Title

Genome-scale Models of Metabolism and Gene Expression : : Construction and Use for Growth Phenotype Prediction

Permalink

https://escholarship.org/uc/item/8zq4p227

Author

Lerman, Joshua Adam

Publication Date

2014

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Genome-scale Models of Metabolism and Gene Expression: Construction and Use for Growth Phenotype Prediction

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioinformatics & Systems Biology

by

Joshua Adam Lerman

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Milton H. Saier, Jr., Co-Chair
Professor Philip E. Bourne
Professor Terence Hwa
Professor Victor Nizet

2014

The dissertation of Joshua Adam Lerman is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2014

DEDICATION

To my mother and father, for your love, guidance, and all the sacrifices you made for Justin, Rachel, and I.

To Lauren, for your love and all those times I told you, “One sec.”

To the loving memory of Bubby.

EPIGRAPH

Tony Stark was able to build this in a cave!
WITH A BOX OF SCRAPS!!
—Obadiah Stane, *Iron Man*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

There are so many people deserving of my thanks. First and foremost, I wish to thank Bernhard Palsson and everyone part of the Systems Biology Research Group at UCSD. We are an amazing assembly of researchers dedicated to systematizing biological discovery, and I believe our impact will be felt in ways we can't even imagine today. I want to start by thanking those that contributed directly (as co-authors) to the manuscripts that form the basis of my dissertation: Roger Chang, Harish Nagarajan, Daniel Hyduke, Haythem Latif, Dae-Hee Lee, Irene Lam, Nathan Lewis, Nicole Fong, Teddy O'Brien, Jeff Orth, Yu Qui, Pep Charusanti, Vasiliy Portnoy, Yekaterina Tarasova, Karsten Zengler, and Bernhard Palsson. Dr. Hyduke, thank you for opening my eyes to the world of Apple products, teaching me how to program in Python, and how and why one should use SVN (and later Git). Haythem, thanks for not getting too fed up working with *Thermotoga maritima*. As a result of your efforts, we had clean and reliable data to work with. Teddy, thank you for being a great friend, mentee, and later my go-to person for deriving constraints and understanding the output of the models we built together. Thanks also to those who came before me, and laid the groundwork upon which my dissertation rests: Timothy Allen and Ines Thiele.

I also had the pleasure of working closely with Jan Schellenberger, Joanne Liu, Steve Federowicz, Ali Ebrahim, Hojung Nam, and Zak King. Jan deserves a special thanks for teaching me the basics of mixed integer linear programming formulation. The project we worked on together (Mini-ME) ran amuck about 6 months after we started it in my second year, so he never got the credit he deserved. Thanks also to Nikolaus Sonnenschein, Aarash Bordbar, Daniel Zielinski, and Donghyuk Kim. You always had open minds and doors I could rely on. Thanks to Adam Feist, José Utrilla Carreri, and Miguel Campodonico for invigorating discussions on the topic of metabolic engineering and modeling. I hope we can collaborate in the future. Adam, I want to additionally thank you for teaching me how to communicate results more clearly and effectively. Thanks also to anyone who ever put delicious food on the center table (Alessandra Gallina, Mallory Embree, and Jennifer Levering come to mind). Finally, a big thanks to Marc Abrams,

Yana Campen, and Kathy Andrews for providing the administrative support for such a large group.

Next, I'd like to thank my dissertation committee. Dr. Palsson, thank you for being so approachable and for providing major course corrections as necessary. Thank you for helping me think bigger and to see the *fundamental* significance in the work we pursued together. Dr. Bourne, thank you so much for a fruitful rotation project centered on repurposing Raloxifene (with Lei Xie and Shannan Ho Sui), and for serving on my dissertation committee. Drs. Hwa, Nizet, and Saier, thanks for contributing periodically at my annual reviews. Although not on my committee, I would also like to thank Dr. Trey Ideker. I enjoyed serving as his teaching assistant for BENG 183.

Without external collaborators, none of this work would have been possible. I thank Thorsten Koch, Matthias Miltenberger, Ambros Gleixner, Ronan Fleming, Michael Saunders, Yuekai Sun, Martina Ma, Daniel Espinoza, Elizabeth Wong, and Bill Cook. These people helped develop algorithmic methods and software tools for solving stiff linear programs. They got me over a huge hurdle to conducting this research. I thank Ronan for writing the grant that initially funded my research. Thanks also to the amazing quantitative proteomics team at the Pacific Northwest National Laboratory (including Joshua Adkins, Alexandra Schrimpe-Rutledge, and Richard Smith). I also wish to thank Heather Mottaz-Brewer for assistance in proteome sample processing.

I wish to thank all the classmates I entered the Bioinformatics and Systems Biology program with: Marcus Kinsella, Stephanie Huelga, Shawn Yost, Allan Wu, Lance Hepler, Jason Bechtel, Gordon Bean, Max Shok, and Boyko Kakaradov. We had an amazing first year that will always be remembered! I would also like to acknowledge Nitin Udpa, Nisha Rajagopal, Josué Pérez, Colin Haynes, and Isabel Canto who welcomed me to San Diego and showed me a good time over the years. Special thanks to the founders and leaders of the Bioinformatics and Systems Biology program (Alexander Hoffmann, Shankar Subramaniam, Pavel Pevzner, Vineet Bafna, and Bing Ren) and its graduate coordinators (Jan Lenington, Laura Gracia, and Kathleen Kazules).

I especially wish to thank my family (Mom, Dad, Justin, Rachel, and my extended family) for supporting me through my formal education years, culminating in the defense of my dissertation. Thanks also to Lauren Sadler for standing beside me and encouraging me in times of need.

Finally, I'd like to thank American taxpayers for funding my research. The dollars provided helped support grant programs at the National Institutes of Health (NIH) and the Department of Energy (DOE). My work was most closely supported by an NIH training grant to the Bioinformatics and Systems Biology program at UCSD, DOE Awards DE-FG02-09ER25917 and DE-FG02-08ER64686, and NIH U01 GM102098. I also used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231.

Chapter 1 is in part adapted from Lerman, J., and Palsson, B. O. (2010). Microbiology. Topping off a multiscale balancing act. Science 330, 1058-1059 (perspective). The dissertation author was the primary author of this perspective.

Chapter 2 in full is a reprint of a published manuscript: Latif H*, Lerman JA*, Portnoy VA, Tarasova Y, Nagarajan H, et al. (2013) The Genome Organization of *Thermotoga maritima* Reflects Its Lifestyle. PLoS Genet 9(4): e1003485. doi:10.1371/journal.pgen.1003485. * indicates equal contribution. The dissertation author was the primary author of this paper responsible for the research. The other authors were Haythem Latif (equal contributor), Vasiliy A. Portnoy, Yekaterina Tarasova, Harish Nagarajan, Alexandra C. Schrimpe-Rutledge, Richard D. Smith, Joshua N. Adkins, Dae-Hee Lee, Yu Qiu, and Karsten Zengler.

Chapter 3 in full is a reprint of a published manuscript: Lerman, J.A. *et al.* *In silico* method for modelling metabolism and gene product expression at genome scale. *Nat. Commun.* 3:929 doi: 10.1038/ncomms1928 (2012). The dissertation author was the primary author of this paper responsible for the research. The other authors were Daniel R. Hyduke (equal contributor), Haythem Latif, Vasiliy

A. Portnoy, Nathan E. Lewis, Jeffrey D. Orth, Alexandra C. Schrimpe-Rutledge, Richard D. Smith, Joshua N. Adkins, Karsten Zengler, and Bernhard Ø. Palsson.

Chapter 4 in full is a reprint of a published manuscript: O'Brien EJ*, Lerman JA*, Chang RL, Hyduke DR, Palsson BO. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol Syst Biol. 2013 Oct 1;9:693. doi: 10.1038/msb.2013.52. The dissertation author was the primary author of this paper responsible for the research. The other authors were Edward J. O'Brien (equal contributor), Roger L. Chang, Daniel R. Hyduke, and Bernhard Ø. Palsson.

Chapter 5 in full is a reprint of a published manuscript: Fong, N. L., Lerman, J. A., Lam, I., Palsson, B. O. and Charusanti, P. (2013), Reconciling a *Salmonella enterica* metabolic model with experimental data confirms that overexpression of the glyoxylate shunt can rescue a lethal *ppc* deletion mutant. FEMS Microbiology Letters, 342: 6269. doi:10.1111/1574-6968.12109. The dissertation author was the second author of this paper, responsible for the computational analysis that inspired the research. The other authors were Nicole L. Fong, Irene Lam, Bernhard Ø. Palsson, and Pep Charusanti.

VITA

2008	B. S. in Biomedical Engineering, The Johns Hopkins University
2008	B. S. in Applied Mathematics and Statistics, The Johns Hopkins University
2014	Ph. D. in Bioinformatics & Systems Biology, University of California, San Diego

PUBLICATIONS

Federowicz, S.A., Kim, D, Ebrahim, A, **Lerman, J.A.**, Nagarajan, H, Cho, B, Zengler, K, Palsson, B.O. Determining the control circuitry of redox metabolism at the genome-scale. Submitted.

Monk, J.M., Charusanti, P, Aziz, R.K., **Lerman, J.A.**, Premyodhin, N., Orth, J.D., Fesit, A.M., Palsson, B.O. Genome-scale metabolic reconstructions of multiple *E. coli* strains highlight strain-specific adaptations to nutritional environments. In press at Proc. Natl. Acad. Sci. USA. Tracking #: 2013-07797R.

O'Brien, E. J.*, **Lerman, J. A.***, Chang, R. L., Hyduke, D. R., and Palsson, B. O. (2013). Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol Syst Biol 9, 693.

Ebrahim, A., **Lerman, J. A.**, Palsson, B. O., and Hyduke, D. R. (2013). COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Syst Biol 7, 74.

Latif, H.*, **Lerman, J. A.***, Portnoy, V. A., Tarasova, Y., Nagarajan, H., Schrimpe-Rutledge, A. C., Smith, R. D., Adkins, J. N., Lee, D. H., Qiu, Y., and Zengler, K. (2013). The genome organization of *Thermotoga maritima* reflects its lifestyle. PLoS Genet 9, e1003485.

Fong, N. L., **Lerman, J. A.**, Lam, I., Palsson, B. O., and Charusanti, P. (2013). Reconciling a *Salmonella enterica* metabolic model with experimental data confirms that overexpression of the glyoxylate shunt can rescue a lethal ppc deletion mutant. FEMS Microbiol Lett 342, 62-69.

Nam, H.*, Lewis, N. E.*, **Lerman, J. A.**, Lee, D. H., Chang, R. L., Kim, D., and Palsson, B. O. (2012). Network context and selection in the evolution to enzyme specificity. Science 337, 1101-1104.

Ho Sui, S. J.*, Lo, R.*, Fernandes, A. R., Caulfield, M. D., **Lerman, J. A.**, Xie, L., Bourne, P. E., Baillie, D. L., and Brinkman, F. S. (2012). Raloxifene attenuates *Pseudomonas aeruginosa* pyocyanin production and virulence. Int J Antimicrob Agents 40, 246-251.

Lerman, J. A.*, Hyduke, D. R.*, Latif, H., Portnoy, V. A., Lewis, N. E., Orth, J. D., Schrimpe-Rutledge, A. C., Smith, R. D., Adkins, J. N., Zengler, K., and Palsson, B. O. (2012). *In silico* method for modelling metabolism and gene product expression at genome scale. Nat Commun 3, 929.

Charusanti, P., Chauhan, S., McAteer, K., **Lerman, J. A.**, Hyduke, D. R., Motin, V. L., Ansong, C., Adkins, J. N., and Palsson, B. O. (2011). An experimentally-supported genome-scale metabolic network reconstruction for *Yersinia pestis* CO92. BMC Syst Biol 5, 163.

Orth, J. D., Conrad, T. M., Na, J., **Lerman, J. A.**, Nam, H., Feist, A. M., and Palsson, B. O. (2011). A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism–2011. Mol Syst Biol 7, 535.

Lerman, J., and Palsson, B. O. (2010). Microbiology. Topping off a multiscale balancing act. Science 330, 1058-1059.

Lewis, N. E., Hixson, K. K., Conrad, T. M., **Lerman, J. A.**, Charusanti, P., Polpitiya, A. D., Adkins, J. N., Schramm, G., Purvine, S. O., Lopez-Ferrer, D., Weitz, K. K., Eils, R., Konig, R., Smith, R. D., and Palsson, B. O. (2010). Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. Mol Syst Biol 6, 390.

Gehlbach, P. L.*, Benson, B. C.*, Cortes, H. M.*, Davenport, M. S.*, Harrison, R. M.*, Hwang, K.*, Kapp, G. M.*, Lee, C. Y.*, **Lerman, J. A.***, Li, T.*, and Wong, J. C.* (2007). Multifunctional neck brace. United States Patent Application. Application number: 11/798,041. Publication number: US 2008/0004556 A1. Filing date: May 9, 2007.

ABSTRACT OF THE DISSERTATION

Genome-scale Models of Metabolism and Gene Expression: Construction and Use for Growth Phenotype Prediction

by

Joshua Adam Lerman

Doctor of Philosophy in Bioinformatics & Systems Biology

University of California, San Diego, 2014

Professor Bernhard Ø. Palsson, Chair
Professor Milton H. Saier, Jr., Co-Chair

In this dissertation, I develop ME-Models. ME-Models are genome-scale models that seamlessly integrate metabolic and gene product expression pathways. They can be used to compute optimal cellular states for growth in steady-state environments. They take as input the availability of nutrients to the cell and produce experimentally testable predictions for: (1) the cell's maximum growth rate (μ*) in the specified environment, (2) substrate uptake/by-product secretion rates at μ*, (3) metabolic fluxes at μ*, and (4) gene product expression levels at μ*. Unlike previous genome-scale models, ME-Models explicitly account for the production of all RNAs and proteins. I first build a prototype ME-Model for the simple microorganism *Thermotoga maritima.* The *T. maritima* genome was sequenced in 1999, and needed correction and complete re-annotation. I developed a framework drawing on multi-omic data to annotate genomic features involved in transcription, translation, and regulation. These features in *T. maritima* were found to display distinctive properties. In addition to basic characterization, the re-annotation was used to build the *T. maritima* ME-Model. Reactions to produce all the RNAs and proteins were added to its metabolic model, and metabolism was linked to gene expression through 'coupling constraints.' In the second part of this dissertation, the method was extended to *E. coli.* Backed by the wealth of phenotypic information available for this organism, I was able to firmly support the statement that ME-Models extend and refine microbial growth phenotype prediction. Next, a previous model predicted a *ppc* knockout of *Salmonella enterica* serovar Typhimurium would grow, but it did not experimentally. Ultimately, network modeling pinpointed the cause of the discrepancy (the inability of cells to route flux through the glyoxylate shunt when *ppc* is removed). The *ppc* project illustrates the importance of considering expression and regulation in genome-scale models. Finally, I demonstrate (albeit preliminarily) that ME-Models begin to bridge systems and synthetic biology approaches for engineering life.

Chapter 1

Introduction

1.1 Life, constraints, and being good enough

What is *life*, exactly?

“Life is about the physical embodiment of knowledge.”
– David Deutsch

“You can’t clone a mountain.”
– Sydney Brenner

As you may be able to tell from the quotations above, I favor a view of life that appreciates that cells are exquisite molecular information storage and processing devices. A cell’s long-term information storage device is its genome. Over time, random changes to the genome produce phenotypic variation, which leads to the emergence of cells that thrive in particular environmental niches. In the absence of competition for limited resources, cells that are ‘good enough’ to merely survive are evolutionary winners. But often it’s the case that competition is fierce, and so what’s good enough today will not be in the future. Fast-forward a few billion years and you can see why we might expect living organisms to be as complex as they are. To keep up with the competition, some cells taught themselves how to operate very close to the boundary separating possible and impossible (hard constraints imposed by the physical laws of the universe). ‘Good enough’ became more and more demanding as time went on.

Then the first full genome sequences of microbes came online in the mid-1990s. We could finally see evolutionary winners and their solutions to fundamental biological problems. And we could build computer models and use them to ask whether cells were growing optimally. Usually, but sometimes the answer was "No." Then again, were we even aware of the evolutionary trade-offs underlying the answer to this question?

1.2 Systems microbiology and its promise

I illustrate systems microbiology in a nutshell in Figure 1.1. Evolution selects among the phenotypes that arise when a cell with a given genotype is placed in a given environment. The fittest cells best optimize multiple trade-offs (not shown). Cyanobacteria are pictured to the right of the top box in Figure 1.1 to serve as a reminder that cells can actively shape their own environments; Cyanobacteria are believed to have oxygenated Earth's atmosphere.

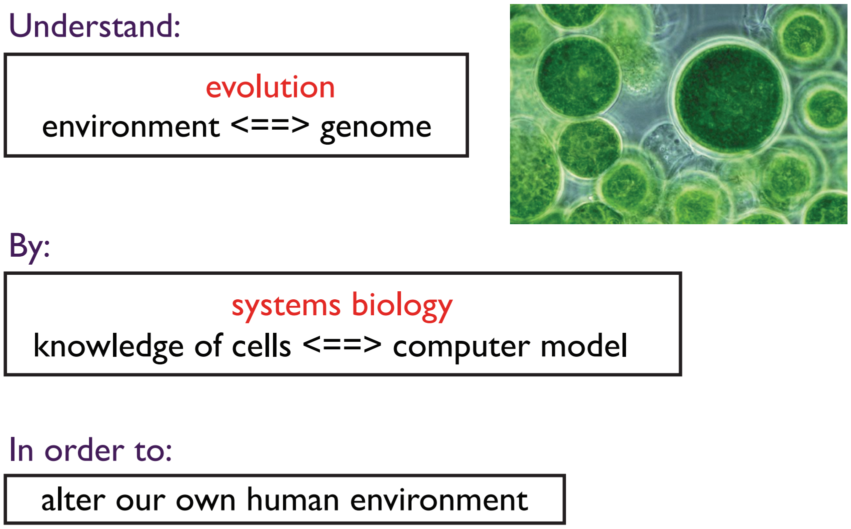


Figure 1.1: **Systems microbiology in a nutshell.** Cyanobacteria image taken from http://sgevurtz.blogspot.com/.

The promise of systems microbiology is shown in the middle box in Figure 1.1. The promise is as follows: If you systematically organize all available knowledge for a particular biological system, the tool you create in the process promises to allow you to learn more about the system. Usually, the tool takes the form of a database and/or a detailed computer model, the specifics of which I will address in the next section.

Systems microbiology has many implications for applied microbiology and synthetic biology. We eventually accumulate enough knowledge about the biological system that we are well-positioned to re-engineer it. Applications typically relate to our desire as a species to alter our own environment, whether it be ridding it of a pathogen or converting one of our waste streams to cellulosic ethanol (see final box of Figure 1.1).

1.3 Systems microbiology as a four-step procedure: Then and Now

When I started my PhD in late 2008, the use of experimental and computational techniques in a synergistic, iterative process had already gained acceptance as the optimal method for understanding the behavior of biochemical systems, ranging from reactions in a single cell to communities and ecosystems.

To practice systems microbiology, one can employ a four-step procedure [1]: (i) the enumeration of the biological components that make up a biological process, (ii) the reconstruction of the network of interactions among these components, (iii) the application of physicochemical equality constraints such as mass and energy balance and the steady-state assumption to determine network capabilities in a simulated environment with specified boundary conditions, and (iv) the comparison of computed network properties with actual phenotypic observations. The procedure is iterative since you learn by going through the four steps, especially when it can be resolved why a computed phenotype does not match an actual phenotype.

As I wrap up my PhD in late 2013, the same four steps remain in place. The difference is that each of the steps is now taken to the extreme. Many more cellular parts and biological processes are considered. As a result, thousands of cellular interactions that were either missing or implicit are now present and explicit. We also have better and more general constraints on network behavior. Finally, there are much improved experimental methods and data sources to validate or invalidate computed functional states of cells.

1.3.1 Then: M-Models (c. 2008)

Genome-scale metabolic models (termed as M-Models) can be built by reconstructing the full complement of metabolic reactions in an organism. The metabolic network reconstruction process is now at an advanced stage of development and has been translated into a 96-step standard operating procedure [2]. M-Models capture basic knowledge of reaction stoichiometry (e.g. '1 A+ 1 B $\rightarrow$ 1 C'). Each reaction is assigned a 'gene-protein-reaction relationship,' or a GPR. GPRs are Boolean logic statements (e.g. '(gene 1 and gene 2) or (gene 3 and gene 4)') that dictate which sets of genes are required to be present for a reaction to carry flux.

M-Models have found a wide range of applications, particularly for model organisms such as *Escherichia coli* [3]. M-Models are great, but they can be improved. The 8 biggest weaknesses of M-Models for microbes are as follows:

1. The cell composition and energy requirements (both growth and non-growth associated) are fixed vs. free variables

2. Absolute rates (such as growth rates) cannot be predicted unless substrate uptake and by-product secretion rates are specified

3. GPRs bridge genotype to phenotype, but insufficiently for many applications

4. Enzyme kinetics and regulation (transcriptional or metabolic) are not accounted for, even though they can significantly influence reaction fluxes

5. Few predictions can be directly experimentally validated (notable exceptions: prediction of growth or no-growth on different carbon sources/media, central carbon fluxes, and gene essentiality calls)

6. Very limited spatial resolution

7. No temporal resolution (methods such as dynamic flux balance analysis are usually inadequate for applications)

8. Missing information: Metabolite damage, enzyme promiscuity, and spontaneous side reactions (all unaccounted for) have major implications for metabolic modeling and engineering

Additionally, when using an M-Model there is often an inherent optimality assumption. This is necessary because the system is underdetermined as it is specified. In order to get around this problem, you have to assume you know something about what the cell has programmed itself to do over millions of years. Often, we assume the cell maximizes its growth rate, minimizes the total flux through all reactions (operates parsimoniously), and/or maximizes energy (ATP) generation. These assumptions get us closer to reality for some growth conditions [4, 5], but severely limit the types of predictions that can be made. Even when these assumptions are made, we may still not be left with a unique flux prediction. In these cases, the best we can do is select a solution randomly.

