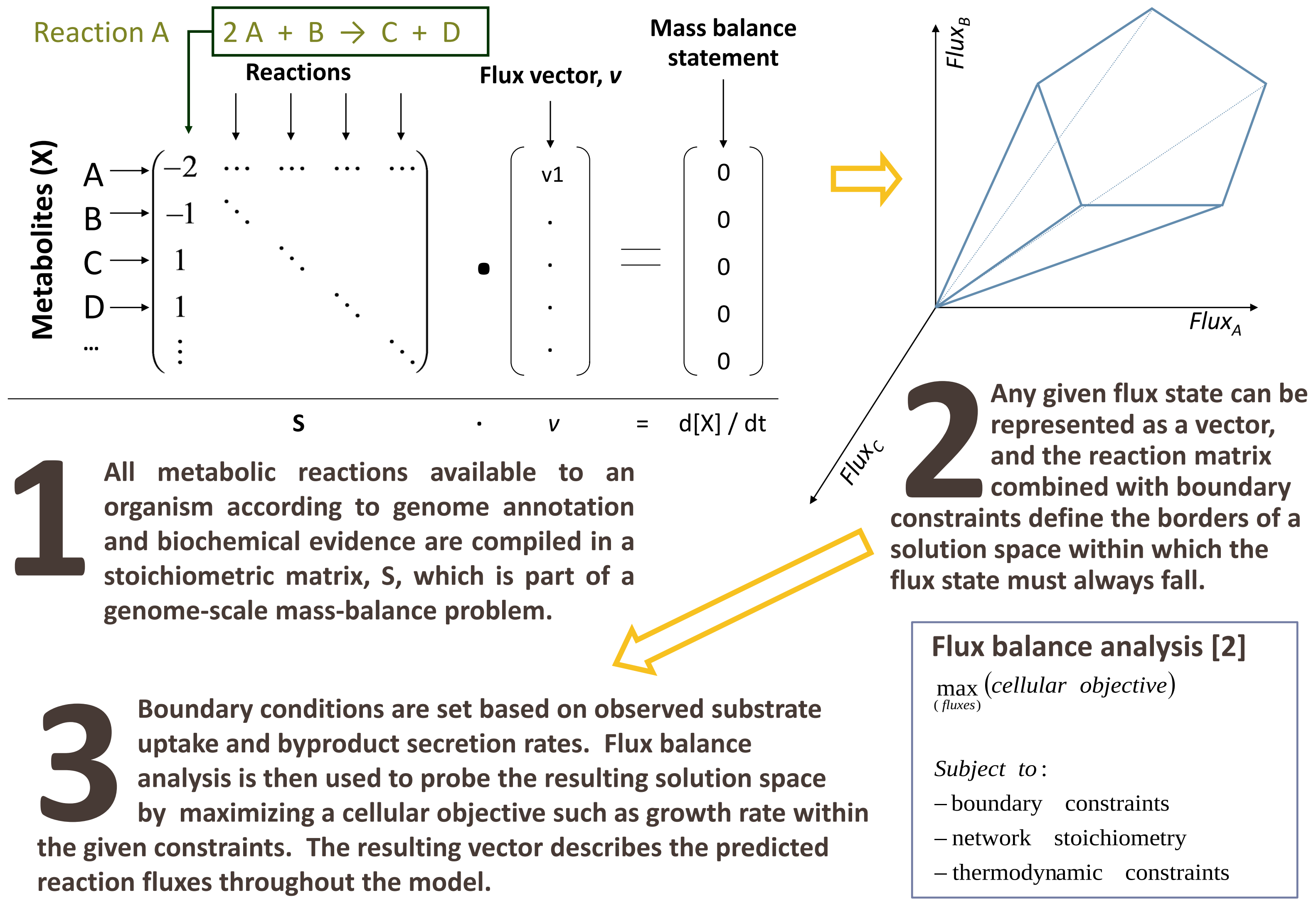
Abstract

Microbial production of solvents such as ethanol from cellulosic biomass could provide a sustainable energy source that is relatively cheap, abundant and environmentally sound, but production costs are driven up by a separate enzymatic treatment that is necessary for releasing fermentable sugars. These costs could be significantly reduced if a microorganism could be engineered to efficiently and quickly convert cellulosic biomass directly to product in a one-step process. The thermophilic anaerobe *Clostridium thermocellum* is a good candidate for this type of process because it naturally possesses the ability to directly convert cellulose to ethanol, obviating both the need for separate saccharification and fermentation steps, as well as the requirement for additional cellulase enzymes.

In silico metabolic reconstructions are versatile computational tools for integrating multiple levels of bioinformatics data, facilitating interpretation of that data, and making functional predictions related to the metabolic behavior of the cell. We have recently developed an *in silico* constraint-based metabolic reconstruction for *C. thermocellum* based on available genome annotations, published phenotypic information, and specific biochemical assays. Furthermore, we have integrated the model with transcriptomic data from high-throughput mRNA sequencing. We show that this model refinement improves the quality of model predictions and provides insights into the internal metabolic flux profile of the cells during exponential growth on cellulose. Additionally, evaluation of existing genome sequence data within the context of this functional model is able to generate testable predictions about the accuracy and completeness of existing genome annotations and regulatory operon structure.