1.3.2 Now: ME-Models (c. 2013)

M-Models can be extended to include the process of gene expression (termed as ME-Models because they are integrated models of metabolism and gene expression). ME-Models are described in detail in Chapters 3 and 4, so I won't say much here. The basic idea is to explicitly account for the production of all RNAs and proteins. Once all the RNAs and proteins are produced, metabolism can be linked to gene expression through additional constraints called 'coupling constraints' (detailed later). As a result, growth phenotype prediction is greatly extended and refined.

In a major departure from the past, the construction and application of ME-Models depends heavily on omics data analysis and integration. Omics data sets describing virtually all biomolecules in the cell are now available. These data can be generally classified into 3 distinct categories [6] (see Figure 1.2). All this

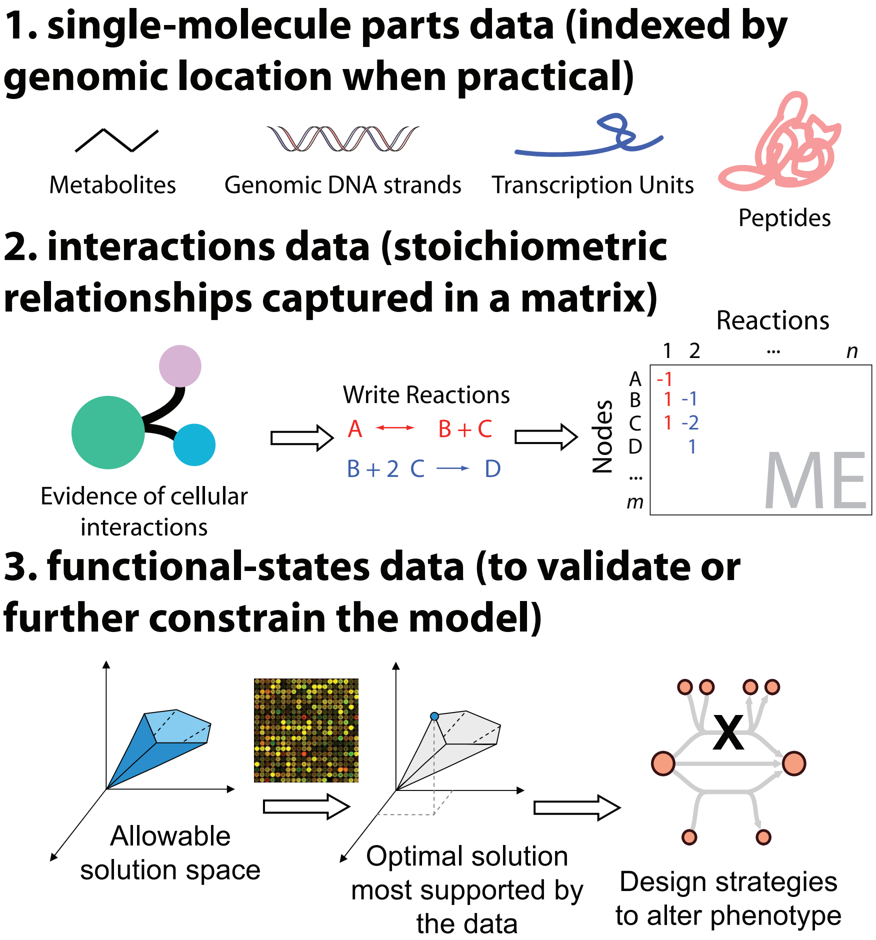

1. single-molecule parts data (indexed by genomic location when practical)
Metabolites
Genomic DNA strands
Transcription Units
Peptides
2. interactions data (stoichiometric relationships captured in a matrix)
Evidence of cellular interactions
Write Reactions
A ⟷ B + C
B + 2 C ⟶ D
Reactions
1 2 ... n
Nodes
A B C D ... m
-1
1 -1
1 -2
1
ME
3. functional-states data (to validate or further constrain the model)
Allowable solution space
Optimal solution most supported by the data
X
Design strategies to alter phenotype

Figure 1.2: **Types of omics data and their uses for constructing and building ME-Models.** Components (or parts) data detail the molecular content of the cell or system by accounting for all metabolites, proteins, RNA molecules, lipids, and the genomic DNA strands. Interactions data specify links between these molecular components. Functional-states data provide insight into cellular phenotype and come primarily in the form of gene and protein expression.

data must be reconciled as a ME-Model is built. For example, determining the transcription units (single RNA molecules) in *Thermotoga maritima* so that its ME-Model could be built was in and of itself a large undertaking (see Chapter 2).

ME-Models overcome some (but not all) of the major weaknesses of M-Models listed in the last section. These changes are summarized in Chapter 7. Before we dive in, let's step back and appreciate that the view of life through the lens of a ME-Model is fundamentally different than through the lens of an M-Model.

With an M-Model, I think its fair to say that a cell is viewed as a sac where "energy transactions through chemical transformations" take place. With a ME-Model, the *in silico* cell additionally carries out the central dogma of molecular biology. This gives us an additional view the cell, this time as a molecular information storage and processing device.

1.4 There's more than one way to model a cell, so where do ME-Models fit in?

The genotype-phenotype relationship is fundamental to biology. Finding general, underlying rules that govern the complex relationship between gene expression and cell growth, however, has proven a challenge. The genotype-phenotype relationship in microbes can be conceptualized as a five-layer hierarchical model (see Figure 1.3). A cell faces myriad constraints on its function at all layers [7, 8]. At the whole-cell level, it may be difficult to determine the constraints that govern cellular functions on a mechanistic basis, but they can be identified from empirical observation. Microbiologists pursued this approach in the 1950s and 1960s, resulting in empirical parameters such as the growth and non-growth maintenance coefficients [9] and yield coefficients that are widely used in the bioprocessing literature [10].

Terry Hwa and colleagues have been progressively expanding on the whole-cell empirical approach by means of an insightful combination of targeted experimentation and mathematical analysis [11, 12, 13, 14, 15, 16]. They predominantly use *Escherichia coli* cells grown under a variety of conditions. Systems microbiologists will always be closing the gap between the new biology uncovered by taking such approaches, and what is currently capable by taking the genome-scale approach. The advantage of taking a genome-scale approach is that you get predictions that are specific and detailed, which sometimes means they are more actionable when pursuing practical applications. We've incorporated some of the information from these pioneering studies to test and parameterize our ME-Models. Interestingly, we find that it may be possible to have the best of both

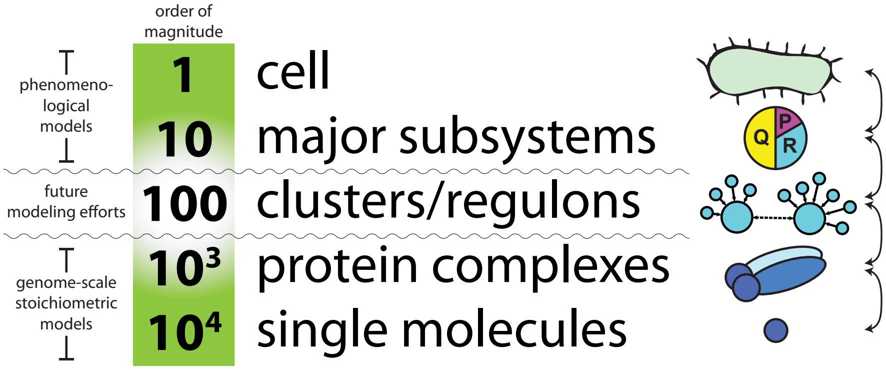


Figure 1.3: **The microbial genotype-phenotype relationship.** Bacterial cell growth and gene expression are linked through a hierarchy that extends from tens of thousands of molecules to a single cell. Each layer in the hierarchy imposes constraints on adjacent layers (arrows, right). At the top, empirical models can predict the relative levels of proteins belonging to major subsystems within a cell (e.g., metabolism (P), macromolecular synthesis (R)). At the bottom, genome-scale models can make predictions by accounting for all single molecules and protein complexes. A future modeling challenge is to characterize the functionality of the approximately 100 coordinately expressed clusters of protein complexes and to determine the evolutionary pressures leading to regulon formation (middle layer).

worlds! Phenomenological models can be embedded inside ME-Models to test their validity in the context of the more detailed description of the cell (more on that in Chapter 4).

Taken together, our combined efforts are leading to a multiscale understanding of the genotype-phenotype relationships underlying metabolism, gene expression, and growth in microbes. At all levels, model structures must be continually developed and re-worked in order to adequately capture new information and constraints allowing for optimization to approximate the cellular objective [17]. Cementing the levels shown in Figure 1.3 into a coherent multiscale framework is a challenge facing the field. The ME-Model is another major step toward meeting this challenge. Clearly, an exciting era is ahead of us, in which a combination of *in silico* and experimental approaches promises to continue the development of mechanistic and principled genotype-phenotype relationships that are akin to the development of fundamental physical laws a century ago. If successful, such development will move microbiology into a fundamentally new realm.

Chapter 1 is in part adapted from Lerman, J., and Palsson, B. O. (2010). Microbiology. Topping off a multiscale balancing act. Science 330, 1058-1059 (perspective). The dissertation author was the primary author of this perspective.

Chapter 2

The genome organization of *Thermotoga maritima* reflects its lifestyle

2.1 Abstract

The generation of genome-scale data is becoming more routine, yet the subsequent analysis of omics data remains a significant challenge. Here, an approach that integrates multiple omics datasets with bioinformatics tools was developed that produces a detailed annotation of several microbial genomic features. This methodology was used to characterize the genome of *Thermotoga maritima*-a phylogenetically deep-branching, hyperthermophilic bacterium. Experimental data were generated for whole-genome resequencing, transcription start site (TSS) determination, transcriptome profiling, and proteome profiling. These datasets, analyzed in combination with bioinformatics tools, served as a basis for the improvement of gene annotation, the elucidation of transcription units (TUs), the identification of putative non-coding RNAs (ncRNAs), and the determination of promoters and ribosome binding sites. This revealed many distinctive properties of the *T. maritima* genome organization relative to other bacteria. This genome has a high number of genes per TU (3.3), a paucity of putative ncRNAs (12), and few

TUs with multiple TSSs (3.7%). Quantitative analysis of promoters and ribosome binding sites showed increased sequence conservation relative to other bacteria. The $5'$UTRs follow an atypical bimodal length distribution comprised of ‘Short’ $5'$UTRs ($11 - 17$ nt) and ‘Common’ $5'$UTRs (26-32 nt). Transcriptional regulation is limited by a lack of intergenic space for the majority of TUs. Lastly, a high fraction of annotated genes are expressed independent of growth state and a linear correlation of mRNA/protein is observed (Pearson r = 0.63, $p < 2.2 \times 10^{-16}$ t-test). These distinctive properties are hypothesized to be a reflection of this organism’s hyperthermophilic lifestyle and could yield novel insights into the evolutionary trajectory of microbial life on earth.

2.2 Author Summary

Genomic studies have greatly benefited from the advent of high-throughput technologies and bioinformatics tools. Here, a methodology integrating genome-scale data and bioinformatics tools is developed to characterize the genome organization of the hyperthermophilic, phylogenetically deep-branching bacterium *Thermotoga maritima*. This approach elucidates several features of the genome organization and enables comparative analysis of these features across diverse taxa. Our results suggest that the genome of *T. maritima* is reflective of its hyperthermophilic lifestyle. Ultimately, constraints imposed on the genome have negative impacts on regulatory complexity and phenotypic diversity. Investigating the genome organization of Thermotogae species will help resolve various causal factors contributing to the genome organization such as phylogeny and environment. Applying a similar analysis of the genome organization to numerous taxa will likely provide insights into microbial evolution.

2.3 Introduction

A fundamental step towards obtaining a systems-level understanding of organisms is to obtain an accurate inventory of cellular components and their interconnectivities [18, 19, 20]. The genome sequence and *in silico* predictions of gene annotation are the starting points for assembling a network. For prokaryotes, these *in silico* approaches detect open reading frames and structural RNAs with varying degrees of accuracy [21]. Recently, multi-omic data generation and analysis studies [22, 23, 24, 25, 26, 27, 28] have revealed an abundance of genomic features that are not detected computationally such as transcription start sites (TSSs), promoters, untranslated regions (UTRs), non-coding RNAs, ribosome binding sites (RBSs) and transcription termination sites [29]. However, the rate at which multi-omic datasets are being generated is substantially outpacing the development of analysis workflows for these inherently dissimilar data types [30]. Here, multi-omic experimental data is generated and analyzed in conjunction with bioinformatics tools to annotate numerous bacterial genomic features that cannot accurately be detected using *in silico* approaches alone. This methodology was applied to study the genome organization of *Thermotoga maritima*-a phylogenetically deep-branching, hyperthermophilic bacterium with a compact 1.86 Mb genome.

Originally isolated from geothermally heated marine sediment, *T. maritima* grows between $60 - 90$°C with an optimal growth temperature of 80°C [31]. This species belongs to the order Thermotogales that have, until recently, been exclusively comprised of thermophilic or hyperthermophilic organisms. Compared to most bacteria, Thermotogales are capable of sustaining growth over a remarkably wide range of temperatures. For instance, *Kosmotoga olearia* can be cultivated between $20 - 80$°C [32]. Recently, the existence of mesophilic Thermotogales [33, 34] was confirmed with the description of *Mesotoga prima*, which grows from $20 - 50$°C with an optimum at 37 °C [35]. Sequencing of *M. prima* revealed that it has the largest genome of all the Thermotogales at 2.97 Mb with $\approx$15% noncoding DNA [36]. *T. maritima*, which grows at the upper-limit known for Thermotogales, has one of the smallest genomes in this order and maintains one of the most compact genomes among all sequenced bacterial species ($<$5% noncoding DNA) [37, 38]. The short intergenic regions in the *T. maritima* genome (5 bp median) resemble those in the genome of *Pelagibacter ubique*, a bacterium that has undergone genome streamlining and has the shortest median intergenic space (3 bp) among free-living bacteria [37]. Although it remains unclear whether *T. maritima* has also undergone streamlining, both organisms encode only a few global regulators (four sigma factors in *T. maritima* versus two in *P. ubique*) and carry just a single rRNA operon. In contrast with *P. ubique*, *T. maritima* displays more metabolic diversity through its ability to ferment numerous mono- and polysaccharides [31, 39].

Thermotogales have been the focus of many evolutionary studies [40, 33, 41]. Organisms in hydrothermal vent communities, where many Thermotogales have been isolated, are thought to harbor traits of early microorganisms [42]. Phylogenetic analysis of 16S rRNA sequences place the Thermotogae at the base of the bacterial phylogenetic tree [43, 44]; however, Zhaxybayeva et al. [41] determined through analysis of 16S rRNA and ribosomal protein genes that Thermotogae and Aquificales (a hyperthermophilic order) are sister taxa. The authors also determined that the majority of Thermotogae proteins align best with those found in the order Firmicutes [41]; therefore, the exact phylogenetic position of Thermotogae is still unresolved. Nevertheless, members of this phylum are among the deepest branching bacterial species and, as such, prime candidates for evolutionary studies.

Thermophiles such as *T. maritima* implement numerous strategies at both the protein and nucleic acid levels to support growth at high temperatures. For instance, intrinsic protein stabilization is achieved by utilizing more charged residues at the protein surface, encoding for a dense hydrophobic core, and increasing disulfide bond usage [45, 46]. DNA is typically kept from denaturing by introducing positive supercoils via reverse gyrase activity while phosphodiester bond degradation is prevented by stabilization through interaction with cations (e.g. K^+, Mg^{2+}) and polyamines [47, 48]. However, the impact of temperature on genome features essential to gene expression such as promoters and RBSs remains largely unexplored. Bacterial transcription initiation is governed by recognition of promoter sequences by sigma factors, which load the RNA polymerase holoenzyme upstream of the transcription start site (TSS). Translation initiation is predominantly reliant on base pairing between the anti-Shine-Dalgarno sequence found near the $3'$-terminus of the 16S rRNA and the Shine-Dalgarno sequence (i.e. the

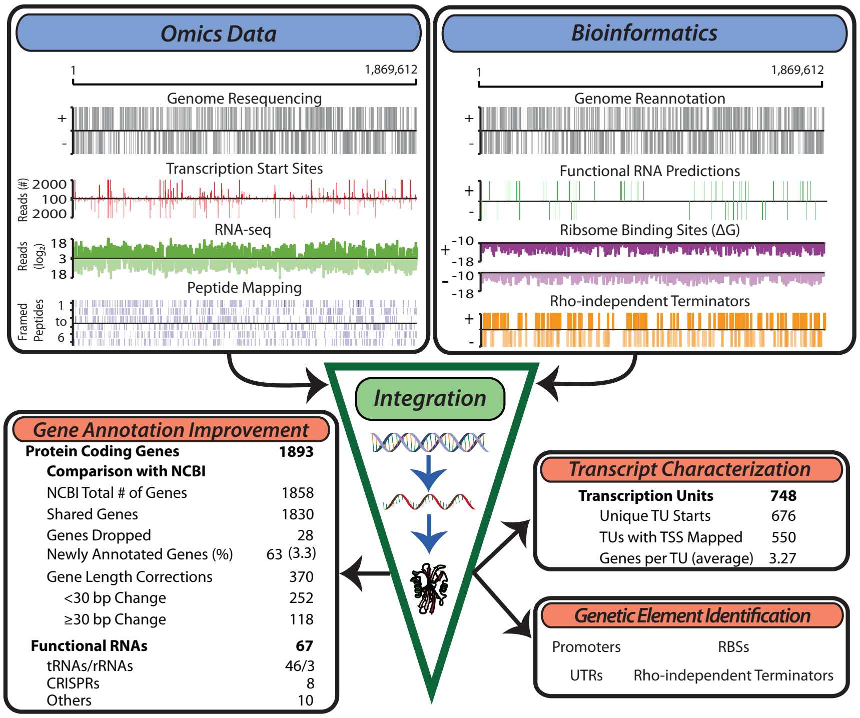
RBS). Therefore, thermophilic macromolecular synthesis machinery must establish and retain contacts with nucleic acids while facing greater thermodynamic challenges.

The integrated approach described here enables an experimentally anchored annotation of several bacterial genomic features including protein-coding genes, functional RNAs, non-coding RNAs, transcription units (TUs), promoters, ribosome binding sites (RBSs) and regulatory sites such as transcription factor (TF) binding sites, $5'$ and $3'$ untranslated regions (UTRs) and intergenic regions. This is achieved through the simultaneous analysis of genomic, transcriptomic and proteomic experimental datasets with complementary bioinformatics approaches. In addition to providing a valuable resource to the research community, this analysis framework facilitates quantitative and comparative analysis of annotated features across microbial species. For the genome of *T. maritima*, several distinguishing characteristics were identified and their potential causal factors are discussed.

2.4 Results

2.4.1 An integrative, multi-omic approach for the annotation of the genome organization

An integrative workflow was developed to re-annotate the genome of *T. maritima*. The re-annotated genome is the result of the simultaneous reconciliation of multiple omics data sources (Figure 2.1, upper left) with bioinformatics approaches (Figure 2.1, upper right). Omics data generated included: (1) genome resequencing, (2) transcription start site (TSS) identification using a modified 5'RACE (Rapid Amplification of cDNA Ends) protocol, (3) transcriptome profiling using both high-density tiling arrays and strand-specific RNA-seq, and (4) LC-MS/MS shotgun proteomics. Transcriptome data were generated from cultures grown in diverse conditions including log phase growth, late exponential phase, heat shock, and growth inhibition by hydrogen (See Materials and Methods). Proteomic datasets include log phase growth and late exponential phase



Gene Annotation Improvement	
Protein Coding Genes	**1893**
Comparison with NCBI	
NCBI Total # of Genes	1858
Shared Genes	1830
Genes Dropped	28
Newly Annotated Genes (%)	63 (3.3)
Gene Length Corrections	370
<30 bp Change	252
≥30 bp Change	118
Functional RNAs	**67**
tRNAs/rRNAs	46/3
CRISPRs	8
Others	10

Transcript Characterization	
Transcription Units	**748**
Unique TU Starts	676
TUs with TSS Mapped	550
Genes per TU (average)	3.27

Genetic Element Identification

Promoters RBSs

UTRs Rho-independent Terminators

Figure 2.1: **Generation of multiple genome-scale datasets integrated with bioinformatics predictions reveals the genome organization.** Experimental data generated for the study of the *T. maritima* genome include genome resequencing, TSS determination, RNA-seq, tiling arrays (not shown) and LC-MS/MS peptide mapping (top left). Bioinformatics approaches used include genome re-annotation, functional RNA prediction, ribosome binding site energy calculations, and determination of intrinsic terminators (top right). Integration of these distinct datasets involves normalization and quantification to genomic coordinates. This experimentally anchors gene annotation improvements, defines the TU architecture, identifies non-coding RNAs and serves as a basis for the identification of additional genetic elements such as promoters and ribosome binding sites.

growth conditions. In combination with various bioinformatics approaches, integration of these omics datasets allowed for the definition of gene and transcription units (TU) boundaries with single base-pair resolution. The updated and expanded annotation served as the basis for genome-wide identification of promoters, ribosome binding sites (RBSs), intrinsic transcriptional terminators and UTRs.

Annotation of open reading frames (ORFs)

Reannotation of the *T. maritima* MSB8 genome began with whole genome resequencing of the ATCC derived strain. Genome resequencing was prompted by the recent identification of a ≈9 kb chromosomal region in the DSMZ derived strain (DSMZ genomovar, Genbank Accession AGIJ00000000.1) that is not present in the original genome sequence derived from a TIGR strain (TIGR genomovar, Genbank Accession AE000512.1) [49]. Resequencing the ATCC derived strain (presented as the ATCC genomovar, Genbank Accession CP004077) ensured that subsequent analyses referenced an accurate genome sequence. The ATCC genomovar sequence consists of 1,869,612 bp and, like the DSMZ genomovar, carries an ≈9 kb chromosomal region found between TM1847 and TM1848 of the TIGR annotation. The draft genome was annotated using the RAST Pipeline [50] and was then reconciled with the existing TIGR genomovar annotation. The RAST draft annotation had 1,887 protein-coding sequences while the TIGR annotation contained 1,858. Comparison of these two annotations with transcriptome, proteome and bioinformatics datasets resulted in a final annotation containing 1,893 protein-coding sequences (Table S1 in [51]). The final gene annotation retained a total of 1,830 NCBI annotated genes while 28 NCBI annotated genes were dropped (or replaced) due to a lack of experimental support. An additional 63 genes were annotated based on evidence found in multiple data-types. Furthermore, 370 genes varied in length when comparing the final gene annotation to the NCBI annotation. These discrepancies in gene length were predominantly due to differences in the start codon assignment, thus changing the amino acid sequence at the N-terminus. Gene length annotation differences of less than 10 amino acids were not resolved using the generated datasets without the presence of direct proteomic evidence to support one annotation over the other. However, 118 of these 370 genes (32%) had large discrepancies in their gene length annotation, equaling or exceeding 10 amino acids. For these cases, annotation conflicts were resolved using data from peptide mapping, transcript presence and bioinformatics tools.