Constraint-based models for rational strain design

Constraint-based *in silico* metabolic reconstructions combined with flux balance analysis (FBA) have proven to be useful tools for directing metabolic engineering decisions[1] by capturing large-scale network topology information while generating predictions.



1 All metabolic reactions available to an organism according to genome annotation and biochemical evidence are compiled in a stoichiometric matrix, S, which is part of a genome-scale mass-balance problem.

2 Any given flux state can be represented as a vector, and the reaction matrix combined with boundary constraints define the borders of a solution space within which the flux state must always fall.

3 Boundary conditions are set based on observed substrate uptake and byproduct secretion rates. Flux balance analysis is then used to probe the resulting solution space by maximizing a cellular objective such as growth rate within the given constraints. The resulting vector describes the predicted reaction fluxes throughout the model.

Flux balance analysis [2]

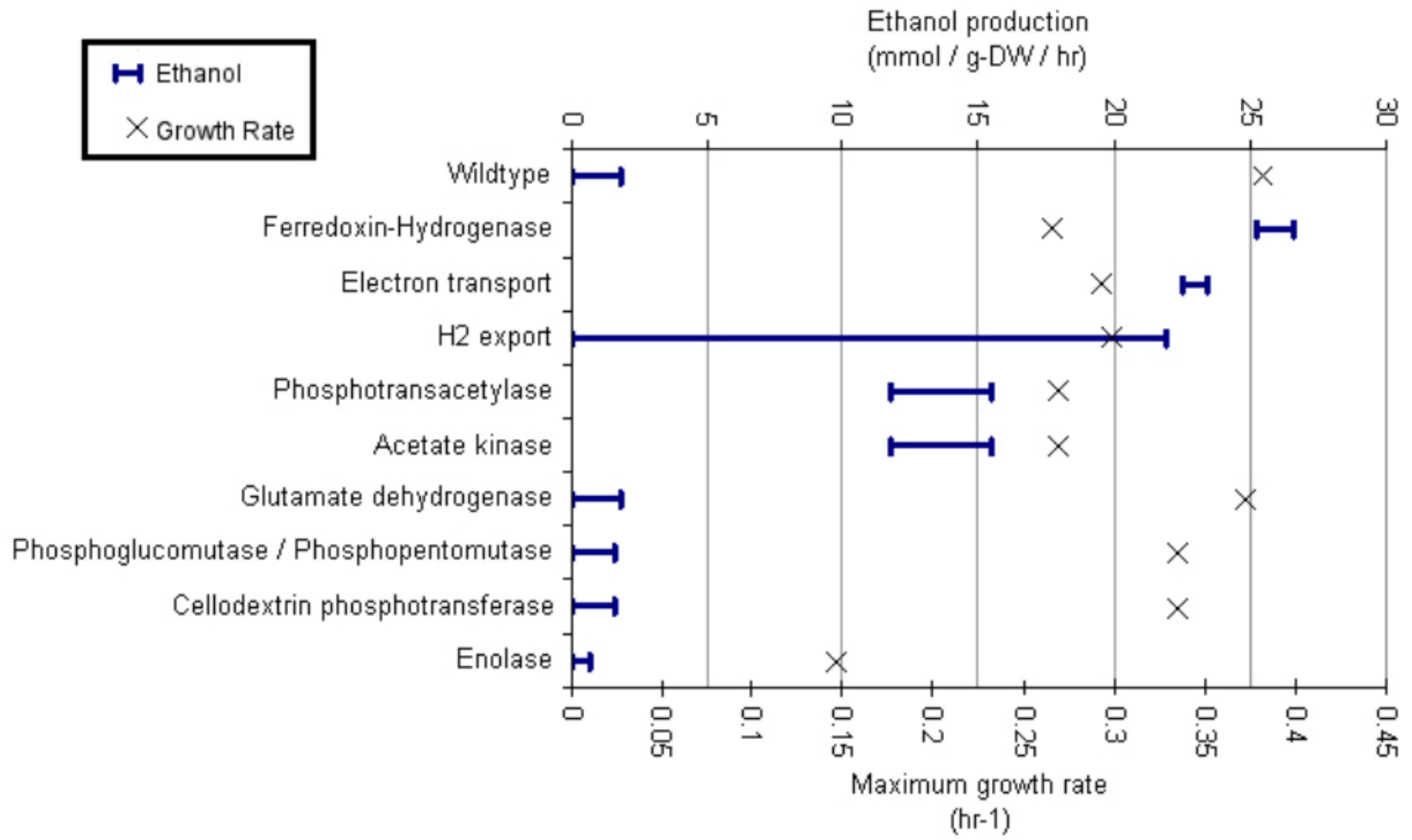
$$\max_{(fluxes)} (cellular\ \ objective)$$

Subject to:
– boundary constraints
– network stoichiometry
– thermodynamic constraints

Previous Work

A constraint-based genome-scale metabolic model has been constructed based on the published annotation of *Clostridium thermocellum's* genome as well as the incorporation of biochemical observation.[2] The statistics of the resulting model are shown in the table to the right. The model is unique in its incorporation of proteomics data to account for substrate-dependent production of the cellulosome. Potential genetic modifications predicted to increase ethanol yield based on FBA are shown below.

Genome size	**3.8 Mb**
ORFs	**3307**
Included genes	**432**
Enzyme complexes	**72**
Isozyme cases	**70**
Reactions (excluding exchanges)	**563**
Transport	56
Gene associated	463
Non-gene associated intracellular	61
Non-gene associated transports	37
Distinct metabolites	**529**



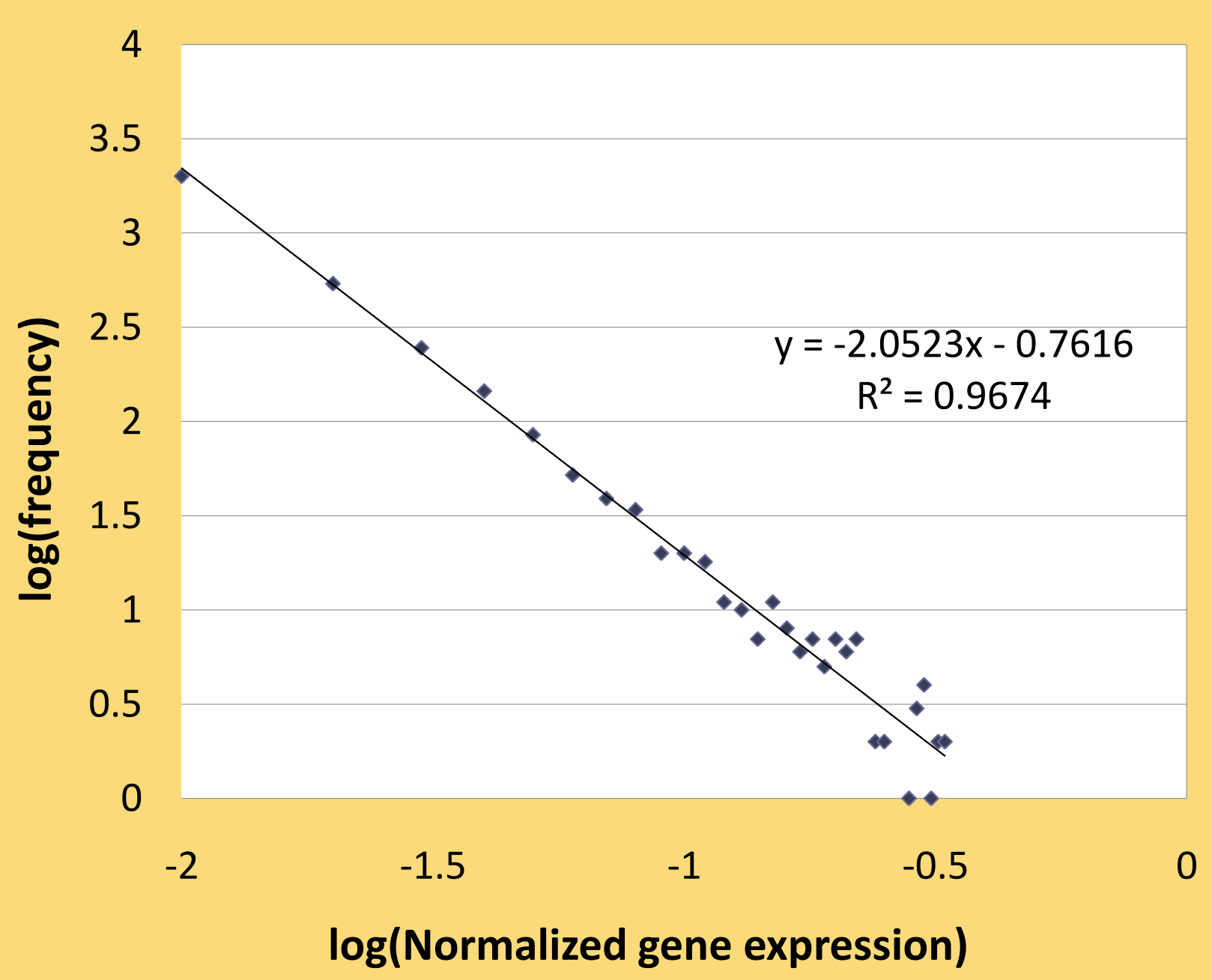
Single gene deletions for which increased ethanol production is predicted. Ethanol production is predicted as the range possible at the corresponding growth rate over alternate optimal solutions.

RNA-Seq analysis *C. thermocellum*

cDNA from a *C. thermocellum* culture grown in batch on minimal MJ medium[3] with 5g/L cellobiose was isolated and sequenced using the Roche 454 sequencer, resulting in over 230,000 unique reads with an average length of 367bp. These reads were blasted against the RefSeq database[4] using MegaBLAST[5] and filtered according to the following rules:

- min length = 30
- max e-value = 0.001
- min % identity = 95.00
- Subject id = NC_009012.1
  (C. thermocellum refseq code)

Distribution of expression levels across entire genome



The expression count of each gene locus was then normalized to gene length and the resulting distribution is shown above.