Annotation of transcription units (TUs)

In addition to the annotation of ORFs, the genome annotation was expanded to include the TU architecture. The TU architecture is defined here to be the genomic coordinates of all RNA molecules in the transcriptome. To expand the annotation to include TUs, transcript bounds were resolved with single base pair resolution using data from RNA-seq and TSS determination. Definition of these bounds was facilitated by bioinformatics approaches; for example, the prediction of intrinsic transcriptional terminators was used to aid in assigning $3'$ bounds of transcripts. This approach resulted in the assignment of 748 TUs with a total of 676 unique TSSs (Table S2 in [51]). The majority of TUs were found to be polycistronic (427, 57%) while the rest of the TUs contain only a single gene (321, 43%). The average TU contains 3.3 genes which is greater than the typical 1-2 genes per transcript observed in other bacteria [24, 52, 53] but similar to those found in archaea [26, 54]. Previous high-resolution studies of microbial transcriptomes have identified the transcription of suboperonic regions as a source of transcriptional complexity [22, 25, 52]. In *T. maritima* 165 TUs (22%) are suboperonic, having their initiation site within a longer TU. This fraction of suboperons observed in *T. maritima* is within the range observed in other bacteria; however, some organisms such as *Helicobacter pylori* have similarly sized genomes (1.67 Mb) but use suboperonic transcription much more frequently (47%, excluding antisense suboperons) [25]. Another source of transcriptional complexity comes from the use of multiple start sites, however, only a small number of *T. maritima* TUs (28, Table S3 in [51]) were observed to utilize them.

Annotation of non-coding RNAs

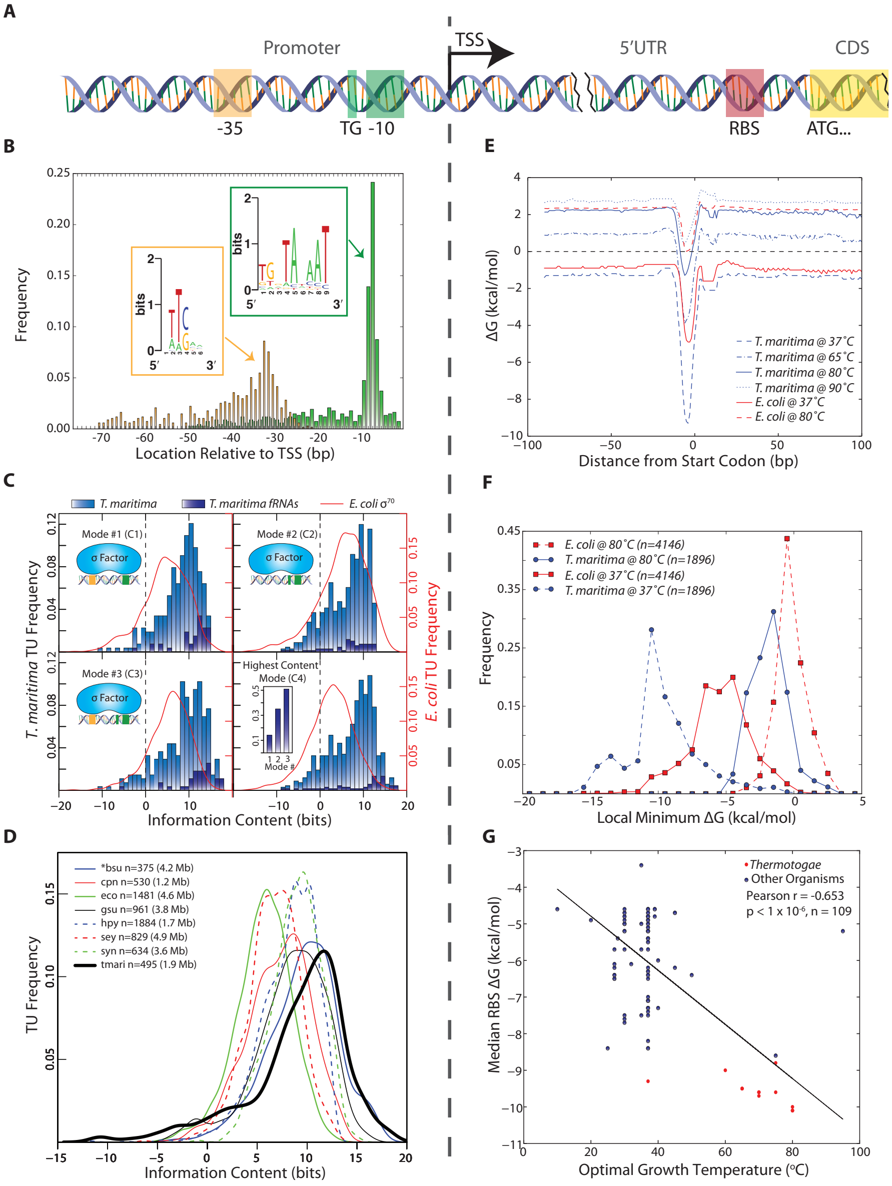
Beyond facilitating protein-coding gene annotation, transcriptome data provided experimental evidence supporting the bioinformatics prediction of 46 tRNAs, 3 rRNAs, 8 CRISPR cassettes and an additional 10 non-coding RNAs which include riboswitches, leader sequences, RNase P RNA, tmRNA and SRP RNA. These features are included in the final annotation presented here (CP004077, Table S1 in [51]). Transcription was detected antisense to 19% of annotated genes (Table

S4 in [51]). However, 3′UTRs account for 52% of these antisense transcripts and only 62 antisense transcripts have an experimentally identified TSS. Furthermore, the median log phase FPKM (Fragments Per Kilobase of transcript per Million mapped reads) values are much lower for antisense transcripts (4.5) than those found for protein-coding genes (117). Transcriptome data also enabled identification of 13 putative non-coding RNAs (ncRNAs, Table S5 in [51]). No secondary structures could be defined for these putative ncRNAs using the prediction algorithms RNAfold [55] and Infernal [56] at 80°C. Four of these putative ncRNAs contain small ORFs (<40 amino acids) but no peptide evidence for these small ORFs was found in the proteomic datasets.

2.4.2 Identification of promoters and RBSs followed by quantitative intra- and interspecies analysis of binding free energies

The genome-wide identification of promoter and RBS sites was facilitated by the annotated TU start loci and protein start codons (Figure 2.2A). Promoter and RBS sequences were then quantitatively analyzed using thermodynamic principles. These same quantitative measures were applied to numerous organisms for interspecies comparison.

Figure 2.2: **Identification and quantitative comparison of genetic elements for transcription and translation initiation.** (A) Schematic showing the position of the promoter upstream of the TSS and the RBS upstream of the translation start codon. (B) The genomic position of the 3′ end of each promoter element is shown relative to the TSS for all *T. maritima* TUs. Promoter elements were identified using a gapped motif search for a -35 hexamer and a -10 nonamer. This revealed an *E. coli* σ70 promoter architecture for the housekeeping sigma factor of *T. maritima*, RpoD. The motif for both promoter elements is displayed as a sequence logo (insets). (C) The relative binding free energy of σ70 is captured using information content. Each panel shows the distribution of promoter information content for *T. maritima* and *E. coli.* Mode 1 (C1) calculates information content based on σ70 contacts with the -35 and -10 hexamer promoter elements $n_{tmari} = 265$, $n_{tmari\_fRNA} = 38$, $n_{eco} = 650$). Mode 2 (C2) represents binding to the extended -10 promoter ($n_{tmari} = 676$, $n_{tmari\_fRNA} = 57$, $n_{eco} = 1{,}481$). Mode 3 (C3) represents σ70-binding to both the -35 and the extended -10 promoter elements ($n_{tmari} = 274$, $n_{tmari\_fRNA} = 37$, $n_{eco} = 657$). (C4) shows the distribution of information content for all promoters when only the highest scoring mode is considered ($n_{tmari} = 676$, $n_{tmari\_fRNA} = 57$, $n_{eco} = 1{,}481$). The inset shows the highest distribution of functional RNAs across the modes. (D) The σ70 binding modes from (C) were used to calculate the promoter information content for seven additional bacterial species. Analogous to (C4), the distribution of information scores when only the highest bit score mode is considered is shown. The organism abbreviations correspond to the following: bsu, *Bacillus subtilis*; cpn, *Chlamydophila pneumoniae* CWL029; eco, *Escherichia coli* K12 MG1655; gsu, *Geobacter sulfurreducens* PCA; hpy, *Helicobacter pylori* 26695; sey, *Salmonella enterica* subsp. enterica serovar Typhimurium SL1344; syn, *Synechocystis* sp. PCC 6803; tmari, *T. maritima* MSB8. The genome size is given in paranthesis. *bsu data is extracted from a highly curated source that is a collection of small-scale experiments and, as such, this distribution is not a genome-scale assessment of promoter strength. (E) The calculated median RBS ΔG for all genes based on the position relative to the start codon. Temperature profiles are shown for *T. maritima* at 37°C (for comparison), 65°C (lower growth limit), 80°C (growth optimum) and 90°C (upper growth limit). Similar profiles are shown for *E. coli* at 37°C (optimal) and 80°C (for comparison). (F) The local minimum RBS ΔG for all genes in a 30 nt window upstream of the annotated start codon generated for *T. maritima* and *E. coli* at 37°C and 80°C. (G) Similar to (F), the median of the local minimum RBS ΔG was calculated and plotted for 109 bacteria against their optimal growth temperature. Species in the Thermotogae phylum (n = 15) are shown in red.


A
Promoter
TSS
5′UTR
CDS
-35
TG -10
RBS
ATG...
B
Frequency
Location Relative to TSS (bp)
bits
5′
3′
E
ΔG (kcal/mol)
Distance from Start Codon (bp)
T. maritima @ 37˚C
T. maritima @ 65˚C
T. maritima @ 80˚C
T. maritima @ 90˚C
E. coli @ 37˚C
E. coli @ 80˚C
C
T. maritima
T. maritima fRNAs
E. coli σ70
Mode #1 (C1)
Mode #2 (C2)
Mode #3 (C3)
Highest Content Mode (C4)
σ Factor
Mode #
T. maritima TU Frequency
E. coli TU Frequency
Information Content (bits)
F
E. coli @ 80˚C (n=4146)
T. maritima @ 80˚C (n=1896)
E. coli @ 37˚C (n=4146)
T. maritima @ 37˚C (n=1896)
Frequency
Local Minimum ΔG (kcal/mol)
D
*bsu n=375 (4.2 Mb)
cpn n=530 (1.2 Mb)
eco n=1481 (4.6 Mb)
gsu n=961 (3.8 Mb)
hpy n=1884 (1.7 Mb)
sey n=829 (4.9 Mb)
syn n=634 (3.6 Mb)
tmari n=495 (1.9 Mb)
TU Frequency
Information Content (bits)
G
Thermotogae
Other Organisms
Pearson r = -0.653
p < 1 x 10-6, n = 109
Median RBS ΔG (kcal/mol)
Optimal Growth Temperature (ºC)

Annotation-guided search for motifs reveals promoter structures that enable many contacts with RNA polymerase holoenzyme

Bacterial RNA polymerase is recruited predominantly through the binding of sigma factors to promoter regions. A promoter motif search was performed upstream of all unique *T. maritima* TU start sites. This revealed a strongly conserved, *E. coli* σ70-like consensus sequence for the housekeeping sigma factor RpoD (Tmari_1457). No motifs were detected for the alternate sigma factors RpoE, SigH and FliA (See Materials and Methods). The RpoD motif has three distinct promoter elements: a -10 hexamer, a -35 hexamer and a $5'$TGn element directly upstream of the -10 hexamer (Figure 2.2B). Individual promoters identified carried combinations of these three elements. The distance between the TSS and the $3'$ end of the -10 element was found to be 7 bp (Figure 2.2B). This is in strong agreement with the expected spacing for the consensus sequence of *E. coli* σ70. The same is true of the -35 element though the location of the -35 hexamer is more variable compared with the -10 hexamer partly due to the variability of the spacing between the -10 and -35 promoter elements. Plotting the spacer between the -10 and -35 promoter elements yields a distribution centered around 17 bp, which also is in agreement with the *E. coli* σ70 consensus (Figure S1 in [51]). Furthermore, plotting of genomic AT content upstream and downstream of aligned -10 promoter elements reveals an increase in AT content $\approx$75 bp upstream of the -10 promoter element (Figure S2 in [51]). This suggests the presence of UP elements for a subset of *T. maritima* promoters. The α-subunits of RNA polymerase bind to UP elements, facilitating initiation of transcription [57, 58].

Quantitative assessment of *T. maritima* promoters indicates high information content across multiple σ70 binding modes

The identification of σ70 promoter elements enabled the quantitative study of the relative binding free energy associated with individual promoters. The sequence conservation of an individual promoter element (i.e. the information content measured in bits [59]) can be computed through application of molecular information theory and is achieved through quantitative comparison of a given sequence to the average sequence conservation across the genome as measured through the position weight matrix [60] (See Materials and Methods). Information content has been correlated to binding free energy (ΔG) through the second law of thermodynamics [61, 62, 63], where sequences with high information content are closer to consensus and, therefore, have stronger relative binding free energy (more negative ΔG). Experimental results, both *in vitro* and *in vivo*, have shown that information content is moderately predictive of promoter strength and activity [64].

The information content for individual *T. maritima* promoters was computed using a model of σ70 promoters that accounts for the information content of each promoter element and the variation in spacing between the -10 and -35 elements [63]. Using this approach, the information content of each *T. maritima* promoter was determined for three, σ70-binding modes that represent the potential contacts between σ70 and the promoter elements (Figure $2.2C1 - C3$). Plotting the maximum information carrying binding mode for all promoters (Figure 2.2C4) shows that the vast majority of promoters (90%) have information content greater than zero. This indicates that, for these TUs, σ70 binding and transcription initiation is thermodynamically favorable (ΔG<0). Furthermore, the distribution of information content indicates that the median *T. maritima* promoter has 8.7 bits compared to *E. coli* σ70 promoters whose median is 5.9 bits. Comparison of *T. maritima* promoters across all modes shows that the extended -10 promoter (-10 hexamer and upstream $5'$TGn, Mode 2) provides the highest information for most TUs (63%). Furthermore, an extended -10 promoter combined with a -35 box (Mode 3) yields the highest information content in 25% of all promoters and 51% of functional RNA promoters (Figure 2.2C4 inset). These RNAs, which are among the most actively transcribed genes, encode promoters with exceptionally high information content (median 12.1 bits).

Interspecies comparative analysis reveals that *T. maritima* promoters have high relative sequence conservation

The surprisingly high sequence conservation of *T. maritima* promoters prompted a comparative analysis of information content across multiple bacterial species. The scope of the comparative analysis was limited by the lack of datasets detailing bacterial TSS locations and the association of those TSSs with $\sigma70$. Publically available datasets for only seven additional, diverse microorganisms met this criteria. The organisms included in the analysis are the Gammaproteobacteria *Escherichia coli* K12 MG1655 [65] and *Salmonella enterica* subsp. enterica serovar Typhimurium SL1344 [66], the Deltaproteobacterium *Geobacter sulfurreducens* PCA [24], the Epsilonproteobacterium *Helicobacter pylori* 26695 [25], the Chlamydiae *Chlamydophila pneumoniae* CWL029 [67], the Cyanobacterium *Synechocystis* sp. PCC 6803 [68] and the Firmicute *Bacillus subtilis* [69]. Since these datasets contain only experimentally confirmed TSS loci, only *T. maritima* TUs with an experimentally confirmed TSS were included in this interspecies comparison (495 TUs out of 676). As before, the information content across all three $\sigma70$-binding modes was calculated. The distribution of the highest information content mode (Figure 2.2D) indicates that *T. maritima* promoters are the strongest among all organisms studied, carrying a median of 10.2 bits of information. Thus, among bacteria, *T. maritima* promoter information content associated with $\sigma70$-binding is relatively high.

Analysis of *T. maritima* RBS binding strength reveals strong binding free energies that support translation initiation at 80 °C

The RNA/RNA binding free energy of the Shine-Dalgarno with the anti-Shine-Dalgarno was calculated in a temperature-dependent manner using the gene annotation as a reference point. Across all protein coding genes, the median RBS ΔG was calculated ±100 nucleotides (nt) from the start codon at temperatures ranging from 37 °C to 90 °C (Figure 2.2E). The position of the lowest ΔG is shown to be $4-6$ nt upstream of the start codon, which is in agreement with the optimal RBS location for translation initiation [70]. *T. maritima* is shown to maintain a thermodynamically favorable median ΔG up to its growth temperature maximum of 90 °C [31]. Plotting the distribution of local minimum ΔG's at 80 °C (Figure 2.2F) reveals that 93% of *T. maritima* protein-coding genes have a RBS with ΔG<0. Calculating RBS free energy distributions at different temperatures (Figure 2.2F) reveals that at higher temperatures there is a narrowing in the range of observed free energies. *T. maritima* RBSs have a median absolute deviation of 1.30 kcal/mol at 37 °C compared to 0.87 kcal/mol at 80 °C ($p = 4.4 \times 10^{-33}$, Wilcoxon rank-sum test). Comparison of *E. coli* and *T. maritima* RBSs reveals that *T. maritima* RBSs are substantially weaker at their respective optimal growth temperatures (Figure 2.2F). A large fraction (36%) of *E. coli* genes have a ΔG>0 at 80 °C and would not be capable of supporting hyperthermophilic life. When compared at equal temperatures (Figure 2.2F, 80 °C) *T. maritima* RBSs are stronger.

Interspecies analysis indicates that RBS binding strength is influenced by both optimal growth temperature and phylogeny

To more rigorously test for a relationship between RBS strength and optimal growth temperature, RBS ΔG's were calculated for all genes in 108 additional bacterial species spanning numerous phyla (including 14 members of the Thermotogae phylum). These organisms include psychrophilic, mesophilic, thermophilic and hyperthermophilic microorganisms. A significant linear correlation was found between optimal growth temperature and median RBS ΔG (Pearson r = -0.653, $p < 1 \times 10^{-6}$ random permutation test), where increasing optimal growth temperatures trend with a lower median RBS ΔG calculated at 37 °C (Figure 2.2G). However, the energetic analysis of RBSs applied here is based on the 16S rRNA sequence of the anti-Shine-Dalgarno and, as such, phylogeny is a potential contributing factor to this correlation. To test this, three distance matrices were constructed: (1) for local minimum median RBS ΔG (across all genes in a given genome), (2) for optimal growth temperatures, and (3) for phylogenetic distances determined from 16S rRNA sequences. The Mantel test was then applied to evaluate the correlations among the pairwise distance matrices (Figure S3 in [51]) allowing for the contribution of optimal growth temperature to be decoupled from phylogeny with respect to RBS strength. This test indicated that both phylogeny and optimal growth temperature impact median RBS strength, with temperature slightly more significant than phylogeny (Mantel Statistic r = 0.37 vs 0.35, $p = 1 \times 10^{-4}$ random permutation test).

2.4.3 *T. maritima* promoter-containing intergenic regions reveal a unique distribution of $5'$UTRs and spatial limitations on regulation

Regulation in *T. maritima* was studied from the vantage point of an organism with extremely short intergenic regions. In both microbes [71] and higher organisms [72] it was shown that the regulatory complexity of an operon positively correlates with the amount of intergenic space found upstream of that operon. Promoter-containing intergenic regions (PIRs) served as well-defined genomic regions for this analysis (Figure 2.3A). PIRs contain target sites for transcriptional regulation (e.g. promoters and TF binding sites) as well as translational regulation (e.g. RBSs). Each PIR can be divided into two components in relation to the TSS: the sequence downstream of the TSS (the $5'$UTR) and the sequence upstream of the TSS.

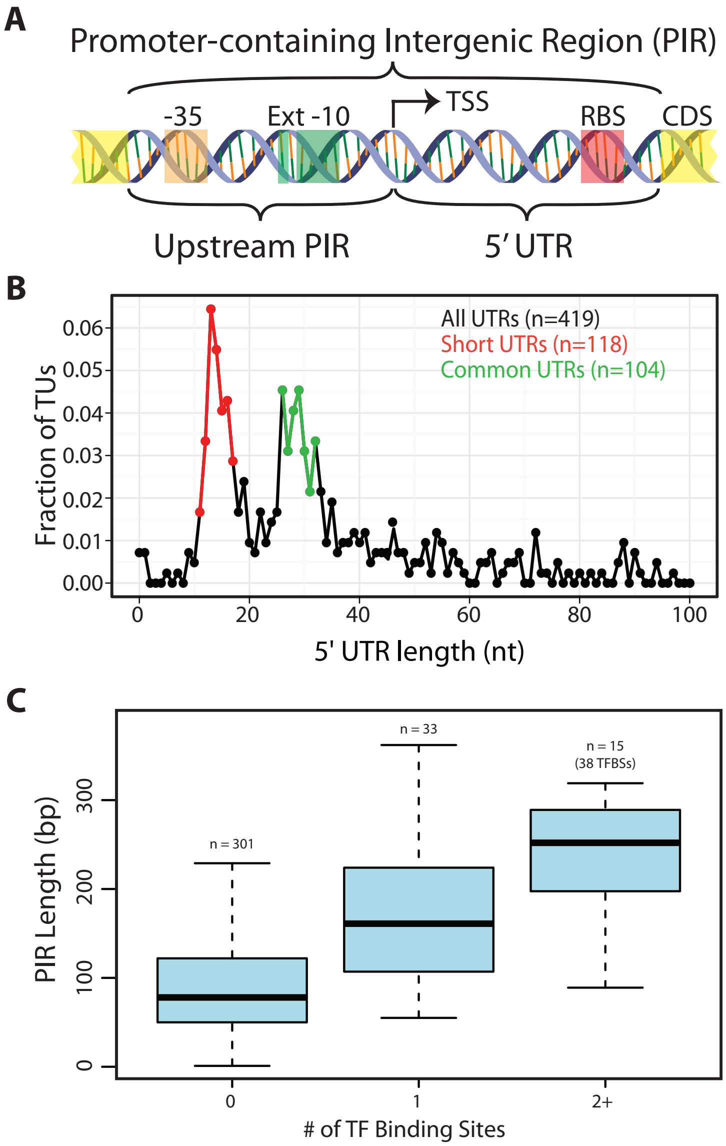


Figure 2.3: Arrangement of genomic features contained within promoter-containing intergenic regions (PIRs). (A) Schematic of the two subdivisions of the PIR and the genetic elements they typically carry. (B) The 5′UTR distribution is shown for all TUs with an experimentally identified TSS. The Short 5′UTR group (11 – 17 nt) is shown in red. The Common 5′UTR group (26 – 32 nt) is shown in green. Transcripts with an annotated functional RNA as the first feature were omitted from the analysis. Though only the first 100 nt are plotted, frequencies are based on the entire set of 5′UTR lengths. (C) A quartile plot of the length distribution of PIRs is shown. PIRs are grouped according to the number of TF binding sites they contain (no TF, a single TF or multiple TFs).