Integration of RNA-seq data with existing model

Relative gene expression levels (g) are converted to discrete gene states using the rule:

$$gene\ state = \begin{cases} -1 & g = 0 \\ 0 & 0 < g \le \gamma \\ 1 & \gamma < g \end{cases}$$

where γ is a specified cutoff value. Mixed integer linear programming (MILP) was then used to search for a flux state that maximized agreement with the determined gene states.[6] The table below compares model predictions from FBA with those determined by incorporating RNA-seq data along with actual fermentation measurements for growth on cellobiose.[7] The Euclidean distance from the measured "Reality" vector is shown for each of the predictions.

	FBA	With RNAseq	Reality
Biomass	0.54	**0.29**	0.29
Acetate	43.6	32.1	16.0
Cellobiose	**-12.8**	**-12.8**	-12.8
Fructose	**0.0**	**0.0**	0.0
CO_2	45.0	17.9	21.4
Ethanol	0.0	0.0	17.6
Formate	0.0	17.0	7.7
H_2	63.5	18.6	18.4
Succinate	0.0	1.0	0.0
Lactate	0.0	0.0	0.0
Distance	61.0	25.8	-

Conclusions

- RNA-seq analysis produces data that is especially suitable for integration with *in silico* metabolic reconstructions because of its discrete, absolute nature.
- Incorporation of RNA-seq data significantly improves the quality of predictions from *in silico* reconstructions, especially when the organism is relatively poorly understood or it has not undergone adaptive evolution and therefore using FBA to maximize growth rate is not necessarily appropriate.
- Incorporation of RNA-seq data into rational strain design decisions improves reliability by demonstrating the gene deletions which are likely to have the greatest impact.

References and Acknowledgements

The authors would like to thank J. Paul Brooks for his computational and operations expertise and the Nucleic Acids Research Facility at VCU for valuable discussions on cDNA preparation. This study was funded in part by a grant from the NASA Applied Information Systems Research Program (#08-AISR08-0008) to SSF.

1. Fong, S. S., Burgard, A. P., Herring, C. D., Knight, E. M., Blattner, F. R., Maranas, C. D., et al. (2005). In silico design and adaptive evolution ofescherichia coli for production of lactic acid. *Biotechnology and Bioengineering*, *91*(5), 643-648.
2. Roberts S, Gowen C, Brooks JP, Fong S: **Genome-scale metabolic analysis of Clostridium thermocellum for bioethanol production.** BMC Systems Biology 2010, **4**(1):31.
3. Johnson EA, Madia A, Demain AL: **Chemically Defined Minimal Medium for Growth of the Anaerobic Cellulolytic Thermophile Clostridium thermocellum.** Appl Environ Microbiol 1981, **41**(4):1060-1062.
4. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** Nucl Acids Res 2007, **35**(suppl_1):D61-65.
5. Zhang Z, Schwartz S, Wagner L, Miller W: **A Greedy Algorithm for Aligning DNA Sequences.** Journal of Computational Biology 2000, **7**(1-2):203-214.
6. Shlomi, T., Cabili, M. N., Herrgard, M. J., Palsson, B. Ø., & Ruppin, E. (2008). Network-Based prediction of human tissue-Specific metabolism. *Nature Biotechnology*, 26(9), 1003-1010.
7. Lamed, R. J., Lobos, J. H., & Su, T. M. (1988). Effects of Stirring and Hydrogen on Fermentation Products of Clostridium thermocellum. *Applied and Environmental Microbiology*, *54*(5), 1216-1221.

American Society of Microbiology 2010

Integration of RNA-Seq data with an *in silico* metabolic reconstruction of *Clostridium thermocellum* for rational strain design

Control Number: 4306

Poster Presentation Number: O-2282

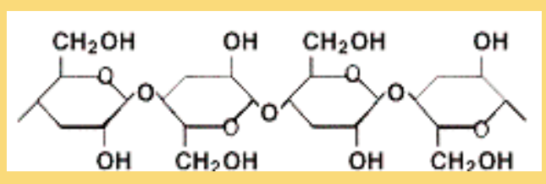
Chris M Gowen[1], **Stephen S Fong**[1,2]

[1]Department of Chemical and Life Science Engineering, Virginia Commonwealth University, Richmond, VA, USA
[2]Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA, USA

Contact:
gowencm@vcu.edu
(804) 827-7000 ext 414

Motivation

Clostridium thermocellum (ATCC 27405) is one of a number of organisms capable of **direct fermentation of cellulose to ethanol**, and its cellulolytic system is one of the most efficient known to researchers, but native ethanol productivity and yield is not currently sufficient for commercialization, partly due to diversion of carbon and electron flows towards competing fermentation end products like those shown below. Directed pathway engineering has the potential to improve yields drastically by controlling both carbon and electron flows in the cell.