T. maritima **has a bimodal distribution of $5'$UTRs comprised of uncharacteristically 'Short' $5'$UTRs and 'Common' $5'$UTRs**

T. maritima exhibits an unusual bimodal distribution with respect to the length of $5'$UTRs (Figure 2.3B). To date, the $5'$UTRs of all other microorganisms follow a unimodal distribution centered at approximately 30 nt [24, 25, 52, 53]. Though *T. maritima* has a distinct peak (local maxima) from 26-32 nt (Common $5'$UTR Group), it has a second peak containing shorter $5'$UTRs with lengths between $11 - 17$ nts (Short $5'$UTR Group). Interestingly, there is underrepresentation of $5'$UTRs with lengths between $18 - 25$ nt. Leaderless transcripts were not detected in *T. maritima*, echoing the RNA/RNA binding energy analysis that indicated exclusive use of RBSs for translation initiation.

To better understand the bimodal nature of the $5'$UTR distribution, various factors were tested that could differentiate the Short $5'$UTR Group from the Common $5'$UTR Group and provide insights into the lack of $5'$UTRs between $18 - 25$ nt. Factors tested for over- or underrepresentation of the different $5'$UTR groups included: (1) gene expression level (both mRNA and protein levels), (2) protein expression normalized to mRNA expression, (3) phylogenetic origin of genes, (4) RBS and promoter strengths, (5) divergent vs. convergent operons, and (6) cellular functional categorization. These factors yielded no discrimination between the Short $5'$UTR Group and the Common $5'$UTR Group and could not explain the bimodal nature of the $5'$UTR length distribution.

T. maritima **PIRs are predominantly too short to permit transcription factor regulation**

To enable regulation of transcription, space in the genome must be dedicated to operator sites, which serve as docking locations for TF recruitment. Typically, these sites reside upstream of the TSS, but can also be found downstream of the TSS (in the $5'$UTR). An analysis centered on PIRs was chosen to capture the potential for TF binding sites both upstream and downstream of the TSS. A total of 31 TF regulons with a combined total of 91 genomic binding sites were extracted from the RegPrecise database [73]. Mapping of the TF binding sites to the

T. maritima genome showed that 71 were within PIRs, 12 mapped to intergenic regions not carrying a promoter and the remaining 8 were within or overlapped an annotated gene (Table S6 in [51]). The length distribution of PIRs without a TF binding site was compared to that of PIRs with TF binding sites (Figure 2.3C). The median length of PIRs that do not contain a TF binding site is 78 bp. This is significantly shorter than the length of PIRs that carry a single TF binding site (median = 161 bp, Wilcoxon rank-sum test $p = 6.9 \times 10^{-8}$) or multiple TF binding sites (median = 252 bp, Wilcoxon rank-sum test $p = 2.8 \times 10^{-7}$). Thus, the majority of *T. maritima* PIRs do not contain the typical space required to encode a TF binding site.

2.4.4 *T. maritima* has an actively transcribed genome that is tightly correlated to protein abundances

Transcriptome data indicate that the genome of *T. maritima* is exceptionally active irrespective of growth condition (Figure 2.4A) with 91-96% of genes expressed above an FPKM threshold of 8. This fraction of genes transcribed is uncharacteristically high compared to other free-living bacteria (see Table S7 in [51]). Furthermore, translational evidence supporting the high gene expression activity of *T. maritima* is found in the proteomic datasets. In each condition tested, peptide evidence was detected for 74% of the annotated proteins. It is also found that mRNA and protein abundances are tightly linked (Pearson r = 0.63, $p < 2.2 \times 10^{-16}$ t-test) (Figure 2.4B). This correlation is stronger and more significant than those reported in comparable studies for other bacteria [74, 75].

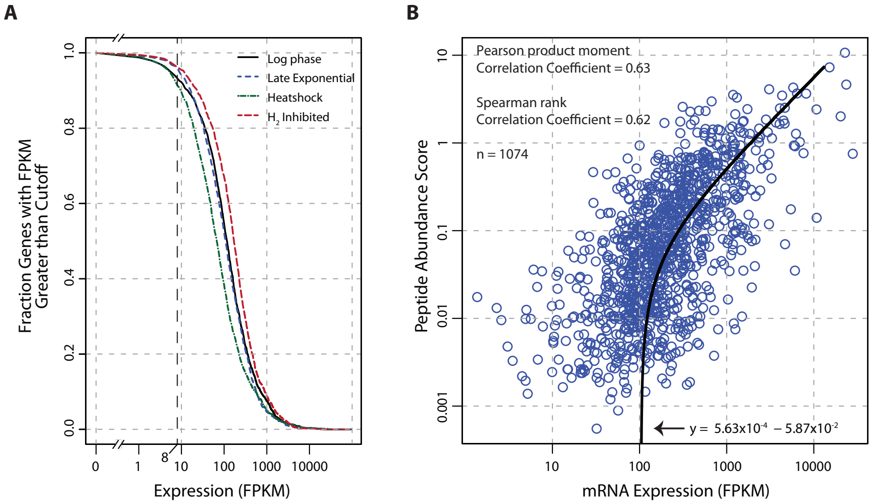


Figure 2.4: **Global analysis of mRNA and protein expression levels.** (A) The fraction of transcribed genes as a function of the FPKM threshold. Under growth promoting conditions (log-phase) and early in the transition to stressed conditions (carbon-limited late exponential phase, heat shock, and hydrogen inhibition), 91-96% of the genome is expressed using a conservative FPKM threshold of $\geq$8. (B) Correlation of mRNA expression and protein abundance. The line of best fit indicates a strong linear relationship (Pearson r = 0.63, $p < 2.2 \times 10^{-16}$ t-test) between transcription and translation. The peptide abundance score for each protein was derived by dividing the total spectral count by the number of possible tryptic peptides (400-2000 m/z up to a charge state (z) of 3, hence a maximum fragment mass of 6000). Abbreviations: FPKM, Fragments Per Kilobase of transcript per Million mapped reads; m/z, mass-to-charge ratio.

2.5 Discussion

Genome-scale technologies have provided researchers unprecedented access to large volumes of data detailing the composition of a cell. However, approaches for data analysis and interpretation have lagged behind due to the scope and complexity of these data types. Here, we present a framework for multi-omic data analysis that annotates genomic features involved in transcription, translation and regulation. This methodology integrates genome-scale datasets with bioinformatics predictions to produce 1) an improvement of the gene annotation, 2) an experimentally validated TU architecture and 3) the identification of putative antisense, non-coding transcripts and alternative TSSs. Using these annotated genomic features enabled the genome-wide identification of promoters and RBSs, which are difficult to identify solely using *in silico* approaches [76, 77]. Furthermore, the relative binding strength of individual promoters and RBSs was quantitatively measured using thermodynamic principles enabling multi-species comparison of these sequence features. The annotated genome organization served as a scaffold for analyzing regulatory features. Transcription factor regulation was examined with respect to promoter containing intergenic regions while the translational impact of the 5′UTR distribution was considered. The multi-omic data generation and analysis demonstrated here is applicable to many microbial species.

Applying this methodology to study the genome organization of *T. maritima* revealed that it has many distinctive properties compared to other organisms. Genome-scale analysis of promoters showed that *T. maritima* encodes a highly conserved, robust architecture that ensures transcription initiation. Similarly, RBS sequence conservation was shown to be thermodynamically sufficient for translation initiation for almost all *T. maritima* genes at 80°C compared with only a fraction of *E. coli* genes. The distinctive properties of the *T. maritima* genome extend beyond sequence composition and are apparent at the organizational level. The high protein-coding density and minimal intergenic space found in this organism have resulted in a high number of genes per TU, a paucity of putative ncRNAs and few TUs with multiple start sites. Furthermore, transcriptional regulation appears to be limited to a few TUs due to a lack of genomic space in PIRs. Interestingly, the $5'$UTR component of the PIR was found to be uncharacteristically bimodal and was comprised of an unusually short grouping of $5'$UTRs. Lastly, the constrained genome organization of *T. maritima* is reflected in the physiological state of the cell. Transcription of the vast majority of genes is detected independent of culture condition and the correlation between protein and mRNA is stronger than previously observed in other bacteria.

We hypothesize that the hyperthermophilic lifestyle of *T. maritima* could potentially explain the distinctive characteristics of this organism's genome organization. For instance, the increased sequence conservation of promoter elements and RBSs throughout the *T. maritima* genome may be attributed to the need to maintain gene expression under extreme temperature conditions. Macromolecular interactions (e.g. protein/protein, protein/DNA and RNA/RNA) are intrinsically harder to maintain at higher temperatures. In the case of TF binding sites, it has been shown that each nucleotide deviation from consensus results in a $\approx$2kbT penalty to the maximum binding free energy for a given TF (where kb is Boltzmann's constant and T is temperature) [78]. Increasing the temperature amplifies the binding free energy penalty for every non-conserved base pair. Therefore at 80°C, mismatches between the Shine-Dalgarno and anti-Shine-Dalgarno sequence are especially severe. Thus, *T. maritima* must overcome the intrinsic challenge of recognizing and retaining contact at the initiation site for both transcription and translation. Our data suggests that high sequence conservation of promoter and RBS sequences is one of the mechanisms used by *T. maritima* to ensure sufficient gene expression. This sequence-level adaptation could be analogous to many others observed in thermophilic organisms such as the amino acid composition of proteins [45, 46] and the GC content of structural RNAs [79].

The minimal intergenic space found in the *T. maritima* genome is reminiscent of a streamlined genome, which could explain the limited regulatory capacity observed in this organism. Inflexibility of metabolic regulons has been previously alluded to for other Thermotogales [80]. Here it is demonstrated that, for most TUs, a lack of physical space exists for transcriptional regulation by TFs. Furthermore, the Short $5'$UTR group carries the minimum number of nucleotides needed to recruit the ribosome based on Shine-Dalgarno/anti-Shine-Dalgarno interactions [70]. Further reduction in 5′UTR length would abolish translation. Short 5′UTRs also reduce the capacity to regulate by limiting 5′UTR interactions [81, 82].

Though thermodynamics and physical space are hypothesized to contribute to the characteristic features of the *T. maritima* genome, the phylogenetic contribution cannot be dismissed. These potential causal factors are difficult to decouple. For RBSs, we were able to determine the impact of phylogeny and optimal growth temperature on RBS binding strength. By analyzing RBSs from 109 bacterial species spanning many phyla and having a diverse range of optimal growth temperatures we were able to demonstrate that both phylogeny and optimal growth temperature were significant determinants of RBSs sequence composition. However, a recent analysis of genome size among species of the order Thermotogales could not resolve the impact of phylogeny from optimal growth temperature [36]. The authors found that a negative correlation between genome size and optimal growth temperature exists within this order but the correlation did not hold when phylogeny was accounted for in the analysis. Interestingly, this study also found that the number of predicted transcriptional regulators and intergenic space is higher in *Mesotoga prima*, a mesophilic member of the Thermotogales. Thus, the relationship between phylogeny and the genome organization is difficult to elucidate without the generation of more datasets similar to the one presented here.

Thermotogae are an ideal phylum for future investigations on the causal impact of factors such as temperature, intergenic space and phylogeny on genome organization. This phylum contains organisms that are found in many diverse environments with a wide range of optimal growth temperatures. Generating multi-omic datasets and analyzing them using an integrated, quantitative workflow for numerous Thermotogae species would enable assessment of various environmental factors in the context of phylogenetic distance. Furthermore, given their phylogenetic depth, characterization of the Thermotogae will also provide insights in the evolutionary trajectory of microbial life on earth.

2.6 Materials and Methods

2.6.1 Culture conditions and physiology

T. maritima MSB8 ATCC derived cultures were grown at 80°C under anoxic conditions in a chemically defined, minimal medium [83]. Cultures were maintained in either serum bottles or pH-controlled (6.5) fermenters with continuous 80% N2, 20% CO_2 sparging. Maltose and acetate concentrations were measured using an HPLC. HPLC parameters were previously described [84]. The following growth conditions were used for omics analysis: 1) log phase, 2) carbon-limited late exponential phase, 3) heat shock and 4) H_2 inhibition. Log phase samples were collected from mid-exponential phase cultures grown in 125 mL serum bottles with 50 mL working volume of media and 10 mM maltose as the sole carbon source. Carbon-limited late exponential phase cultures were grown in pH controlled fermenters with pH control and continuous stripping of evolved hydrogen. Cultures were monitored for OD and maltose concentration and samples were collected upon depletion of maltose. The heat shock condition was achieved by rapidly heating mid-exponential phase cultures grown in serum bottles (similar to the log phase condition) to 90°C and sampled after 10 minutes for transcriptome analysis. This has been shown to result in the heat shock response [85]. H_2 inhibition was achieved by allowing the native evolution of hydrogen to accumulate in serum bottles (similar to the log phase condition). Arrested growth was indicated by successive OD readings that showed no change measured every 30 minutes. Growth profiles for these conditions are shown in Figure S4 in [51].

2.6.2 Genome resequencing and annotation updates

The recent identification of a 9 kb gap in the *T. maritima* MSB8 genome [49] prompted genome resequencing. Genomic DNA was isolated using Promega's Wizard Genomic DNA Purification Kit. Paired-end resequencing libraries were generated following standard Illumina protocols and sequenced on an Illumina GAIIx platform. The updated genome sequence was assembled as follows: (1) Reads were aligned to the 8.9 kb region identified in the *T. maritima* MSB8 DSMZ genomovar

(AGIJ00000000.1) [49] and the TIGR genomovar (AE000512.1) sequence using SHOREmap [86] and MosaikAligner (http://bioinformatics.bc.edu/marthlab/Mosaik). (2) Unaligned reads were *de novo* assembled using Velvet [87] to ensure no additional assemblies were present. (3) The sequence was corrected for SNPs and indels detected during read alignment.

An updated genome annotation was generated using the RAST pipeline with the default parameters [50]. Predicted gene sequences were mapped to the AE000512.1 annotation using a bidirectional Smith-Waterman alignment to identify the corresponding locus tags. Instances where ≥30 bp separated the predicted gene length between annotations were reconciled through manual inspection of gene expression data and bioinformatics predictions. Gene length differences <30 bp could not be reconciled (unless peptide data supported only one annotation). In these cases, the updated sequence annotation was retained.

2.6.3 Transcription start site determination

Total RNA was isolated from log phase cultures using the hot SDS/phenol approach as previously described (http://www.bio.davidson.edu/projects/GCAT/protocols/ecoli/RNApurification.pdf). DNase-treated total RNA samples were recovered using Fisher SurePrep TrueTotal RNA columns. Two biological replicate TSS sequencing libraries were constructed as previously described [24]. Illumina reads were aligned to the updated *T. maritima* genome using the Mosaik Aligner. The number of sequenced reads and the number of aligned reads can be found in Table S10 in [51]. Only uniquely mapped $5'$ ends with ≥5 reads were retained as potential TSSs.

2.6.4 Transcriptome characterization and gene expression

Tiling array and RNA-seq data were generated under log phase growth, carbon-limiting late exponential phase, heat shock and hydrogen inhibited conditions. Total RNA was isolated using the TRIzol (Invitrogen) extraction procedure followed by DNase treatment and purification using either the Qiagen RNeasy Mini

Kit (Tiling Arrays) or the SurePrep TrueTotal RNA columns (RNA-seq).

Custom tiling arrays were synthesized based on the AE000512.1 genome sequence by Roche Nimblegen to carry 71,548 probes with a mean interval of 25 bp. Probe information was remapped to the updated genome sequence. Of the original 71,548 probes, only 125 did not map. Labeled cDNA was generated and processed as previously described [24]. The Transcription Detector algorithm [88] determined probes expressed above background at a FDR = 0.05.

Paired-end, strand-specific RNA-seq was performed using the dUTP method [89] with the following modifications. rRNA was removed with Epicentre's Ribo-Zero rRNA Removal Kit. Subtracted RNA was fragmented for 3 min using Ambion's RNA Fragmentation Reagents. cDNA was generated using Invitrogen's SuperScript III First-Strand Synthesis protocol with random hexamer priming. Illumina reads were aligned to the updated *T. maritima* genome using Bowtie [90] with up to 2 mismatches per read alignment. The number of sequenced reads and the number of aligned reads can be found in Table S10 in [51]. FPKM values were calculated using Cufflinks [91]. Functional RNA transcripts were excluded from FPKM determination.

2.6.5 Proteomics, peptide mapping, and protein abundance quantitation

Proteomics samples and data were generally prepared as previously described [92]. In summary, triplicate samples of both log phase and late exponential phase culture were lysed by French press, and proteins were extracted into global, soluble, and insoluble fractions. The three protein fractions were digested with trypsin (Promega) for 4 h at 37°C and then cleaned-up using C18 or SCX SPE columns (Supelco), as appropriate. Resulting peptide samples were separated in the first dimension by high pH HPLC (Agilent) and then analyzed by LC-MS/MS using C18 resin (Phenomenex) with an expontial gradient on a custom built LC platform coupled to a linear ion trap (LTQ) or a Velos Orbitrap mass spectrometer (Thermo Scientific) operated in data dependent mode. Peptides were identified by SEQUEST (Thermo Scientific) against a six-frame translation of the *T. maritima* genome with no protease specified in the search. Xcorr values were refined to conform to generally accepted criteria and were applied to result in a false discovery rate of 0.16% at the peptide level. Non-quantitative peptide-level data can be found in Table S8 in [51].

Normalized protein abundances can be found in Table S9 in [51]. Quantitative Peptide-level data was extracted from Lerman et al. [93] and mapped to the CP004077 genome annotation. The following criteria were used to filter proteins for quantitative analysis: 1) the protein has a total spectral count ≥2 across all conditions (minimum of two unique peptides or a single unique peptide with two observations), 2) the protein has ≥1 observed peptide under log phase since our data was correlated against log phase transcriptome data. Redundant peptides (i.e. peptides mapping to multiple protein entries) were excluded from the analysis to minimize potential ambiguity. For quantitative analysis, we normalized the observed spectral counts for each ORF by the number of possible fully tryptic peptides in the ORF. The number of possible fully tryptic peptides for each ORF was determined using the Protein Digestion Simulator (http://omics.pnl.gov /software/ProteinDigestionSimulator.php). Default settings were used, except the parameter Max Missed Cleavages was set to 0 and Minimum Residue Count was set to 6. These options require fully tryptic peptides of at least length 6. This program only considers peptides 400-2000 m/z up to a charge state (z) of 3, hence a maximum fragment mass of 6000.

2.6.6 Promoter element motif analysis and position weight matrix (PWM) generation

The process of determining individual $\sigma70$ promoter elements upstream of each unique TU start in *T. maritima* was an iterative process, involving two software packages: BioProspector [94] and MEME [95]. BioProspector is able to identify gapped motif elements so it was used to initially identify *T. maritima* motifs. In BioProspector, sequences 75 bp upstream of TU starts were searched for bipartite elements (6 and 9 bp in width) with a 10-25 bp allowable gap and visualized through WebLogo [96]. MEME provides deterministic position-weight matrices appropriate for information content calculations. The -10 and extended -10 boxes were searched [-1 to -18] upstream of the TSS while the -35 box was searched [-20 to -44]. *E. coli* TUs annotated with σ70 promoters and experimentally validated TSSs in the EcoCyc Database (version 15.0) [65] were extracted for comparative analysis.

A similar approach was applied to identify promoter motifs for alternative sigma factors. *T. maritima* has three annotated alternative sigma factors: RpoE (Tmari_1606), SigH (Tmari_0531) and FliA (Tmari_0904). For RpoE and SigH, the upstream region of TUs having genes showing high differential expression under a given stress condition (heat shock, hydrogen inhibited and carbon-limited late exponential phase) were searched for motif elements. The upstream regions of flagellar gene encoding TUs were searched for a FliA motif. However, no sequence motif could be detected for any of the three alternate sigma factors.

2.6.7 Information content calculations

Position weight matrices (PWMs) for each promoter element were converted to individual information weight matrices using the following formula established in the field of molecular information theory [60]: Riw(b, i) = 2-(-log2f(b, i)), where f(b, i) is taken to be the probability of observing base b at position i. The individual information of a sequence, Iseq, was calculated by summing the relevant entries of Riw. For any particular sequence, only one entry of Riw is relevant among 4 bases for each position i in the sequence. Iseq is measured throughout in bits since the log was base 2 in converting the PWM to Riw.

Iseq reflects sequence conservation for a single sequence, but natural promoters are often formed by multiple promoter elements, each with their own sequences and corresponding Iseq values. When multiple elements are present, variable length spacers are frequently found between the elements. We applied an approach previously described by Shultzaberger et al. [63] to properly account for all possible promoter elements and the variation in their spacing. This allowed us to assess total sequence conservation for an entire promoter. For each promoter, the information content for a particular binding mode was calculated based on the formulas: (1) Mode 1: Iseq_whole_promoter = Iseq(-10 element)+Iseq(-35 element)-GS(d); (2) Mode 2: Iseq_whole_promoter = Iseq(extended-10 element); (3) Mode 3: Iseq_whole_promoter = Iseq(extended-10 element)+Iseq(-35 element)-GS(d). GS(d) is gap surprisal accounting for variable spacing (of length d) between the -10 and -35 elements. GS(d) penalizes for unexpected spacing given the major groove accessibility of B-form DNA and was defined as in equation (3) in Shultzaberger [65] with no small-sample correction factor as the analysis here is performed at genome scale. In accordance with the Shultzaberger model, the space between the -10 and -35 elements was restricted to $15 - 20$ bp as measured from the $3'$ end of the -35 element and the $5'$ end of the -10 element. This limit on the spacer distance Iseq_whole_promoter is measured in bits.

2.6.8 Ribosome binding site energy calculations

The anti-RBS sequence $5'$-UCACCUCCUU-$3'$ ($3'$ end of the 16S rRNA) was selected for this study. The hybrid-2s program in the UNAFold software package [97] was used to compute hybridization energies (ΔG) for all possible 10-mers over the temperature range 20-100°C. This dictionary was mined for three applications: (1) binding energy values for all 10-mer sequences in the updated *T. maritima* genome were computed to aid in annotation improvement, (2) the median positional ΔG for all CDSs ±100 bp from the start codon, and (3) the local minimum ΔG for all CDSs 30 bp upstream of the start codon. RBS binding energies across 109 organisms were calculated using this dictionary. Optimal growth temperatures for all non-Thermotogae bacteria were collected from Takemoto et al. [98] and the protein coding gene annotation for each bacterium was extracted from NCBI. CDS data for all Thermotogae with a complete genome sequence were extracted from NCBI with the exception of *T. maritima* for which the annotation generated in this study was used. For each organism, the median RBS ΔG was calculated from the set of minimum RBS ΔG's found for each CDS 30 bp upstream of the annotated start codon. Three distance matrices were constructed for analysis of the 109 bacterial species for which optimum growth temperatures were found. The matrices included are as follows: (1) the absolute difference of median RBS strength values, (2) the absolute difference of optimal growth temperatures and (3) the distance matrix generated by aligning full-length 16S rRNA gene sequences using ClustalW2 (slow mode) followed by the phylogenetic tree generation script (http://www.ebi.ac.uk/Tools/phylogeny/) with default settings. Next, the Mantel test, which tests the correlation between two distance matrices, was applied to compute the significance of various correlations. The vegan package of R was used with its default settings.

2.6.9 Rho-independent terminator site determination

Intrinsic terminators were predicted using the TransTermHP program [99]. To avoid bias introduced by annotation, no genome annotation was used in prediction of Rho-independent terminators. Only terminator structures predicted with a 100% confidence score were included in the curation of TUs.

2.6.10 Prediction of small RNAs

Small RNAs were predicted with Infernal [56] using cmsearch with default settings against the Rfam 10.0 Database [100] of small RNA families. sRNAs with an E-value<0.01 were manually curated to verify expression. These sRNAs were checked against the sRNA predictions from Rfam and fRNA-DB (http://www.ncrna.org) based on the AE000512.1 genome sequence.

2.6.11 Transcription unit assembly

TU assembly was accomplished through an iterative procedure beginning with tiling array expression data. Tiling array data was processed with two Bioconductor packages for transcript segmentation based on change point analysis: tilingArray (http://www.bioconductor.org/
packages/2.2/bioc/html/tilingArray.html) and DNAcopy
(http://www.bioconductor.org/packages/2.3/bioc/html/DNAcopy.html).
Manual comparison of the output from both packages with array data was used to refine the automated set of transcriptional segments. Additional datasets and

bioinformatics predictions were added and manually curated to fully characterize the TU assembly. TSS and RNA-seq data provided single-base pair resolution of segment boundaries. Intrinsic terminator predictions were also used for $3'$ boundary definition. ncRNAs were identified using the transcript segments. Transcribed regions not associated with a TU and with length exceeding 68 nt (the combined length of the paired end reads with no insert separating them) were quantified using Cufflinks to generate FPKM values across all RNA-seq conditions. Regions with at least two conditions showing FPKM values >8 were retained as putative ncRNAs.

2.6.12 Transcription factor binding site mapping

TF binding sites were extracted from RegPrecise [73] and coordinates were mapped to the updated genome. Table S6 in [51] has the TF binding sites used in Figure 2.3C.

2.6.13 Data deposition

The *T. maritima* MSB8 ATCC (genomovar) genome and annotation are found under Genbank Accession CP004077. RNA-seq, TSS, and tiling array datasets are available in the Gene Expression Omnibus under Accession GSE37483. Proteogenomic data are made available through PNNL (http://omics.pnl.gov) and in Table S8 in [51].

2.7 Acknowledgments

We acknowledge Dmitry Rodionov and Andrei Osterman of the Sanford-Burnham Medical Research Institute, La Jolla, CA, for providing TF binding site loci, and Daniela Bezdan for assistance with the genome annotation. We also thank Bernhard Palsson, Pep Charusanti, and Ramy K. Aziz for guidance with manuscript preparation. Funding: Funding for this work was provided by the Office of Science of the U.S. Department of Energy (DOE) under grants DE-FG02-

08ER64686 and DE-FG02-09ER25917. HL is supported through the National Science Foundation Graduate Research Fellowship under grant DGE1144086. Proteomics capabilities were developed under support from the DOE Office of Biological and Environmental Research (BER) Pan-omics Project and the NIH National Center for Research Resources (RR018522), and a significant portion of this work was performed in the Environmental Molecular Sciences Laboratory (EMSL), a DOE-BER national scientific user facility at Pacific Northwest National Laboratory (PNNL) in Richland, Washington. PNNL is a multi-program national laboratory operated by Battelle Memorial Institute for the DOE under contract DE-AC05-76RLO 1830. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Conceived and designed the experiments: HL JAL VAP KZ. Performed the experiments: HL VAP YT D-HL KZ. Analyzed the data: HL JAL HN ACS-R JNA YQ. Contributed reagents/materials/analysis tools: RDS JNA. Wrote the paper: HL JAL HN KZ.

Chapter 2 in full is a reprint of a published manuscript: Latif H*, Lerman JA*, Portnoy VA, Tarasova Y, Nagarajan H, et al. (2013) The Genome Organization of *Thermotoga maritima* Reflects Its Lifestyle. PLoS Genet 9(4): e1003485. doi:10.1371/journal.pgen.1003485. * indicates equal contribution. The dissertation author was the primary author of this paper responsible for the research. The other authors were Haythem Latif (equal contributor), Vasiliy A. Portnoy, Yekaterina Tarasova, Harish Nagarajan, Alexandra C. Schrimpe-Rutledge, Richard D. Smith, Joshua N. Adkins, Dae-Hee Lee, Yu Qiu, and Karsten Zengler.

Chapter 3

In silico method for modelling metabolism and gene product expression at genome scale

3.1 Abstract

Transcription and translation use raw materials and energy generated metabolically to create the macromolecular machinery responsible for all cellular functions, including metabolism. A biochemically accurate model of molecular biology and metabolism will facilitate comprehensive and quantitative computations of an organism's molecular constitution as a function of genetic and environmental parameters. Here we formulate a model of metabolism and macromolecular expression. Prototyping it using the simple microorganism *Thermotoga maritima*, we show our model accurately simulates variations in cellular composition and gene expression. Moreover, through *in silico* comparative transcriptomics, the model allows the discovery of new regulons and improving the genome and transcription unit annotations. Our method presents a framework for investigating molecular biology and cellular physiology *in silico* and may allow quantitative interpretation of multi-omics data sets in the context of an integrated biochemical description of an organism.

3.2 Introduction

A goal of systems biology is to provide comprehensive biochemical descriptions of organisms that are amenable to mathematical enquiry [101]. These models may then be used to investigate fundamental biological questions [101], guide industrial strain design [102] and provide a systems perspective for analysis of the expanding ocean of omics data [30]. Over the past decade, there has been steady progress in developing genome-scale models of metabolism (M-Models) for basic research and industrial applications [103, 104, 105]. M-Models are stoichiometric representations of the enzymatic and spontaneous biochemical reactions associated with an organism's metabolic network at the genome scale; however, M-Models do not quantitatively describe gene expression (Figure 3.1a). The lack of an explicit representation for enzyme production precludes quantitative interpretation of omics data and can result in biologically implausible predictions [106, 107]. Because M-Models do not contain chemical representations of transcription and translation, to date, it has only been possible to use omics data as *ad hoc* constraints for enzyme activities [108, 109, 110, 111].

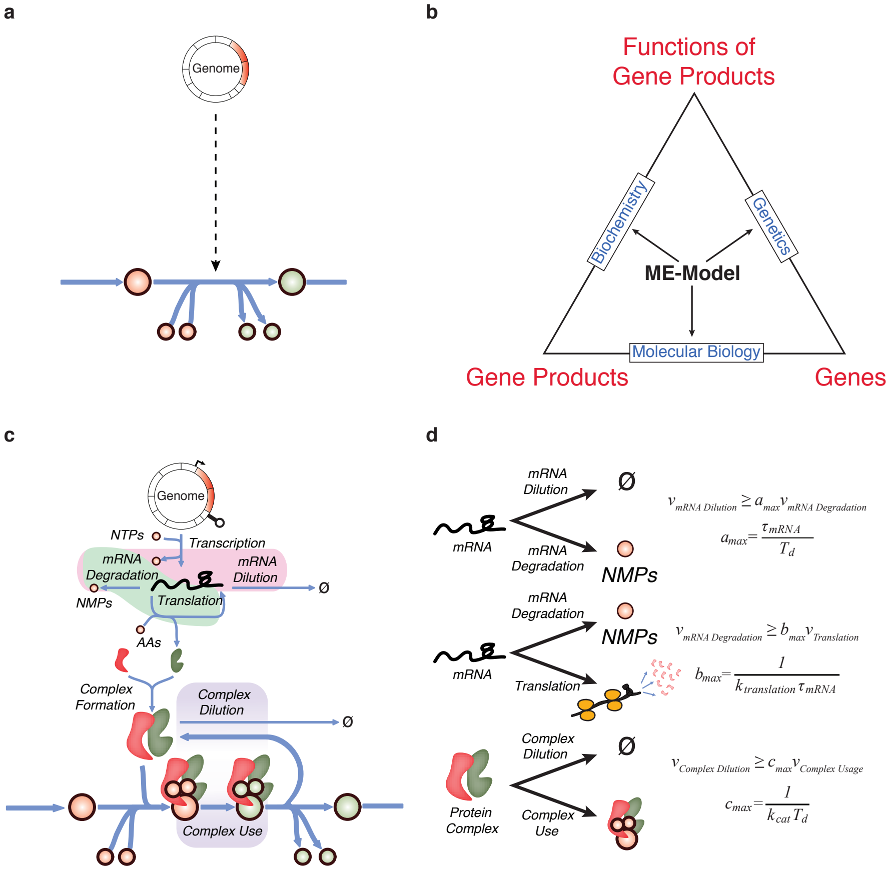


Figure 3.1: **Genome-scale modelling of metabolism and expression.** (a) Modern stoichiometric models of metabolism (M-models) relate genetic loci to their encoded functions through causal Boolean relationships. The gene and its functions are either present or absent. The dashed arrow signifies incomplete and/or uncertain causal knowledge, whereas blue arrows signify mechanistic coverage. (b) ME-Models provide links between the biological sciences. With an integrated model of metabolism and macromolecular expression, it is possible to explore the relationships between gene products, genetic perturbations and gene functions in the context of cellular physiology. (c) Models of metabolism and expression (ME-Models) explicitly account for the genotypephenotype relationship with biochemical representations of transcriptional and translational processes. This facilitates quantitative modelling of the relation between genome content, gene expression and cellular physiology. (d) When simulating cellular physiology, the transcriptional, translational and enzymatic activities are coupled to doubling time (T_d) using constraints that limit transcription and translation rates as well as enzyme efficiency. τ_{mRNA}, mRNA half-life; k_{cat}, catalytic turnover constant; $k_{translation}$, translation rate; v, reaction flux.

A modelling approach that accounts for the production and degradation of a cell's macromolecular machinery will provide a full genetic basis for every computed molecular phenotype (Figure 3.1b). Such computations in turn enable the direct comparison of simulation to omics data and the simulation of variable expression and enzyme activity [112, 113]. In other words, an integrated model of metabolism and macromolecular expression (ME-Model) affords a genetically consistent description of a self-propagating organism at the molecular level and moves us substantially closer to establishing a systems-level quantitative basis for biology.

Here, we developed an ME-Modelling approach for the relatively simplistic microorganism, *Thermotoga maritima*, which metabolizes a variety of feedstocks into valuable products including H_2 [114]. *T. maritima* possess a number of characteristics conducive to systems-level investigations of the genotypephenotype relationship: a compact 1.8-Mb genome [38], wealth of structural proteome data [115], a limited repertoire of transcription factors (TFs) [116] and reduced genome organizational complexity compared with other microbes [51]. Taken together, *T. maritima*'s small set of TFs and reduced genome complexity impose a narrowed range of viable regulatory and functional states [51]. The existence of few regulatory states may simplify the addition of synthetic capabilities and facilitate metabolic engineering efforts by reducing unexpected and irremediable side-effects arising from genetic manipulation [117]. A combination of metabolic versatility and genomic simplicity make *T. maritima* a promising candidate for investigating fundamental relationships between molecular and cellular physiology, both *in silico* and *in vivo*, and for the creation of a minimal chassis for chemical synthesis [118]. Our *T. maritima* ME-Model simulates changes in cellular composition with growth rate, in agreement with previously reported experimental findings [119, 11]. We observed positive correlations between *in silico* and *in vivo* transcriptomes and proteomes for the 651 genes in our ME-Model with statistically significant ($P < 1 \times 10^{-15}$ *t-test*) Pearson correlation coefficients (PCC) of 0.54 and 0.57, respectively. And, when we used our ME-Model as an exploratory platform for an *in silico* comparative transcriptomics study, we discovered putative TF-binding motifs and regulons associated with L-arabinose (L-Arab) and cellobiose metabolism, and improved functional and transcription unit (TU) architecture annotation. Overall, ME-Models provide a chemically and genetically consistent description of an organism, thus they begin to bridge the gap currently separating molecular biology and cellular physiology.

3.3 Results

3.3.1 Genome-scale modelling of metabolism and expression

We developed a network reconstruction and modelling method that includes macromolecular synthesis and post-transcriptional modifications in addition to metabolism (Figure 3.1c; Supplementary Methods in [93]). Specifically, our method accounts for the production of TUs, functional RNAs (that is, transfer RNAs (tRNAs), ribosomal RNAs (rRNAs) and so on) and peptide chains, as well as the assembly of multimeric proteins and dilution of macromolecules to daughter cells during growth. Based on available genomic, structural proteomic and biochemical literature we constructed an ME-Model for *T. maritima* that accounts for the functional activities of 50% of the annotated gene products and, more importantly, mechanistically links these enzyme activities to the genome.

To accurately model self-replicating cells at the molecular level, it is necessary to account for material dilution during cell division as a result of volume doubling, and to provide limits on the number of proteins that may be translated from an messenger RNA before the mRNA decays or is transmitted to a daughter cell. To approximate dilution of transcripts and proteins to daughter cells and prevent infinite translation of peptides from an mRNA, we devised a series of coupling constraints (Figure 3.1d; Supplementary Methods in [93]). These constraints effectively provide upper limits on enzyme expression and activity and are a function of the organism's doubling time (T_d). These coupling constraints may be tuned for specific mRNAs or enzymes if their, respective, degradation rates or catalytic turnover constants (k_{cat}) are known.

Applications of M-Models often involve simulating log-phase cellular growth using flux balance analysis (FBA) [120, 121]. The organism's gross lipid, nucleotide, amino acid (AA) and cofactors, as well as growth-associated and maintenance ATP usage, are experimentally measured. Then, these measurements are integrated with the organism's T_d to define a biomass reaction that approximates the dilution of cellular materials during formation of daughter cells. However, cellular composition is known to vary as a function of T_d and medium [119] –with Schaechter *et al.* indicating that T_d is more influential than growth medium.

Our ME-Model explicitly describes transcription, translation and the dilution of gene products to daughter cells, thus it is unnecessary to use a gross biomass production reaction when simulating growth. Instead, ME-Models contain a structural reaction that accounts for the dilution of structural materials (that is, DNA, cell wall, lipids and so on) during division and the energy cost associated with cellular maintenance of the structure (Supplementary Table S1 in [93]). Conceptually, this structural reaction approximates the production of a cell whose composition varies as a function of environment and growth rate (Figure 3.2a).

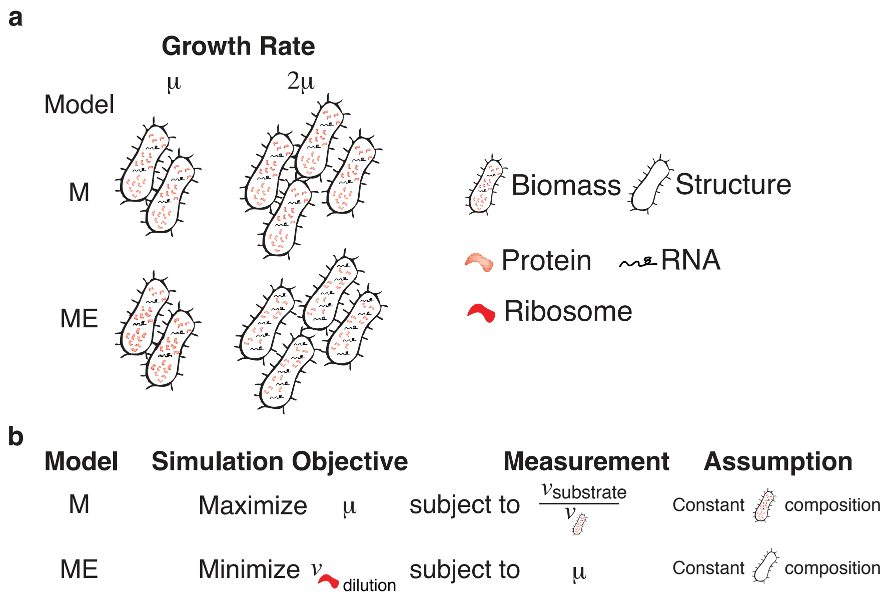


Figure 3.2: **Comparison of M- and ME-Models objective functions and assumptions.** (a) M-Models simulate constant cellular composition (biomass) as a function of specific growth rate (μ), whereas ME-Models simulate constant structural composition with variable composition of proteins and transcripts. (b) Linear programming simulations with M-Models are designed to identify the maximum μ that is subject to experimentally measured substrate uptake rates. Only biomass yields are predicted as μ enters indirectly as an input through the supplied substrate uptake rate (see the measurement column for M-Models). Importantly, the substrate uptake rate is derived by normalizing to biomass production. Linear programming simulations with ME-Models aim to identify the minimum ribosome production rate required to support an experimentally determined μ. μ enters into the coupling constraints and so it must be supplied (or sampled) as the problem would otherwise be a Nonlinear Program (NLP). As all M-Models reactions are contained within the ME-Models, ME-Models can simulate all M-Models objectives in addition to the broad range of objectives associated with macromolecular expression.

3.3.2 Molecularly efficient simulation of cellular physiology

The RNA-to-protein mass ratio (r) has been observed to increase as a function of specific growth rate (μ) [119, 11] and decreases as a function of translation efficiency [11]. Schaechter *et al.* also observed an increase in the number of ribonucleoprotein particles with increasing μ, whereas the translation rate per ribonucleoprotein particle was relatively constant [119]. The increase in r and ribonucleoproteins may be due to the reduced number of translation events mediated by a ribosome as T_d decreases.

To ascertain whether our ME-Model recapitulated the observed increases in r, ribosomal RNA and proteins with increasing μ, we simulated a range of growth rates in a defined minimal medium [83] (Supplementary Table S2 in [93]). To simulate the molecular physiology of *T. maritima* for a particular μ, we used FBA [121] subject to linear programming optimization [122] to identify the minimum ribosome production rate required to support a given μ (Figure 3.2b). Ribosome production has been shown to be linearly correlated with growth rate in *Escherichia coli* [11, 123, 124]. Assuming that efficient use of enzymes contributes to the fitness of an evolutionarily adapted lineage [125], we would expect a successful organism to produce the minimal amount of ribosomes required to support expression of the proteome.

Consistent with experimental observations [119, 11], our ME-Model simulated an increase in r with increasing μ and with decreasing translation efficiency (Figure 3.3a). We observed that the fraction of the transcriptome associated with ribosomal RNA *in silico* increased with μ (Figure 3.3b). In addition, the ribosomal proteins account for a larger proportion of the total proteome as μ increases (Figure 3.3c). These results indicate that it is possible to mechanistically model changes in cellular physiology that have only recently yielded to phenomenological modelling [11].