Cellulose

Cellobiose

Fructose

⇨

CREDIT: DOE Joint Genome Institute
http://genome.jgi-psf.org/finished_microbes/cloth/cloth.home.html

⇨

Ethanol

Acetate

Formate

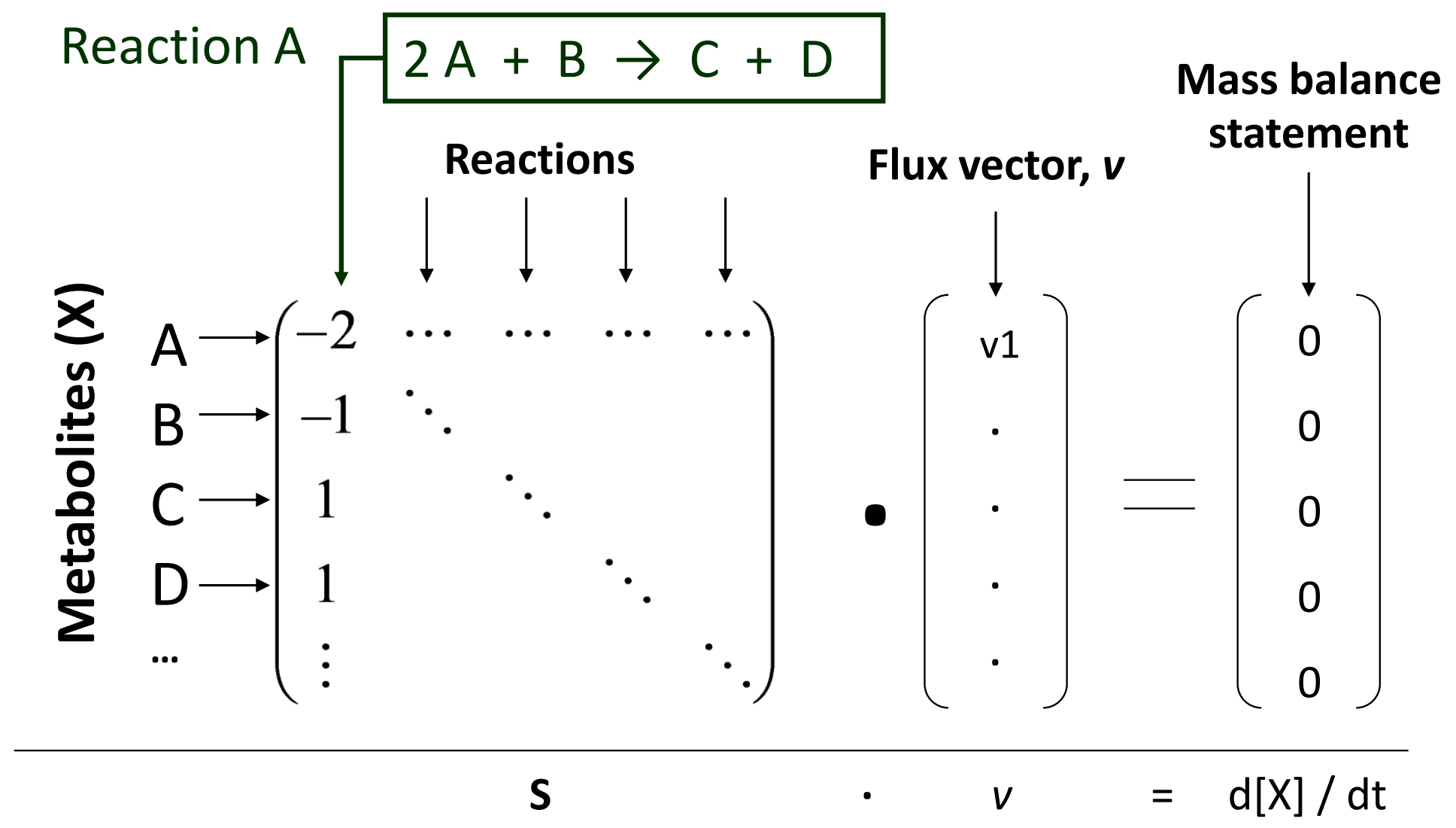
Lactate

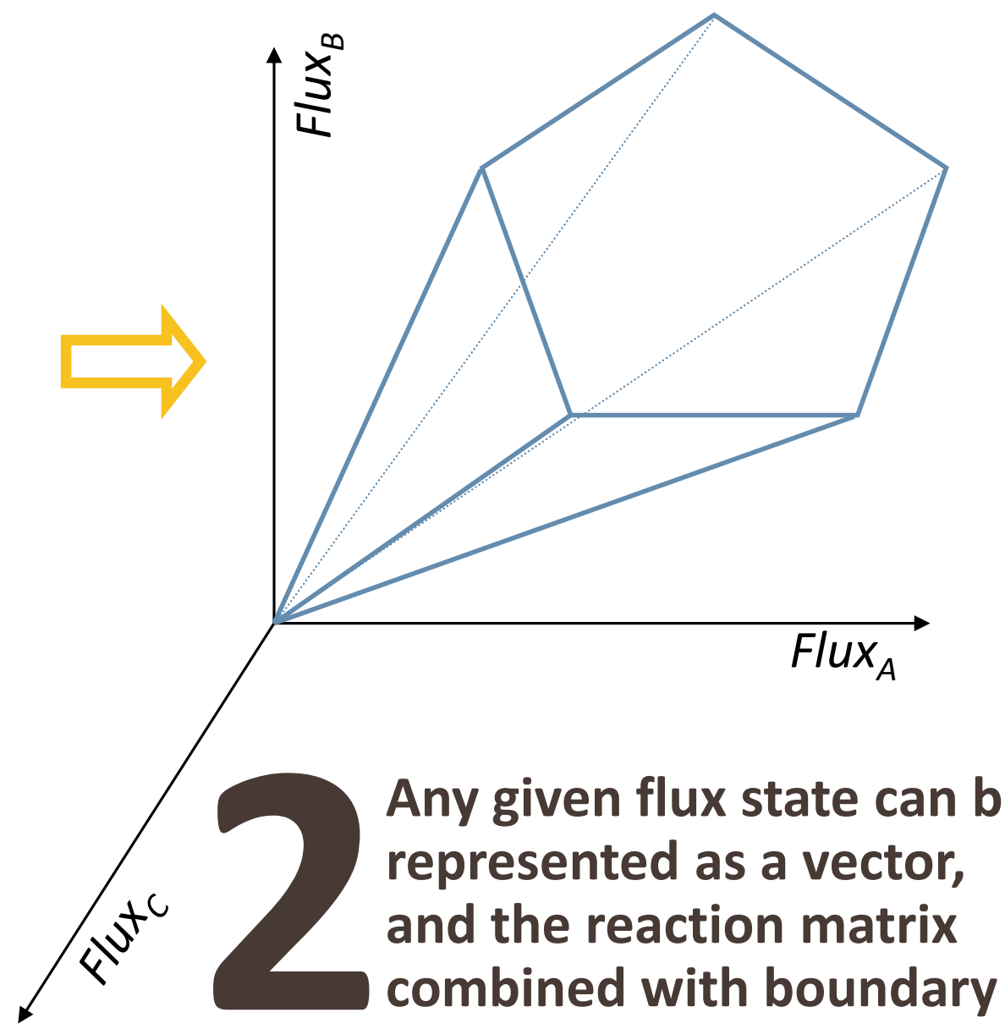
H_2

CO_2

Constraint-based models for rational strain design

Constraint-based *in silico* metabolic reconstructions have proven to be useful tools for directing metabolic engineering decisions by capturing large-scale network topology information while generating predictions





1 All metabolic reactions available to an organism according to genome annotation and biochemical evidence are compiled in a stoichiometric matrix, S, which is part of a genome-scale mass-balance problem.

2 Any given flux state can be represented as a vector, and the reaction matrix combined with boundary constraints define the borders of a solution space within which the flux state must always fall.

3 Boundary conditions are set based on observed substrate uptake and byproduct secretion rates. Flux balance analysis is then used to probe the resulting solution space by maximizing a cellular objective such as growth rate within the given constraints. The resulting vector describes the predicted reaction fluxes throughout the model.

Flux balance analysis [2]

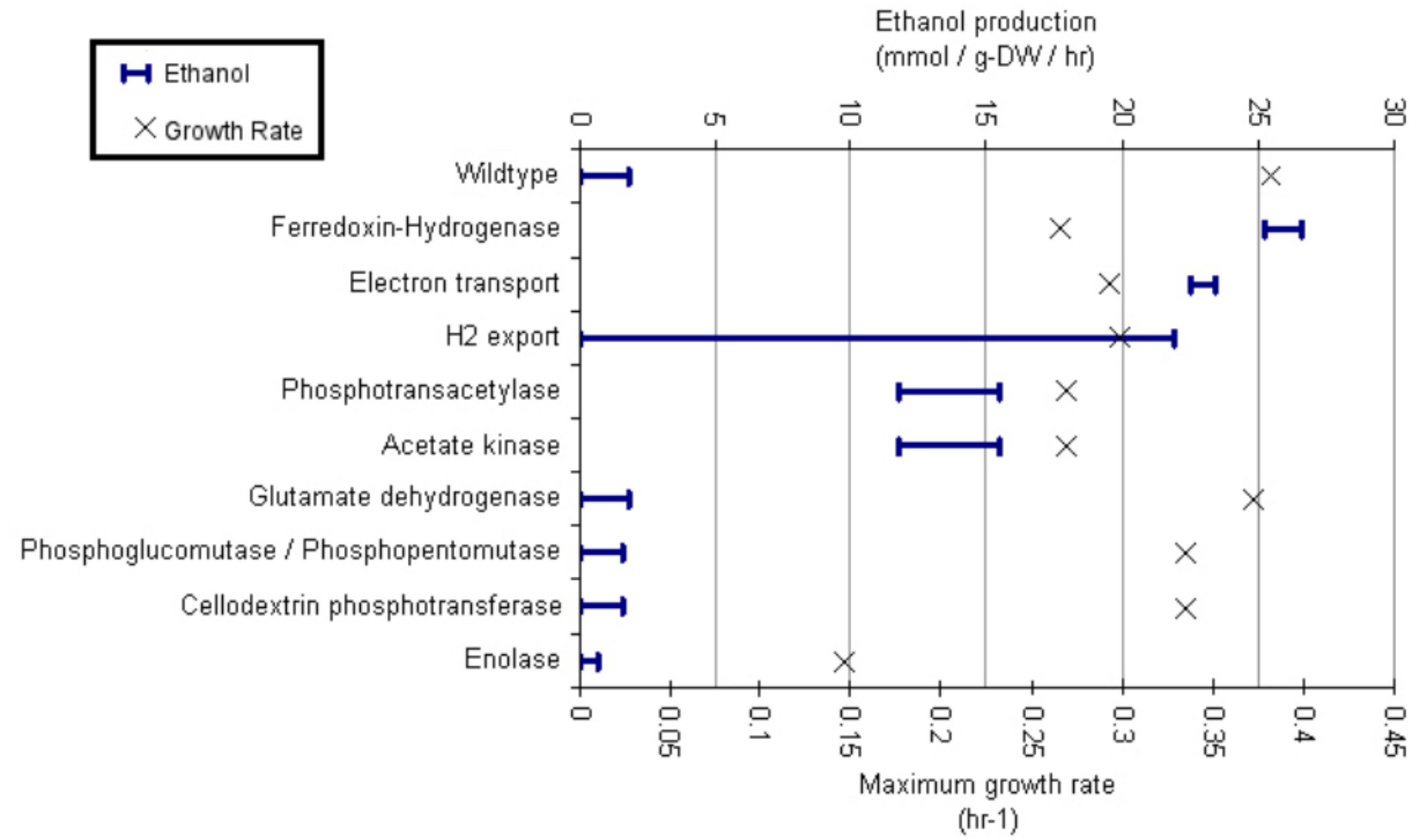
$$\max_{(fluxes)} (cellular\ \ objective)$$