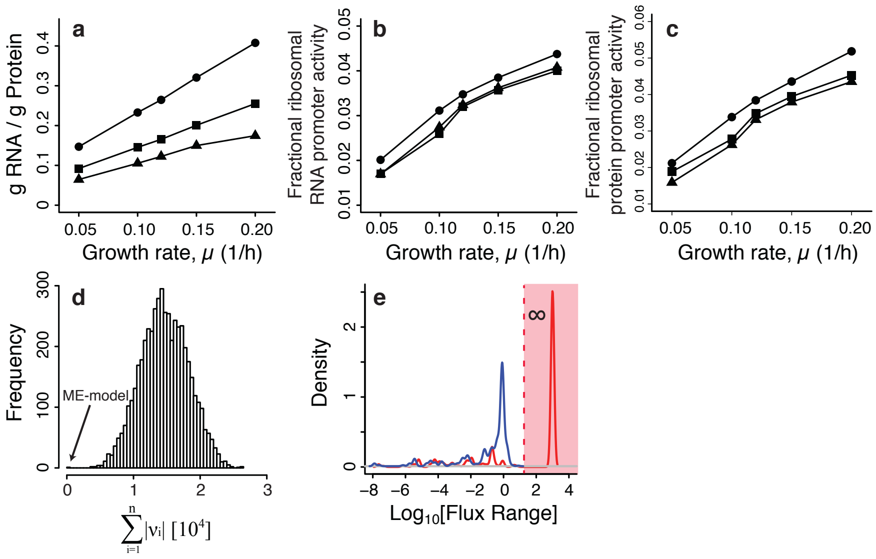


Figure 3.3: **Simulation of variable cellular composition and efficient use of enzymes.** (a) With our ME-model, the RNA/protein ratio increases linearly with growth rate and with a slope proportional to translational capacity in amino acids per second (circles: 5 AA/s, squares: 10 AA/s, triangles: 20 AA/s). (b) Ribosomal RNA (rRNA) synthesis increases, relative to total RNA synthesis, with growth rate (symbols as in a). (c) Ribosomal protein promoter activity increases, relative to total RNA synthesis, with growth rate (symbols as in a). (d) Random sampling of the M-Model solution space indicates that the M-Model solution space contains numerous internal solutions with a broad range of total network flux. The probability of finding an M-Model solution as efficient as an ME-Model simulation is 2.1×10^{-5}; the probability was calculated from a normal distribution constructed from the M-Model sample space. The M-Model sample contains 5,000 flux vectors randomly sampled from the M-Model solution space. (e) Smooth estimate of the density of the flux ranges for the metabolic enzymes that may be simulated while maintaining the objective for efficient growth with a 1% tolerance (M-Model: red line, ME-Model: blue line). The shaded area denotes biologically unrealistic flux values. All simulations were performed with an *in silico* minimal medium with maltose as the sole carbon source.

With M-Models, the cellular macromolecular composition is constant, ergo they cannot reproduce the observed increases in r or ribosomes with increasing μ. Although it is possible to empirically determine a relationship between gross biomass composition and μ and then use this relationship to study variable composition in M-Models [126], the M-Models will compute a solution space where the range of activity for a number of enzymes may be rather broad and even infinite [106], if not specifically constrained. The biologically implausible sections of the M-Model solution space are due, in large part, to unconstrained thermodynamically infeasible internal loops that can operate at an arbitrary flux level [107]. These arbitrary activities contradict previous observations that efficient organisms should maintain a minimal total flux through their biochemical network [125, 127].

By explicitly accounting for enzyme expression and activity, ME-Model simulations should identify the set of proteins that will result in optimally efficient conversion of growth substrates into cells. To determine whether our ME-Model was more economic in terms of enzyme usage than the M-Model, we compared our ME-Model simulation to a random sampling of the M-Model solution space [106]. After we fit a normal distribution to the sampled M-Model space, we found that there is a small (2.1×10^{-5}) probability of finding an M-Model solution as efficient as the ME-Model solution (Figure 3.3d). Because ME-Models explicitly account for the costs of enzyme expression and dilution to daughter cells, the most efficient growth simulations will minimize the materials required to assemble the cell; that is, ME-Models will efficiently use enzymes when simulating a μ.

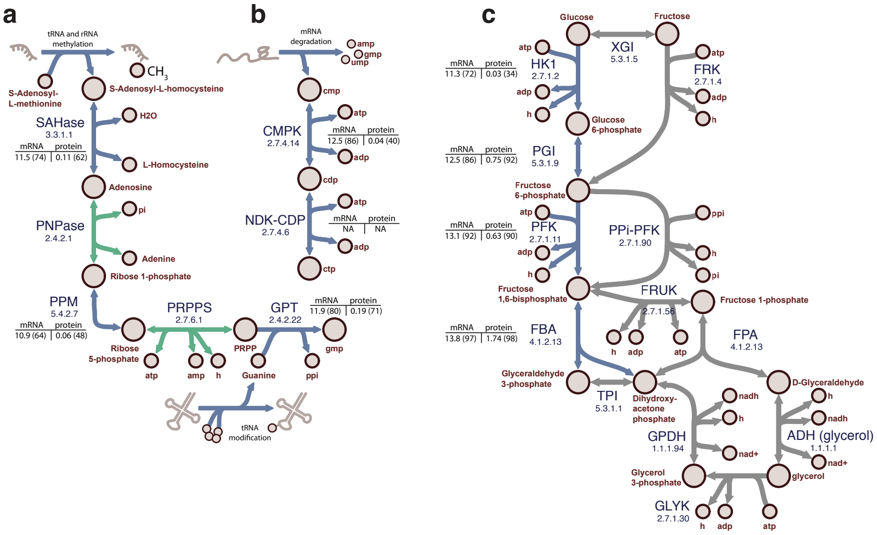
To compare the range of permissible, that is, computationally feasible, activity for each metabolic enzyme in the ME-Model versus the M-Model, we performed flux variability analysis. Flux variability analysis identifies the flux range that each reaction may carry given that the model must also simulate the specified objective value, such as μ, with a set tolerance. The permissible enzyme activities for simulating efficient growth with a 1% tolerance tended to have smaller ranges in the ME-Model compared with the M-Model (Figure 3.3e; Supplementary Data 1 in [93]), highlighting the sharply reduced flexibility in the ME-Model solution space when simulating optimal growth.

Our ME-Model contains gene products that carry out 142 of the 206 functions estimated as essential for a minimal organism [128], whereas the M-Model contains only 65 of these core functions. With the ME-Model, 120 of the 142 functions were essential for ribosome production, whereas only 23 of the 65 functions in the M-Model were essential for biomass production (Supplementary Data 2 in [93]). This broader coverage of cellular functions means that ME-Models may be used for *in silico* investigations of phenotypic states that are inaccessible to M-Models.

3.3.3 Gene product production and turnover alters pathway activity

In addition to simulating variable cellular composition and effectively eliminating the infinite catalysis problem, there are a number of metabolic activities that are required for optimally efficient growth with the ME-Model but not with the M-Model (Figure 3.4). These differences are due to the ME-Model producing small metabolites as by-products of gene expression and explicitly accounting for the material and energy costs of macromolecule production and turnover. The ME-Model includes metabolic activities for recycling S-adenosylhomocysteine, which is a by-product of rRNA and tRNA methylation, and guanine, which is a by-product of queuosine modification of various tRNAs (Figure 3.4a). The ME-Model, also, produces CTP from CMP that is produced during mRNA degradation (Figure 3.4b). Interestingly, the M-Model does not require CDP production to simulate growth, whereas CDP production is essential in the ME-Model. The ME-Model exhibits frugality with respect to central metabolic reactions (Figure 3.4c) and proposes the canonical gylcolytic pathway during efficient growth, whereas the M-Model indicates that alternate pathways are as efficient. When the efficiency requirement is relaxed these less-efficient pathways may be active in the ME-Model solution space (Supplementary Data 1 in [93]). The genes associated with optimal activities tended to be strongly expressed (approximately 60th90th percentile) in transcriptome data.

Figure 3.4: **Metabolic reactions required for efficient growth with the ME-Model but not the M-Model.** (a) Recycling of by-products of RNA modifications. Adenosylhomocysteinease (SAHase) hydrolyses S-adenosylhomocysteine (SAH) to L-homocysteine (L-HCys) and adenosine. Purine nucleoside phosphorylase (PNP) phosphorylases adenosine to adenine and ribose-1-phosphate (Rib-1-P). Rib-1-P is converted to ribose-5-phosphate (Rib-5-P) by phosphopentomutase (PPM). Phosphoribosylpyrophosphate synthetase (PRPPS) phosphorylates Rib-5-P to produce 5-phosphoribosol-1-pyrophosphate (PRPP). Guanine phosphoribosyltransferase (GPT) produces GMP from the reaction of PRPP and guanine, which is a by-product of tRNA metabolism. (b) CMP produced during mRNA degradation is recycled to CTP using cytidylate kinase (CMPK) and nucleoside-diphosphate kinase (NDK-CDP). (c) The ME-model uses the canonical glycolytic pathway, whereas with the M-Model can circumvent portions during optimal growth simulations. The canonical pathway involves phosphorylation of D-glucose (D-Glc) to glucose-6-phospate (G6P) by hexokinase (HK1). G6P is isomerized to fructose-6-phosphate (F6P) by phosphoglucose isomerase (PGI). F6P is phosphorylated to fructose-1,6-bisphosphate (FBP) by phosphofructokinase (PFK). FBP is metabolized to glyceraldehyde-3-phosphate (G3P) and dihydroxyacetone phosphate (DHAP) by FBP aldolase (FBA). The M-Model can circumvent the HK1/PGI portion with glucose/xylose isomerase (GXI) and fructokinase (FRK); however, HK1 or PGI must also be expressed because G6P is an essential metabolite. PFK can be circumvented by diphosphate-fructose-6-phosphate 1-phosphotransferase (PPi-PFK). FBA can be circumvented by a pathway using 1-phosphofructokinase (FRUK), fructose-1-phosphate aldolase (FPA), alcohol dehydrogenase (ADH(glycerol)), glycerol kinase (GLYK), glycerol-3-phosphate dehydrogenase (GPDH) and triose phosphate isomerase (TPI). Enzyme commission numbers are provided for each reaction. mRNA and protein expression (and quantile) values are provided. Flux variability analysis was performed for simulated growth on maltose minimal medium. Blue arrows: reactions required for optimally efficient growth with the ME-Model, but not the M-Model. Green arrows: active reactions in a single maltose minimal medium simulation shown to put results into pathway context. Grey arrows: alternate optimal pathways in the M-Model.


a
tRNA and rRNA methylation
CH_3
S-Adenosyl-L-methionine
S-Adenosyl-L-homocysteine
SAHase
3.3.1.1
mRNA protein
11.5 (74) | 0.11 (62)
H2O
L-Homocysteine
Adenosine
PNPase
2.4.2.1
pi
Adenine
Ribose 1-phosphate
PPM
5.4.2.7
mRNA protein
10.9 (64) | 0.06 (48)
Ribose 5-phosphate
PRPPS
2.7.6.1
atp
amp
h
PRPP
GPT
2.4.2.22
mRNA protein
11.9 (80) | 0.19 (71)
gmp
Guanine
ppi
tRNA modification
b
mRNA degradation
amp
gmp
ump
cmp
CMPK
2.7.4.14
atp
mRNA protein
12.5 (86) | 0.04 (40)
adp
cdp
NDK-CDP
2.7.4.6
atp
mRNA protein
NA | NA
adp
ctp
c
Glucose
Fructose
XGI
5.3.1.5
atp
HK1
2.7.1.2
mRNA protein
11.3 (72) | 0.03 (34)
adp
h
Glucose 6-phosphate
atp
FRK
2.7.1.4
adp
h
mRNA protein
12.5 (86) | 0.75 (92)
PGI
5.3.1.9
Fructose 6-phosphate
atp
PFK
2.7.1.11
mRNA protein
13.1 (92) | 0.63 (90)
adp
h
PPi-PFK
2.7.1.90
ppi
h
pi
Fructose 1,6-bisphosphate
FRUK
2.7.1.56
Fructose 1-phosphate
h
adp
atp
mRNA protein
13.8 (97) | 1.74 (98)
FBA
4.1.2.13
FPA
4.1.2.13
Glyceraldehyde 3-phosphate
TPI
5.3.1.1
Dihydroxy-acetone phosphate
D-Glyceraldehyde
nadh
h
GPDH
1.1.1.94
nad+
h
nadh
ADH (glycerol)
1.1.1.1
nad+
Glycerol 3-phosphate
glycerol
GLYK
2.7.1.30
h
adp
atp

These differences highlight the interplay between macromolecular synthesis and degradation, metabolism and salvage, and optimal use of the proteome. The ME-Models allow a fine resolution view of these processes and their simultaneous reconciliation. Not only can one analyse specific pathways in isolation, such as the three examples given above, but it is now possible to investigate in detail the coordination of functions within an organism's biochemical repertoire.

3.3.4 Simulation of systems-level molecular phenotypes

To assess our ME-Model's ability to simulate systems-level molecular phenotypes, we compared model predictions to substrate consumption, product secretion, AA composition, transcriptome and proteome measurements. With the only external constraints for the ME-Model being the experimentally determined μ during log-phase growth in maltose minimal medium at 80 °C, our model accurately predicted maltose consumption and acetate and H_2 secretion (Figure 3.5a; Supplementary Table S3 in [93]). Predicted AA incorporation was linearly correlated (0.79 PCC; $P < 4.1 \times 10^{-5}$ *t-test*) with measured AA composition (Figure 3.5b).

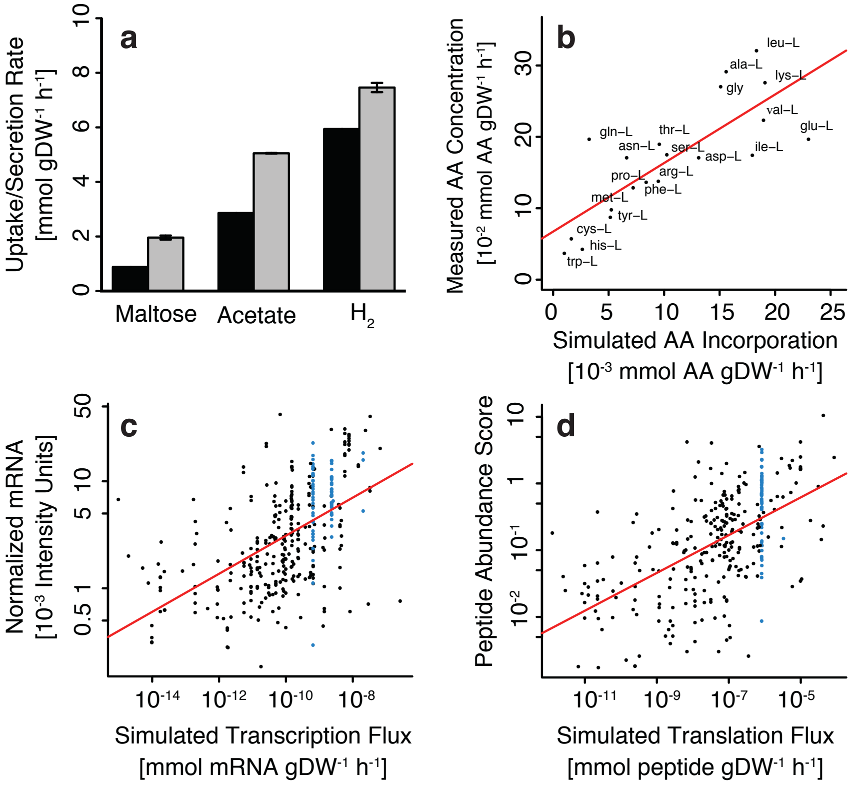


Figure 3.5: **The ME-Model accurately simulates molecular phenotypes during log-phase growth.** (a) The ME-Model accurately simulates H_2 and acetate secretion with maltose uptake when constrained with a measured growth rate (n=2). Experiment: grey bars, simulation: black bars. (b) The *in silico* ribosome incorporates the 20 amino acids at rates proportional (Pearson correlation coefficient=0.79; $P < 4.1 \times 10^{-5}$ *t-test*) to the bulk amino-acid composition of a *T. maritima* cell as measured by high-performance liquid chromatography (n=1). (c) Simulated transcriptome fluxes are significantly ($P < 2.2 \times 10^{-16}$ *t-test*) and positively correlated (Pearson correlation coefficient=0.54) with semi-quantitative *in vivo* transcriptome measurements (n=4). RNAs containing ribosomal proteins (blue) were expressed stoichiometrically in simulations but exhibited variability in measurements. (d) Simulated translation fluxes are significantly ($P < 2.2 \times 10^{-16}$ *t-test*) and positively correlated (Pearson correlation coefficient=0.57) with semiquantitative *in vivo* proteomic measurements (n=3). Ribosomal proteins (blue) were expressed stoichiometrically in simulations but exhibited variability in measurements.

FBA simulates reaction fluxes, whereas transcriptomics and proteomics technologies provide semiquantitative measurements of expressed gene product abundance. Thus, the simulated fluxes through the transcriptome and proteome do not directly approximate the respective omics measurements; however, for macromolecules there should be a positive correlation between gene and protein synthesis fluxes and the respective gene product abundances during log-phase growth. In other words, proteins and genes are relatively stable and when an organism is growing at steady state a relative increase in expression rate for a protein will effectively increase the quantity of that protein.

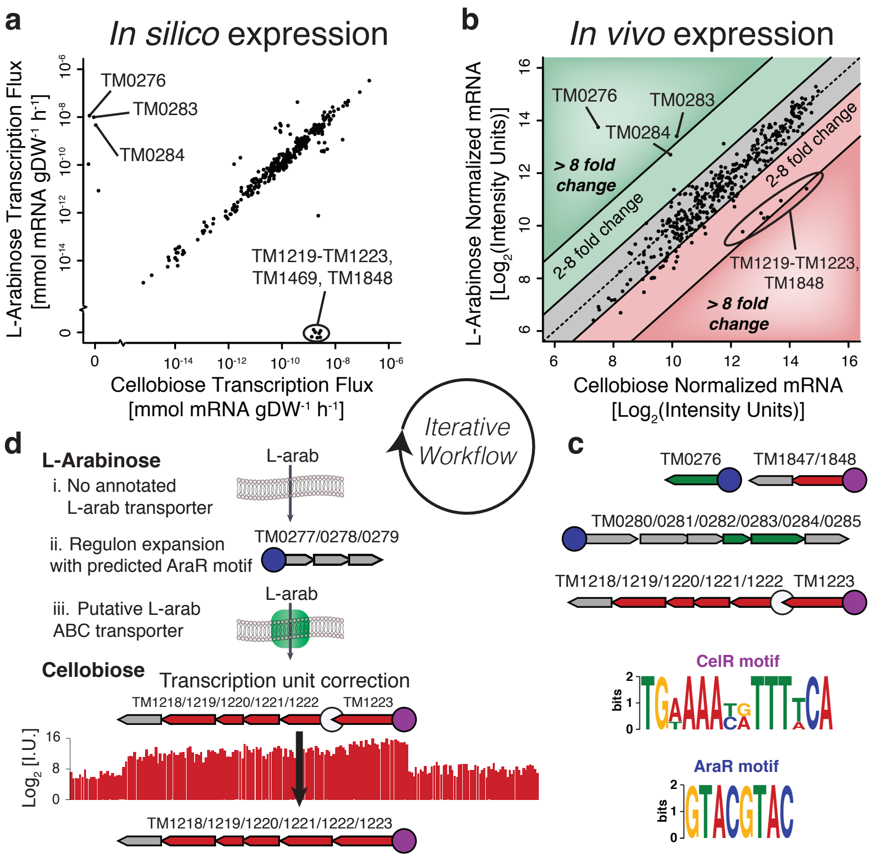
Interestingly, when we compared the simulated transcriptome and proteome fluxes to transcriptome and proteome measurements, respectively, there were statistically significant ($P < 2.2 \times 10^{-16}$ *t-test*) positive correlations for both the transcriptome (0.54 PCC; Figure 3.5c) and the proteome (0.57 PCC; Figure 3.5d). This degree of concordance was unexpected because the model does not account for transcriptional regulation or transcript-specific RNA degradation rates. However, this concordance may be the result of our simulation objective being aligned with *T. maritima*'s regulatory programme, whereas a decreased concordance would be expected if the regulatory network was responding to a stress. We have previously observed a tendency to increase the expression of metabolically efficient pathways, and decrease inefficient alternatives, by *E. coli* after adaptive evolution under growth selection pressure [127]. Also, we have observed that *T. maritima*'s genome is highly active with >89% of the protein-coding genes expressed in diverse conditions [51], which could indicate a general eschewal of complex and expensive circuitry within the global regulatory strategy.

Approximately 30% of *T. maritima*'s genome is not functionally annotated and 50% of the functionally annotated genes fall outside of the scope of our ME-Model. A number of genes not accounted in our model were expressed *in vivo* (Supplementary Fig. S1 in [93]), and the costs of their expression as well as their functional activities may contribute to the differences between simulation and measurement. In addition, unknown regulatory features might be responsible for irregularities observed when comparing simulation to the measurement. For instance, ribosomal RNAs and proteins are expected to be expressed at stoichiometric ratios, as occurs with the simulation, yet there is sizable variability in their measured values (Figure 3.5c,d, blue colouring). These results illustrate that it is possible to sketch a molecular description of a replicating organism solely from simple, but stoichiometrically accurate, chemical equations represented on a genome scale.

3.3.5 *In silico* gene expression profiling drives discovery

With our ME-Model it is now possible to compute the gene expression profile associated with growth in a specific condition or for a specific mutant. These gene expression profiles may then be compared to identify genes that are likely differentially regulated. The set of differentially expressed *in silico* genes may then be used to drive biological discovery or improve our model (Figure 3.6).