Subject to:
– boundary constraints
– network stoichiometry
– thermodynamic constraints

Previous Work

A constraint-based genome-scale metabolic model has been constructed based on the published annotation of *Clostridium thermocellum's* genome as well as the incorporation of biochemical observation [cite]. The statistics of the resulting model are shown in the table to the right. The model is unique in its incorporation of proteomics data to account for substrate-dependent production of the cellulosome. Potential genetic modifications predicted to increase ethanol yield based on FBA are shown below.

Genome size	**3.8 Mb**
ORFs	**3307**
Included genes	**432**
Enzyme complexes	**72**
Isozyme cases	**70**
Reactions (excluding exchanges)	**563**
Transport	56
Gene associated	463
Non-gene associated intracellular	61
Non-gene associated transports	37
Distinct metabolites	**529**



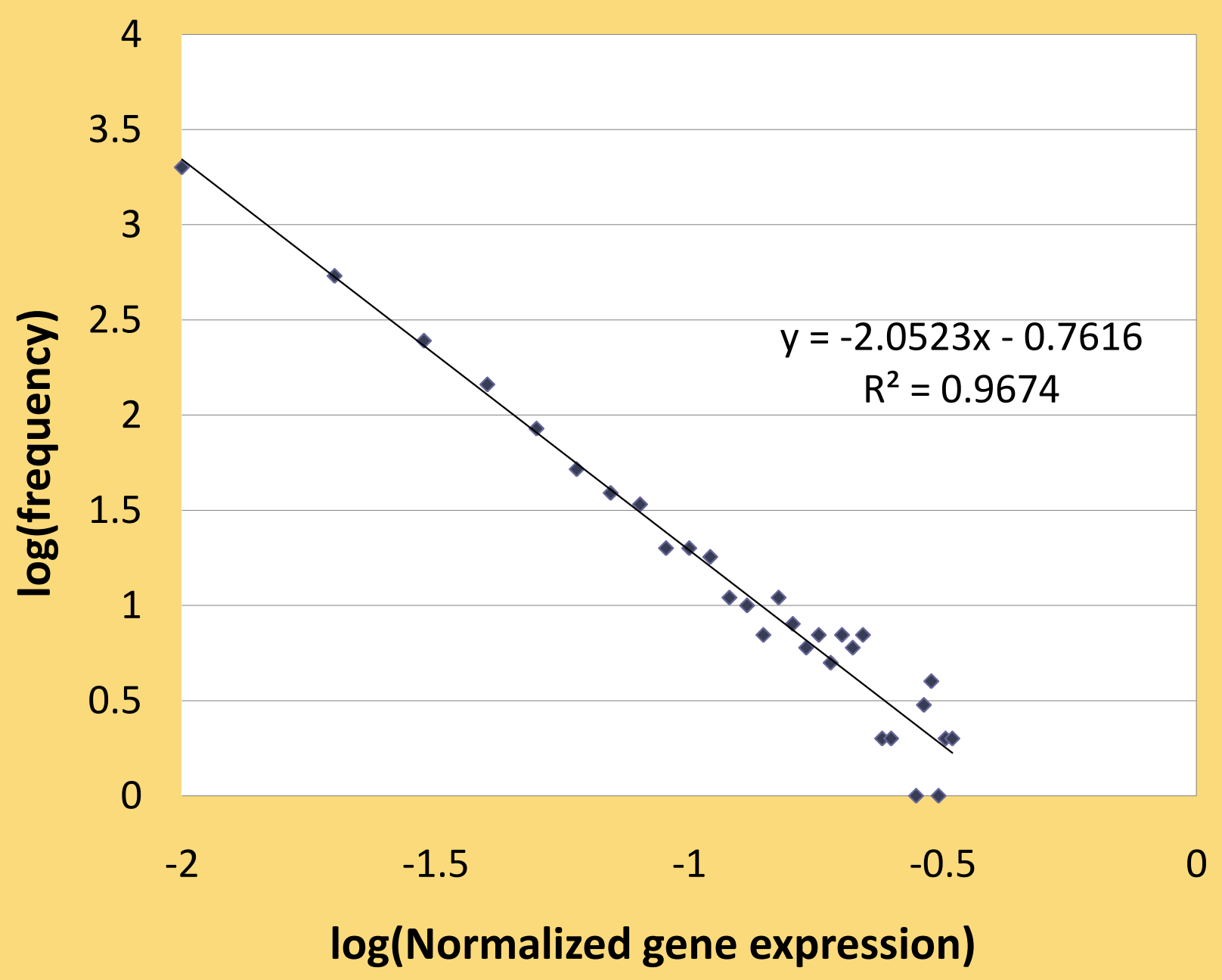
Single gene deletions for which increased ethanol production is predicted. Ethanol production is predicted as the range possible at the corresponding growth rate over alternate optimal solutions.