Figure 3.6: ***In silico*** **transcriptome profiling drives biological discovery.** (a) *In silico* comparative transcriptomics identifies sets of genes that are differentially regulated for growth in L-arabinose (L-Arab) versus growth in cellobiose minimal media. TM0276, TM0283 and TM0284 are essential for metabolizing L-Arab, whereas TM1219TM1223, TM1469 and TM1848 are essential for metabolizing cellobiose. (b) *In vivo* transcriptome measurements (n=2) confirm the *in silico* transcriptomics predictions for differential expression of genes when metabolizing L-Arab or cellobiose. (c) Two distinct putative TF-binding motifs are present upstream of the TUs containing genes differentially expressed *in silico* when simulating growth in L-Arab versus cellobiose minimal media. The motif upstream of the genes upregulated during growth in L-Arab medium is termed AraR, whereas the motif of the genes upregulated during growth in cellobiose medium is termed CelR. Genes (grey: not in the model, green: upregulated by L-arabinose, red: upregulated by cellobiose) organized into TUs involved in the shift are shown. Each TU contains a promoter region (circle) arbitrarily taken to be 75 base pairs upstream of the first gene in the TU. Promoters found to contain the AraR or CelR motifs are coloured blue and purple, respectively. (d) Searching *T. maritima*'s genome for additional AraR and CelR motifs results in new biological knowledge. Although *T. maritima* can metabolize L-Arab, there is no annotated transporter in the current genome. We identified a putative AraR motif in a single TU (TM0277/0278/0279) not contained in the ME-Model. Analysis of the TM0277/0278/0279 TU with the SEED RAST server indicated that the genes are likely components of an ABC transporter that may be associated with L-Arab transport. The CelR motif was not present in the promoter region upstream of the cellobiose transporter operon (TM1218/1219/1220/1221/1222); however, the CelR motif was present in the promoter of the TU (TM1223) directly upstream of the cellobiose transport operon. Examination of the *in vivo* transcriptome measurement indicates that the cellobiose transporter operon belongs to the same TU as that of TM1223.


a
In silico expression
L-Arabinose Transcription Flux
[mmol mRNA gDW-1 h-1]
10-6
10-8
10-10
10-12
10-14
0
TM0276
TM0283
TM0284
TM1219-TM1223,
TM1469, TM1848
0
10-14
10-12
10-10
10-8
10-6
Cellobiose Transcription Flux
[mmol mRNA gDW-1 h-1]
b
In vivo expression
L-Arabinose Normalized mRNA
[Log2(Intensity Units)]
16
14
12
10
8
6
TM0276
TM0283
TM0284
> 8 fold
change
2-8 fold change
2-8 fold change
TM1219-TM1223,
TM1848
> 8 fold
change
6
8
10
12
14
16
Cellobiose Normalized mRNA
[Log2(Intensity Units)]
Iterative
Workflow
c
TM0276
TM1847/1848
TM0280/0281/0282/0283/0284/0285
TM1218/1219/1220/1221/1222
TM1223
CelR motif
bits
2
1
0
TGAAAATGTTTTCA
AraR motif
bits
2
1
0
GTACGTAC
d
L-Arabinose
i. No annotated
L-arab transporter
L-arab
ii. Regulon expansion
with predicted AraR motif
TM0277/0278/0279
iii. Putative L-arab
ABC transporter
L-arab
Cellobiose
Transcription unit correction
TM1218/1219/1220/1221/1222
TM1223
Log2 [I.U.]
16
8
0
TM1218/1219/1220/1221/1222/1223

Towards this end, we computed the transcriptome profiles for *T. maritima* grown in a minimal medium with either L-Arab or cellobiose as the sole carbon source (Figure 3.6a). Our computations identified genes that were exclusively expressed and essential for growth with each carbon source. Because these genes are essential for growth on the respective substrate they are conditionally essential genes. Conditionally essential genes are often subject to transcriptional regulation, however, they may be constitutively expressed. To assess whether the genes were differentially expressed *in vivo*, we measured the transcriptome of *T. maritima* growing in minimal medium with L-Arab or cellobiose as the carbon source. The genes with the strongest differential expression *in vivo* were among the set of differentially expressed genes *in silico* (Figure 3.6b) providing supporting evidence for the presence of transcriptional regulation.

Conditionally expressed genes may be regulated by the same TF [129]. The presence of a common motif in the promoter regions of a set of genes may indicate regulation by a common TF. To identify potential TF-binding motifs, we scanned the promoter and upstream regions of the *in silico* differentially expressed genes with MEME (Multiple Expectation Maximum for Motif Elicitation) [130]. Surprisingly, there was a high-scoring motif for the genes essential for growth on L-Arab and a high-scoring motif for the genes essential for growth on cellobiose (Figure 3.6c). The motif found upstream of the L-Arab upregulated genes is similar to the AraR motif from *Bacillus subtilis* [131] (Supplementary Fig. S2 in [93]). Also, the motif upstream of the cellobiose upregulated genes bears resemblance to catabolite-responsive elements (*cres*), known to have an important global role in catabolite repression through the binding of the CcpA protein in *B. subtilis* [132]. Here, we term the motif the CelR motif, as the regulated genes are involved in cellobiose metabolism. These discoveries highlight how ME-Model simulations can guide discovery of new regulons.

After identifying the putative AraR and CelR motifs, we scanned *T. maritima*'s genome for the presence of other members of the putative regulons. For the nondegenerate AraR motif 5′-GTACGTAC-3′, we identified a single additional instance in an intergenic region upstream of the TU-containing genes TM0277, TM0278 and TM0279 (Figure 3.6d). These genes were induced when L-Arab was the carbon source, but not when cellobiose or maltose serves as the carbon source (Supplementary Fig. S3 in [93]). L-Arab transport is an orphaned activity in our model, which means that *T. maritima* may import L-Arab, however, the responsible loci are not known. When we examined these genes using the SEED RAST server [50], TM0278 and TM0279 were classified as permeases of an ABC transporter putatively involved in L-Arab utilization, whereas TM0277 was not classified because it was annotated as containing an authentic frameshift [133]. Recent resequencing of *T. maritima*'s genome [51] refute the initial annotation that TM0277 contains a frameshift mutation; and the SEED RAST annotation for TM0277 is a predicted sugar-binding protein for an arabinoside ABC transporter. Interestingly, the TUs containing ABC transporters for maltose and chitobiose are organized in the same manner: a binding protein followed by two permeases. The presence of the AraR motif, the strong upregulation of the TM0277/TM0278/TM0279 TU in response to L-Arab *in vivo*, the SEED RAST classification and resequenced genome strongly suggest that we have identified a functional L-Arab transport system in this organism. This discovery illustrates how *in silico* molecular biology at the genome scale can be used to expand regulons and improve genome annotation.

When we scanned *T. maritima*'s genome for matches to the degenerate CelR motif TGWAAAYRTTTWCA, the promoter regions of TUs associated with cellobiose metabolism were identified. Interestingly, the promoter region of the TU-containing TM1222, TM1221, TM1220, TM1219 and TM1218 did not contain a CelR motif (Figure 3.6c,d). TM1222, TM1221, TM1220 and TM1219 encode for a cellobiose ABC transporter, while TM1218 is annotated as a LacI family transcription regulator. However, the promoter region of the TU for TM1233, which is directly upstream of TM1222, contains the CelR motif. TM1233 encodes for the cellobiose-binding protein that facilitates cellobiose transport. In the TU architecture of our model, there was a predicted Rho-independent terminator following TM1223 that resulted in a new TU starting with TM1222. However, no promoter was detected in the intergenic region between TM1223 and TM1222 using PromBase [77]. This result leads us to believe that the initial assignment of TM1223 and TM1222 to separate TUs was incorrect (Figure 3.6d). The presence of the cellobiose transport system in the updated TU, the strong CelR motif and the annotation of TM1218 as a TF suggest that TM1218 may encode for CelR.

3.4 Discussion

Our ME-Modelling approach represents a fundamental advance in the evolution of genome-scale biochemical models of life and significantly broadens the scope of microbial systems biology. It is now possible to ask systems-level questions *in silico* beyond metabolism and quantitatively analyse, in a bottom-up and mechanistic manner, a variety of omics data in the context of a growing organism. For instance, we can use a systems perspective to identify the minimal number of genes required to support homeostasis and replication–120 of the 142 of the proposed minimal bacterial genome [128] were essential for ribosome production in maltose minimal medium (Supplementary Data 2 in [93]).

Not only can ME-Models predict global phenotypes that are traditionally employed with M-Models, such as maximal growth rate in a defined medium, but they can also be used to calculate whether the system has any material and energy reserves available for ancillary functions. For example, the measured maltose consumption rate was greater than the one that we calculated for economically efficient growth (Figure 3.5a). This discrepancy between measurement and simulation could indicate that *T. maritima* does not strive for economic efficiency or represent the portion of sugar used to support the activities of the unannotated genes or regulatory circuitry. Given that the expression levels for the gene products associated with the more efficient pathways were highly expressed (Figure 3.4c), we are disposed towards the latter. Although the ME-Model does not account for regulatory events, the presence of a strong discordance between simulation and measurement would indicate that factors other than economic efficiency are influencing the expressome, thus informing hypothesis generation. For example, if a more expensive isozyme was expressed *in vivo* than *in silico*, then it would be possible to estimate the improvement in k_{cat} required for the expensive isozyme to offset its higher materials and energy costs.

Technological advances have contributed to an expanding ocean of omics data that has been under-explored [30]. Omics data have been under-analysed, in part, due to the lack of a mechanistic systems-level framework for analysing myriad molecular components in the context of cellular physiology. To date, with the notable exception of C13 metabolic flux analysis, it has only been possible to perform indirect comparative analysis between omics data and M-Models [127] or to neglect the complexity of the genotypephenotype relationship and use omics data as *ad hoc* constraints for M-Model enzyme activities [108, 109, 110, 111]. Because ME-Models explicitly represent gene expression, directly investigating omics data in the context of the whole is now feasible.

Viewing multi-omics data in the context of biochemically and genomically consistent ME-Models may allow us to extract more value from legacy and future omics data. Comparing *in silico* and *in vitro* transcriptomes, or proteomes, can highlight under-explored areas of molecular biology. For example, a set of genes highly expressed *in silico* but not expressed *in vivo* may indicate the presence of transcriptional regulation. Differential expression of a class of genes may indicate incompleteness in our knowledge of how those gene products interact or allude to, heretofore unknown, moonlighting functions. For instance, in the case of ribosomal proteins (Figure 3.5c,d, blue) the model predicts uniform expression, whereas omics measurements exhibit variability. The model was designed based on evidence that ribosomal protein synthesis is highly coordinated [134], and does not account for feedback circuits affecting degradation rates that have yet to be fully elucidated [134, 135].

Although there is a positive correlation between the simulated transcriptome fluxes and semiquantitative transcriptome data there was still a substantial amount of dispersion (Figure 3.5c). When comparing *in silico* and *in vivo* transcriptome measurements it is important to realize that both are approximations of the transcript levels in an organism, and that omics technologies have been inherently noisy to date [136]. Incomplete knowledge, such as a lack of specific translation efficacy for each protein and degradation rates for each mRNA, and lack of signalling and regulatory circuitry will contribute to deviations from reality by ME-Model simulations. Similarly, probe-binding and sample-labelling efficacies, as well as other technical issues, serve as barriers to absolute quantitative transcriptome measurements [137].

Although it is a non-trivial endeavour to identify the source of all variation between the simulated and measured transcriptomes, it is possible to use the ME-Model for comparative transcriptomics approaches similar to two-channel DNA microarray studies. Despite the early technological limitations of DNA microarrays, biological discovery was enabled by performing comparative transcriptomics [138, 139, 140, 141]. Transcriptome profiling has been used extensively to identify genes that are differentially regulated as a function of genetics and environment [138]. Analysis of differentially expressed genes has contributed to the identification of gene products responsible for unannotated enzymatic activities [139]. In combination with sequence analysis, differential gene expression data can be used to investigate transcriptional regulation [140, 141].

We devised and implemented a workflow for *in silico* comparative transcriptomics, which resulted in the discovery of new regulons and improved both genome and TU annotation (Figure 3.6ad). The similarities between the comparative transcriptomics *in silico* (Figure 3.6a) and *in vivo* (Figure 3.6b) studies are striking, given the variation observed between the simulated and measured transcriptomes (Figure 3.5c) –this emphasizes that, in spite of its shortcomings, the ME-Modelling framework is a powerful tool for biological research.

Finally, ME-Models enable integrated molecular biology on a genome scale while accounting for the metabolic requirements, which partially fulfills the challenge of Project K [142] and moves us one step closer to a molecular representation of CellMap [101].

3.5 Methods

3.5.1 Network reconstruction procedure

The detailed procedure and formalism are described in detail in the Supplementary Methods in [93]. Our method accounts for biochemical reactions associated with transcription of TUs, TU degradation, translation, protein maturation, RNA processing, protein complex formation, ribosomal assembly, rRNA modification, tRNA modification, tRNA charging, aminoacyl-tRNA synthetase charging, charging EF-Tu, cleavage of polycistronic TUs to release stable RNA products, sources, sinks and tRNA activation (EF-TU) as well as metabolism. In our formalism, metabolic reactions are represented as multi-step processes including substrate binding by the enzyme and dissociation of substrateenzyme complex to enzyme and products. The metabolic content for our reconstruction was based on the previously published model [115], with updates to correct errors and incorporate new data (Supplementary Data 3 in [93]).

The molecular machinery (for example, proteins, genes, RNAs) involved in macromolecular synthesis were identified from the genome annotation [38], SEED subsystem analysis [143], comparative genomics analysis of the *E. coli* model [124] and KEGG [133]. The functions of each of the 159 proteins associated with macromolecular synthesis in *T. maritima* were determined by primary literature when available. When no primary literature was available, the Uniprot [144] and SEED [143] databases were used to infer function by homology. All proteins currently believed to be used for macromolecular synthesis by *T. maritima* are enumerated in Supplementary Data 4 in [93], and 93% of these genes are mechanistically linked in our ME-Model.

The reactions associated with transcription and translation, including initiation, biopolymerization and termination, were generated from the genome sequence and a set of *T. maritima* template reactions (Supplementary Methods in [93]). In our modelling formalism, reversible reactions were represented as two unique reactions: one for the forward direction and one for the reverse.

3.5.2 Protein complexes

For each functional protein, we used primary literature and the RCSB Protein Data Bank [145] to determine whether the machine was a monomer or oligomer. The Protein Data Bank entries provided an opportunity to integrate 3D structural data into our reconstruction (this model includes structures for 32 additional open reading frames compared with Zhang et al.). When data for multimeric state were unavailable for a protein of interest, state data for orthologs from closely related organisms were used; otherwise, the Uniprot database [144] was consulted. In the absence of data providing insight into the multimeric state of the protein, we assumed that the functional protein was a monomer.

3.5.3 Genetic code determination

From inspection of tRNA sequences and structures downloaded from the transfer RNA database [146], we determined that *T. maritima* uses uniform-GUC decoding with only 46 tRNA genes (see Supplementary Data 5 in [93]). In both Archaea and Bacteria, but not in Eukarya, the conversion of C34 of a CAU-anticodon to lysidine (k2C) or analogue generates an anticodon for isoleucine [147]. TMtRNA-Met-2 was assigned this role based on a strong sequence alignment to *E. coli* tRNAs containing k2C. The *T. maritima* genome encodes two additional tRNA genes with CAU anticodons, TMtRNA-Met-1 and TMtRNA-Met-3. Based on structural similarity [148] to those found in a crystal structure of E. coli's formyl-methionyl-tRNAfMet55, TMtRNA-Met-1 may be involved in translation initiation, therefore, TMtRNA-Met-3 was designated to participate in translation elongation.

3.5.4 TU architecture determination

We assembled a draft TU architecture (Supplementary Data 6 in [93]) for *T. maritima* based on a series of rules (Supplementary Methods in [93]). In short, we assumed all TUs start with a gene start and proceed until one of the following conditions is met: (1) two genes are found in convergent orientation on different strands, (2) two genes are found in divergent orientation on different strands, (3) a high-confidence Rho-independent transcription terminator is found separating two genes oriented in series on the same strand, (4) more than 55 base pairs separate two genes in series on the same strand or (5) experimental evidence indicates a TU boundary. Finally, to reflect the possibility of internal transcription start sites in TUs reconstructed using the rules above, we added an additional TU in cases where a high-confidence promoter was found in the region separating two genes oriented in series on the same strand.

3.5.5 *In silico* molecular biology

Log-phase growth simulations were performed using FBA [121]. Linear programming was used to identify the maximum μ or minimum ribosome production flux supporting a particular μ from the components of the *in silico* minimal media. Because of the presence of fast (metabolic) and slow (macromolecular synthesis) timescale reactions, the parameters in the ME-Model span a wide range that can result in inaccurate simulations due to floating point limitations of currently available floating point linear programming software (Supplementary Methods in [93]). To remove the possibility of simulation results being artefacts arising from floating point limitations, we used the exact simplex routines available in the QSopt_ex package [122], with default parameter settings for ME-Model simulations. The predicted transcription level of a gene was determined by summing across the sink fluxes of TUs containing the gene, which is equivalent to the transcription fluxes less the TU degradation fluxes. Translation levels were reported as the sum across the relevant translation initiation fluxes, as many TUs can contribute to the production of a given protein. These values were compared with each other in the case of simulated nutrient shifts or to the abundances reported experimentally.

3.5.6 *In vivo* methods

T. maritima MSB8 (ATCC: 43589) was grown in 500 ml serum bottles containing 200 ml of anoxic minimal media with 10 mM maltose, L-arabinose or cellobiose as the sole carbon source at 80 °C. All samples were collected during log-phase growth. Substrate uptake and by-product secretion rates, compositional analyses, and transcriptome and proteome measurements were performed as described in the Supplementary Methods in [93]. Transcriptome data have been submitted to the NCBI Gene Expression Omnibus (accession ID: GSE28822) and processed values are in Supplementary Data 7 in [93]. Proteomics data are available through Pacific Northwest National Laboratory (http://omics.pnl.gov) and processed values are in Supplementary Data 8 in [93].

3.5.7 RNA modifications

A variety of post-transcriptional modifications of rRNAs are represented in our model. For 16S rRNA, there was experimental evidence for ten modifications [149] in this organism (Supplementary Table S4 in [93]). The locations of pseudouridines, which are mass silent, were not available, but an 11th modification, U to Y at position 516, was included in the reconstruction based on the fact that it is well conserved in bacteria and the alignment (Supplementary Data 9 in [93]) supports its inclusion. An unusual derivative of cytidine-designated N-330 has been sequenced to position 1,404 [149] in the decoding region of the 16S rRNA. This modified nucleoside was excluded from the reconstruction as the exact chemical composition of the modification is unknown. We were unable to find organism-specific literature supporting modifications to the 5S and the 23S rRNA. Modifications to 5S rRNA are infrequent in bacteria [150]. Attempting to extrapolate 23S rRNA modifications from *E. coli* was relatively unsuccessful as alignment via ClustalW2 [151] showed significant differences near many of the putative modification sites (Supplementary Data 10 in [93]). The alignment reveals that the 23S rRNA of *T. maritima* is significantly longer (>100 bp) than that of *E. coli.* Only three proteins with annotated roles in modifying the 23S rRNA were added to the model for a total of six modifications (Supplementary Table S5 in [93]). Those were TM0940, TM0462 and TM1715.

Post-transcriptional modification of tRNA also requires a significant investment in genes, enzymes, substrates and energy [152]. We included a variety of modifications (Supplementary Table S6 in [93]) in our model based on bioinformatics predictions and literature evidence (Supplementary Table S7 in [93]).

3.5.8 Sensitivity analysis

To explore the influence of some of the newly introduced parameters on model output, the bulk parameters used for the coupling constraints (Supplementary Methods in [93]) were varied (two-, four- and eight-fold increases and decreases away from the parameter set used). The results are summarized in Supplementary Fig. S4 in [93].

3.5.9 File formats

Our final model is available as a Systems Biology Markup Language (SBML) XML file (Supplementary Data 11 in [93]). The model is also available as an LP file (Supplementary Data 12 in [93]) for use with linear programming solvers.

3.5.10 Accession codes

Transcriptome data have been submitted to the NCBI Gene Expression Omnibus under accession code GSE28822.

3.6 Acknowledgements

We thank Jan Schellenberger, Daniel Espinoza, Bill Cook and Michael Saunders for invigorating discussions on solving stiff LPs. Heather Mottaz-Brewer for assistance in proteome sample processing. This work was supported in part by the US National Institute of Allergy and Infectious Diseases and the US Department of Health and Human Services through interagency agreement Y1-AI-8401-01, DOE Awards DE-FG02-09ER25917 and DE-FG02-08ER64686. Proteomic analyses were performed in the Environmental Molecular Sciences Laboratory, a US DOE BER national scientific user facility at Pacific Northwest National Laboratory. D.R.H. is supported in part by a Seed Award from the San Diego Center for Systems Biology funded by NIH/NIGMS (GM085764).

Experiments and simulations were conceived and designed by J.A.L. and D.R.H. J.A.L. and J.O. led the network reconstruction. Transcriptomics experiments were performed by H.L. and V.A.P. Proteomics data were generated by A.C.S.-R. and J.N.A. Peptides were called and mapped by A.C.S.-R., J.N.A. and R.D.S. Data were normalized by J.A.L and D.R.H. The manuscript was written by J.A.L. and D.R.H. with input from H.L., N.E.L., K.Z. and B.O.P.

Chapter 3 in full is a reprint of a published manuscript: Lerman, J.A. *et al.* *In silico* method for modelling metabolism and gene product expression at genome scale. *Nat. Commun.* 3:929 doi: 10.1038/ncomms1928 (2012). The dissertation author was the primary author of this paper responsible for the research. The other authors were Daniel R. Hyduke (equal contributor), Haythem Latif, Vasiliy A. Portnoy, Nathan E. Lewis, Jeffrey D. Orth, Alexandra C. Schrimpe-Rutledge, Richard D. Smith, Joshua N. Adkins, Karsten Zengler, and Bernhard Ø. Palsson.

Chapter 4

Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction

4.1 Abstract

Growth is a fundamental process of life. Growth requirements are well-characterized experimentally for many microbes; however, we lack a unified model for cellular growth. Such a model must be predictive of events at the molecular scale and capable of explaining the high-level behavior of the cell as a whole. Here, we construct an ME-Model for *Escherichia coli*—a genome-scale model that seamlessly integrates metabolic and gene product expression pathways. The model computes ∼80% of the functional proteome (by mass), which is used by the cell to support growth under a given condition. Metabolism and gene expression are interdependent processes that affect and constrain each other. We formalize these constraints and apply the principle of growth optimization to enable the accurate prediction of multi-scale phenotypes, ranging from coarse-grained (growth rate, nutrient uptake, by-product secretion) to fine-grained (metabolic fluxes, gene ex-

pression levels). Our results unify many existing principles developed to describe bacterial growth.

4.2 Introduction

The genotype-phenotype relationship is fundamental to biology. Historically, and still for most phenotypic traits, this relationship is described through qualitative arguments based on observations or through statistical correlations. Understanding the genotype-phenotype relationship demands vantage points at multiple scales, ranging from the molecular to the cellular. Reductionist approaches to biology have produced 'parts lists', and successfully identified key concepts (e.g., central dogma) and specific chemical interactions and transformations (e.g., metabolic reactions) fundamental to life. However, reductionist viewpoints, by definition, do not provide a coherent understanding of whole cell functions. For this reason, modeling whole biological systems (or subsystems) has received increased attention.