RNA-Seq analysis *C. thermocellum*

cDNA from a *C. thermocellum* culture grown in batch on minimal MJ medium[4] with 5g/L cellobiose was isolated and sequenced using the Roche 454 sequencer, resulting in over 230,000 unique reads with an average length of 367bp. These reads were blasted against the RefSeq database[5] using MegaBLAST[6] and filtered according to the following rules:

- min length = 30
- max e-value = 0.001
- min % identity = 95.00
- Subject id = NC_009012.1
  (C. thermocellum refseq code)

Distribution of expression levels across entire genome



The expression count of each gene locus was then normalized to gene length and the resulting distribution is shown above.

Integration of RNA-seq data with existing model

Relative gene expression levels (g) are converted to discrete gene states using the rule:

$$gene\ state = \begin{cases} -1 & g = 0 \\ 0 & 0 < g \le \gamma \\ 1 & \gamma < g \end{cases}$$

where γ is a specified cutoff value. Mixed integer linear programming (MILP) was then used to search for a flux state that maximized agreement with the determined gene states [CITE]. The table below compares model predictions from FBA with those determined by incorporating RNA-seq data along with actual fermentation measurements for growth on cellobiose [cite]. The Euclidean distance from the measured "Reality" vector is shown for each of the predictions.

	FBA	With RNAseq	Reality
Biomass	0.54	**0.29**	0.29
Acetate	43.6	32.1	16.0
Cellobiose	**-12.8**	**-12.8**	-12.8
Fructose	**0.0**	**0.0**	0.0
CO_2	45.0	17.9	21.4
Ethanol	0.0	0.0	17.6
Formate	0.0	17.0	7.7
H_2	63.5	18.6	18.4
Succinate	0.0	1.0	0.0
Lactate	0.0	0.0	0.0
Distance	61.0	25.8	-

Conclusions

- RNA-seq analysis produces data that is especially suitable for integration with *in silico* metabolic reconstructions because of its discrete, absolute nature.

- Incorporation of RNA-seq data significantly improves the quality of predictions from *in silico* reconstructions, especially when the organism is relatively poorly understood or it has not undergone adaptive evolution and therefore using FBA to maximize growth rate is not necessarily appropriate.

- Incorporation of RNA-seq data into rational strain design decisions improves reliability by demonstrating the gene deletions which are likely to have the greatest impact.

References and Acknowledgements

The authors would like to thank J. Paul Brooks for his computational and operations expertise, …

•Edwards, J. S., Covert, M., & Palsson, B. Ø. (2002). Metabolic modelling of microbes: the Flux-balance approach. *Environmental Microbiology*, 4(3), 133-140.

•Lamed, R. J., Lobos, J. H., & Su, T. M. (1988). Effects of Stirring and Hydrogen on Fermentation Products of Clostridium thermocellum. *Applied and Environmental Microbiology*, *54*(5), 1216-1221.

•Shlomi, T., Cabili, M. N., Herrgard, M. J., Palsson, B. Ø., & Ruppin, E. (2008). Network-Based prediction of human tissue-Specific metabolism. *Nature Biotechnology*, 26(9), 1003-1010.

•Burgard, A. P., Pharkya, P., & Maranas, C. D. (2003). Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and Bioengineering*, *84*(6), 647-657.

•Fong, S. S., Burgard, A. P., Herring, C. D., Knight, E. M., Blattner, F. R., Maranas, C. D., et al. (2005). In silico design and adaptive evolution ofescherichia coli for production of lactic acid. *Biotechnology and Bioengineering*, *91*(5), 643-648.

•Desai, S., Guerinot, M., & Lynd, L. (2004). Cloning of l-Lactate dehydrogenase and elimination of lactic acid production via gene knockout in thermoanaerobacterium saccharolyticum jw/sl-Ys485. *Applied Microbiology and Biotechnology*, *65*(5), 600-605.

American Society of Microbiology 2010