A number of modeling approaches have been developed to predict systems-level phenotypes. What distinguish these models from each other are the underlying assumptions they make, the input data they require, and the scope and precision of their predictions [153]. The type of modeling formalism employed is influenced by all of these distinguishing characteristics [154]. Genome-scale optimality models of metabolism (termed as M-Models) have made much progress in recent years as they require only basic knowledge of reaction stoichiometry, are genome-scale in scope, and have fairly accurate predictive power. Recently, M-Models have been extended to include the process of gene expression (termed as ME-Models) [93, 155], opening up completely new vistas in the development of microbial systems biology. On the heels of these developments, a whole-cell model (WCM) of the human pathogen *Mycoplasma genitalium* appeared [156]. The WCM integrates many more cellular processes and can be used to simulate dynamic cellular states; however, it depends on detailed molecular measurements of an initial state (e.g., growth rate, biomass composition, and gene expression). While the model described by Karr et al is a major advance toward whole-cell computation, many practical applications rely on the ability to compute optimal phenotypic states. The WCM does not have this ability owing to the disparate mathematical formalisms it employs. The WCM and genome-scale optimality models thus have different capabilities and will find use to predict and explain different biological phenomena.

Here, we construct an ME-Model for *E. coli* K-12 MG1655. The ME-Model is a microbial growth model that computes the optimal cellular state for growth in a given steady-state environment. It takes as input the availability of nutrients to the cell and produces experimentally testable predictions for: (1) the cell's maximum growth rate (μ*) in the specified environment, (2) substrate uptake/by-product secretion rates at μ*, (3) metabolic fluxes at μ*, and (4) gene product expression levels at μ*. The creation of this model required the development of a new modeling formalism and optimization procedure to couple gene expression with metabolism, which provided new insight into growth rate- and nutrient limitation-dependent changes in enzymatic efficiency. The model predicts three distinct regions of microbial growth, defined by the factors (nutrient and/or proteome) limiting growth. We show that proteomic constraints improve predictions of metabolism itself, rectifying dominant failure modes in M-Models. Finally, we compute gene expression changes as the cell transitions through and between the different growth regions. The ME-Model computes measurable coarse- and fine-grained cellular and molecular phenotypes, and provides unity in the field by reconciling a variety of principles related to cellular growth at various scales of complexity.

4.3 Results

4.3.1 Integration of genome-scale reaction networks of protein synthesis and metabolism

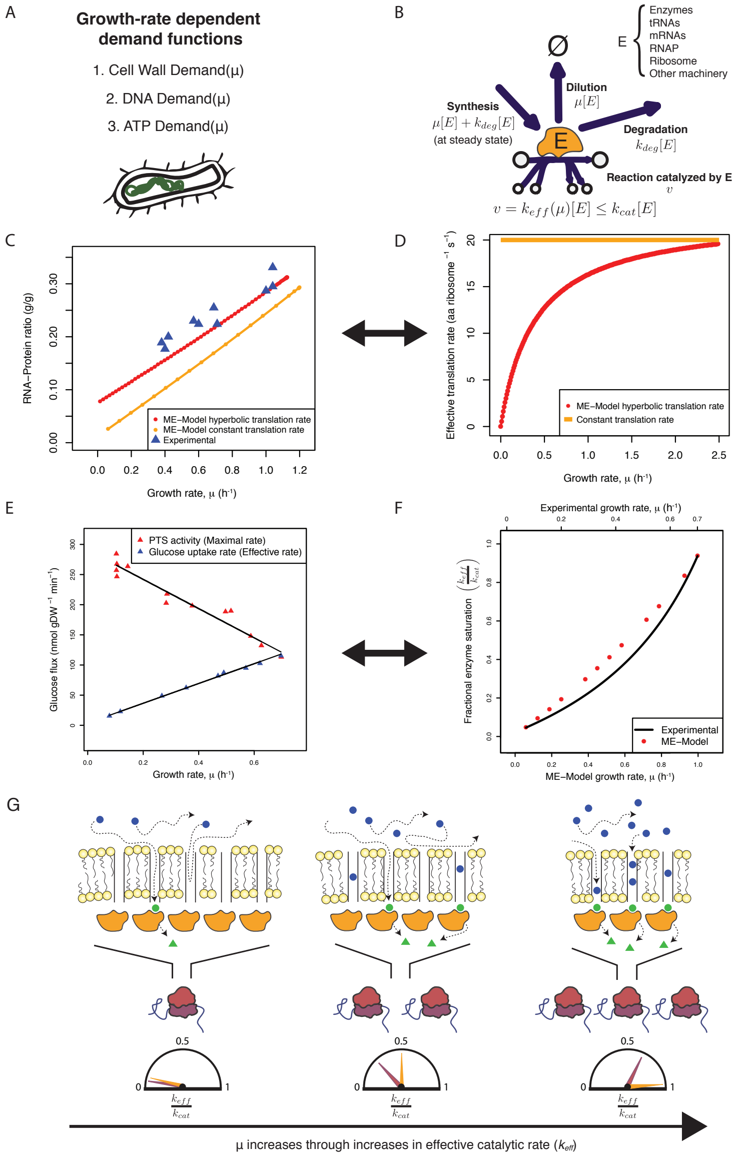
To create an ME-Model for *E. coli*, we started from two previous network reconstructions. The first reaction network includes all known metabolic pathways as of late 2011 [157] and is referred to as the M-Model throughout. The second accounts for reactions that describe gene expression and the synthesis of functional macromolecules in a mechanistically detailed manner [124]. The two reaction networks were integrated (see Materials and methods), and reactions and gene functions in both networks were updated to reflect gaps in knowledge that have been filled since their creation. We updated subunit stoichiometries for hundreds of multiprotein complexes and expanded the types of prosthetic groups, cofactors, and post-translational modifications required for catalytic activity (Materials and methods; Supplementary Table S1 in [158]). The scope and coverage of cellular processes in the integrated network is extensive. The integrated network mechanistically links the functions of 1541 unique protein-coding open reading frames (ORFs) and 109 RNA genes; it thus accounts for ∼35% of the 4420 protein-coding ORFs, ∼65% of the functionally well-annotated ORFs [159], and 53.7% of the non-coding RNA genes identified in *E. coli* K-12 [160]. In total, 1295 unique functional protein complexes are produced. Taken together, these complexes account for 80-90% of *E. coli*'s expressed proteome by mass (Supplementary Table S2 in [158]).

The integrated reaction network covers and accurately predicts a large proportion of essential cellular functions. It includes 223 of the 302 (73.8%) genes classified as essential for cell growth under any condition [161] (Supplementary Table S3A in [158]), and 166 of the 206 functions (80.6%) estimated as essential for a minimal organism [128] (Supplementary Table S3B in [158]). *In silico* prediction of gene essentiality in glucose M9 minimal media results in an accuracy of 88.8% (precision=60.4%, recall=75%, Supplementary Table S4 in [158]). One of the dominant failure modes of essentiality predictions is due to the assumption that all tRNA and rRNA modifications are essential; removing these genes from predictions increases performance notably (accuracy=92.3%, precision=75.3%, recall=75%, Supplementary Table S4 in [158]). This accuracy is on par with previous approaches using the metabolic reaction network alone (accuracy=91.2%, precision=81%, recall=68%) [157]. Many of the key differences between the M-Model and the ME-Model essentiality predictions are due to the mechanistic treatment of cofactor and prosthetic group synthesis and utilization in the ME-Model. Specifically, for a protein complex to be functional in the ME-Model it has to contain the embedded prosthetic groups required for function; while this change in model structure results in some false predictions of essentiality compared with M-Models (which include all prosthetic groups in a biomass objective function that does not change across conditions), the essentiality predictions in the ME-Model can be directly related to the essential enzymes requiring the prosthetic group.

4.3.2 Growth demands and general constraints on molecular catalysis

To compute functional states of the integrated network, growth demands are first imposed. Growth requires the replication of the organism's genome and synthesis of a new cell wall to contain the replicated DNA. In the ME-Model, growth rate-dependent DNA and cell wall demand functions formalize these requirements (Figure 4.1A; Supplementary information in [158]). We derived these demand functions from growth rate-dependent trends in cell size [162] and DNA content [163, 164] (Supplementary information in [158]). In addition, as in previous models, we imposed growth-associated and non-growth-associated ATP utilization demands [9] as the ostensible energy requirements [165, 166].

Figure 4.1: **Growth demands and coupling constraints leading to growth rate-dependent changes in enzyme and ribosome efficiency.** (A) Three growth rate-dependent demand functions derived from empirical observations determine the basic requirements for cell replication (detailed in Supplementary information in [158]). (B) Coupling constraints link gene expression to metabolism through the dependence of reaction fluxes on enzyme concentrations. (C, D) RNA:protein ratio predicted by the ME-Model with two different coupling constraint scenarios, one for variable translation rate versus growth rate (red lines) and one for constant translation rate (orange lines). Experimental data in (C) obtained from Scott et al (2010). (E) Phosphotransferase system (PTS) transient activity following a glucose pulse in a glucose-limited chemostat culture (red) and glucose uptake before the glucose pulse (blue) is plotted as a function of growth rate. The data shown were obtained from O'Brien et al (1980)). Data from $\mu > 0.7$ h^{-1} were omitted. (F) Data from (E) are used to plot glucose uptake as a fraction of PTS activity. The resulting value is the fractional enzyme saturation (black line). The fractional enzyme saturation predicted by the ME-Model is plotted as a function of growth rate under carbon limitation (red dots). (G) The cartoon depicts changes in extra- (blue) and intra- (green) cellular substrate (circle) and product (triangle) concentrations and metabolic enzyme (orange) and ribosome (purple/maroon) levels as the concentration of a growth-limiting nutrient (and growth rate) increases. The dials show k_{eff}/k_{cat}, the effective catalytic rate over the maximum for metabolic enzymes (orange) and ribosomes (purple/maroon).


A
Growth-rate dependent demand functions
1. Cell Wall Demand(μ)
2. DNA Demand(μ)
3. ATP Demand(μ)
B
Ø
E
Enzymes
tRNAs
mRNAs
RNAP
Ribosome
Other machinery
Dilution
$\mu[E]$
Synthesis
$\mu[E] + k_{deg}[E]$
(at steady state)
Degradation
$k_{deg}[E]$
E
Reaction catalyzed by E
v
$v = k_{eff}(\mu)[E] \leq k_{cat}[E]$
C
RNA–Protein ratio (g/g)
0.00
0.10
0.20
0.30
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Growth rate, μ (h^{-1})
ME–Model hyperbolic translation rate
ME–Model constant translation rate
Experimental
D
Effective translation rate (aa ribosome $^{-1}$ s^{-1})
0
5
10
15
20
0.0
0.5
1.0
1.5
2.0
2.5
Growth rate, μ (h^{-1})
ME–Model hyperbolic translation rate
Constant translation rate
E
Glucose flux (nmol gDW $^{-1}$ min^{-1})
0
50
100
150
200
250
300
0.0
0.2
0.4
0.6
Growth rate, μ (h^{-1})
PTS activity (Maximal rate)
Glucose uptake rate (Effective rate)
F
Experimental growth rate, μ (h^{-1})
0
0.2
0.4
0.6
0.7
Fractional enzyme saturation $\left(\frac{k_{eff}}{k_{cat}}\right)$
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ME–Model growth rate, μ (h^{-1})
Experimental
ME–Model
G
0.5
0
1
$\frac{k_{eff}}{k_{cat}}$
0.5
0
1
$\frac{k_{eff}}{k_{cat}}$
0.5
0
1
$\frac{k_{eff}}{k_{cat}}$
μ increases through increases in effective catalytic rate (k_{eff})

One large improvement is that RNA and protein are not included as demand functions (as they are in M-Models; [120]; instead, expression of specific RNA and protein molecules are free variables determined during ME-Model simulations. 'Coupling constraints' [167, 93] relate the synthesis of RNA- and protein-based molecules to their catalytic functions in the cell (Figure 4.1B). The coupling constraints are based on parameters that define the effective catalytic rate (k_{eff}) and degradation rate constant (k_{deg}) of molecular machines (Supplementary information in [158]).

A nutritional environment is then defined by setting constraints on the availability and uptake of nutrients. For a particular nutritional environment, there is a maximum growth rate at which the cell can no longer produce enough RNA and protein machinery to meet the demands of growth. The computed cellular state (biomass composition, substrate uptake and by-product secretion, metabolic flux, and gene expression) at this maximum growth rate is the predicted optimal response of the cell to the specified nutritional environment.

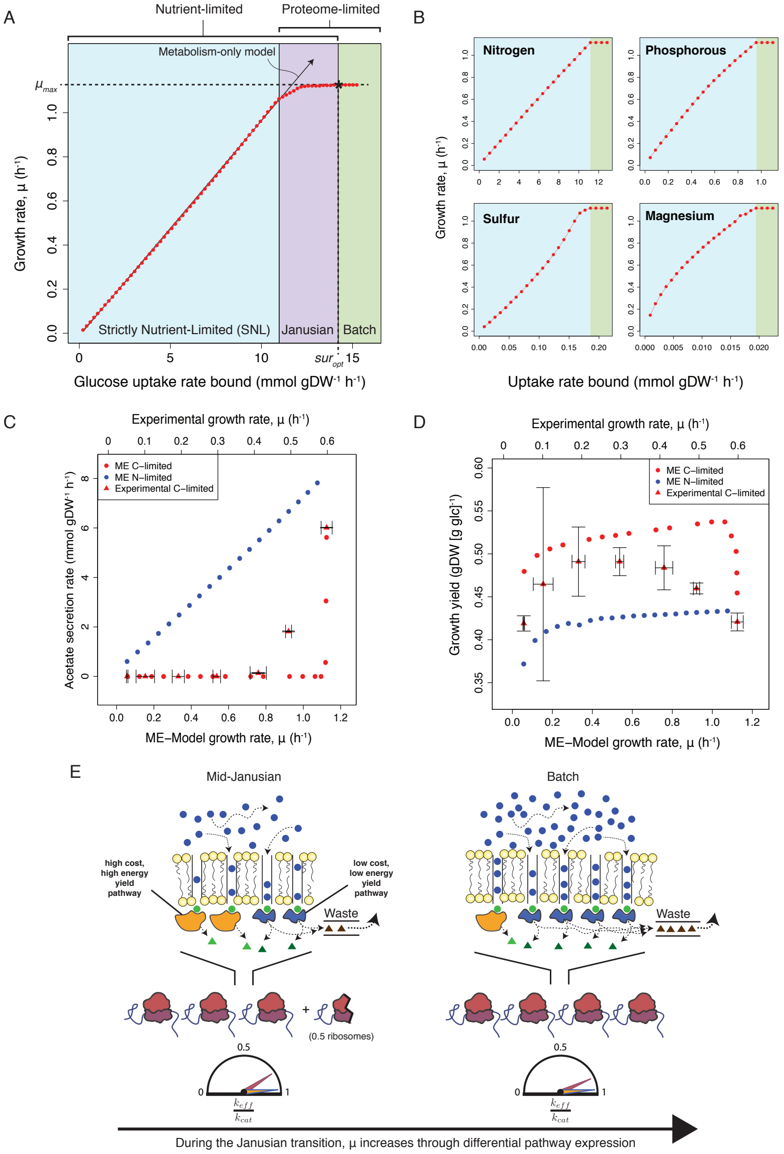
4.3.3 Derivation of constraints on molecular catalytic rates

Previous studies disagree as to if ribosomes translate with the same efficiency (amino acids per ribosome per second) across growth conditions [168, 11]. Here, we use the ME-Model and available data to determine an appropriate constraint for ribosomal efficiency as a function of growth rate. We find that if a constant translation rate of 20 amino acids per second is imposed as a constraint in the ME-Model, the model predicts a linear growth rate-dependent RNA-to-protein ratio (Figure 4.1C), consistent with the previous measurements [11]; however, the predicted RNA content does not quantitatively match measured values. In particular, a constant translation rate results in no RNA production in the limit of no growth. We therefore hypothesized that ribosomal translation rate systematically varies with growth rate, and back-calculated a growth rate-dependent translation rate using measured growth rate-dependent RNA content (Supplementary information in [158]). Ultimately, we recovered a Michaelis-Menten-type rate law (Figure 4.1D) with a maximal rate (Vmax) of ∼20 amino acids per second, consistent with previous findings for maximal ribosomal speed [164]; the rate law results in a quantitative match of RNA content compared with experimental data (Figure 4.1C, Pearson's r=0.96). This rate law causes translation efficiency to increase under nutrient-richer conditions, which recent experimental evidence supports [169, 170]. Interestingly, when we applied the same Michaelis-Menten-type equations to constrain tRNA and mRNA catalytic rates, we recovered maximal turnover rates highly consistent with previous estimates (Supplementary information in [158]). The catalytic rates of metabolic enzymes are variable as well, and tend to decrease when nutrients are limited. Both metabolomics [171] and proteomics [170] data sets suggest a large-scale scaling of enzyme efficiencies under nutrient limitation. We approximate these changes in metabolic catalysis in the ME-Model with two minimal assumptions: (1) when the cell is nutrient-limited, protein content is maximized (at a given growth rate) and (2) this protein content specifically is metabolic enzymes not operating at their maximal catalytic rate [170] (i.e., $k_{eff}/k_{cat} < 1$, see Figure 4.1G and Supplementary information, Optimization procedure in [158]). These two assumptions allow us to predict average catalytic rates of metabolic enzymes under nutrient limitation. The nutrient limitation-dependent shape of our computed catalytic rates matches assays for glucose transporters under glucose limitation [172] (Figures 4.1E and F), LacZ under lactose limitation [173] Supplementary Figure S1A in [158]), and the enzyme efficiency in a small-scale optimality model accounting for substrate concentrations with Michaelis-Menten kinetics [174] (Supplementary Figure S1B in [158]). However, because the current ME-Model simulation procedure assumes that k_{eff} decreases uniformly across metabolism, the model does not capture the importance of specific enzymes for particular nutrient limitations; recent data sets [170] and kinetic models [14] can help us understand and model these trends better at the genome-scale.

4.3.4 Growth regions under varying nutrient availability

Upon derivation of the growth demands and molecular efficiencies, we investigate high-level model behavior to variable nutrient availability. Unlike previous genome-scale models [157, 155], growth rate in the ME-Model is a non-linear function of the substrate uptake rate bound (Figure 4.2A), and eventually reaches a maximum. This behavior is consistent with long-standing empirical models of microbial growth [175, 176], in which growth is first nutrient-limited, but then limited by some intra-organismal bound.

Figure 4.2: **Predicted growth, yield, and secretion.** (A) Predicted growth rate is plotted as a function of the glucose uptake rate bound imposed in glucose minimal media. Three regions of growth are labeled Strictly Nutrient-Limited (SNL), Janusian, and Batch (i.e., excess of substrate) based on the dominant active constraints (nutrient and/or proteome limitation). The proteome-activity constraint inherent in the ME-Model results in a maximal growth rate and substrate uptake rate. The behavior of a genome-scale metabolic model (M-Model) is depicted with an arrow. (B) Predicted growth rates as a function of uptake of a limiting nutrient with glucose in excess. The shaded regions correspond to those as labeled in (A). (C) Experimental (triangle) and ME-Model-predicted (circle) acetate secretion in Nitrogen- (blue) and Carbon- (red) limited glucose minimal medium are plotted as a function of growth rate. Data were obtained from Zhuang et al (2011). The root-mean-square error (RMSE) between data and the ME-Model is 0.12 (for comparison, RMSE=0.40 for the M-Model). (D) Experimental (triangle) and ME-Model-predicted (circle) carbon yield (gDW Biomass/g Glucose) in Carbon- (red) and Nitrogen- (blue) limited glucose minimal medium are plotted as a function of growth rate. Data were obtained from Zhuang et al (2011). RMSE between data and the ME-Model is 0.04 (for comparison, RMSE=0.07 for the M-Model). (E) The cartoon depicts changes in extra- (blue) and intra- (green) cellular substrate (circle) and product (triangle) concentrations and metabolic enzyme (blue/orange) and ribosome (purple/maroon) levels during the Janusian region. Metabolic enzymes are saturated throughout the entire Janusian region. To increase the growth rate, the cell expresses metabolic pathways that have lower operating costs. (Pathways with the smaller blue proteins taken to be 0.25 the cost of the pathways with larger orange proteins.) A higher glucose uptake and turnover results, but energy yield is lower and some carbon is 'wasted' and secreted (brown triangles). The dials show k_{eff}/k_{cat}, the effective catalytic rate over the maximum for metabolic enzymes (blue/orange) and ribosomes (purple/maroon).


A
Nutrient-limited
Proteome-limited
Metabolism-only model
μ_{max}
Growth rate, μ (h⁻¹)
0.0
0.2
0.4
0.6
0.8
1.0
Strictly Nutrient-Limited (SNL)
Janusian
Batch
0
5
10
sur_{opt}
15
Glucose uptake rate bound (mmol gDW⁻¹ h⁻¹)
B
Nitrogen
Phosphorous
Sulfur
Magnesium
Growth rate, μ (h⁻¹)
Uptake rate bound (mmol gDW⁻¹ h⁻¹)
C
Experimental growth rate, μ (h⁻¹)
ME C–limited
ME N–limited
Experimental C–limited
Acetate secretion rate (mmol gDW⁻¹ h⁻¹)
ME–Model growth rate, μ (h⁻¹)
D
Experimental growth rate, μ (h⁻¹)
ME C–limited
ME N–limited
Experimental C–limited
Growth yield (gDW [g glc]⁻¹)
ME–Model growth rate, μ (h⁻¹)
E
Mid-Janusian
Batch
high cost, high energy yield pathway
low cost, low energy yield pathway
Waste
Waste
(0.5 ribosomes)
0.5
0
1
$\frac{k_{eff}}{k_{cat}}$
0.5
0
1
$\frac{k_{eff}}{k_{cat}}$
During the Janusian transition, μ increases through differential pathway expression

Under nutrient-excess conditions, growth in the ME-Model is limited by internal constraints on protein production and catalysis—the cell is ‘proteome-limited’—resulting in a corresponding maximal growth rate (Figure 4.2A). This feature allows Batch culture growth to be simulated without specifying nutrient uptake bounds; instead, the ME-Model predicts a maximum batch growth rate and optimal substrate uptake rate.

Supporting the validity of the proteomic constraints limiting growth in Batch culture, optimal Batch growth rates, substrate uptake rates, and biomass yields correlate with experimental data for growth on different carbon sources (Supplementary Table S5 in [158]). The ME-Model predicted substrate uptake and biomass yield closely matches laboratory evolved strains (Pearson’s r=0.89 and r=0.91, respectively) (Supplementary Table S5C in [158], sensitivity analysis in Supplementary Table S6 in [158]). Though less accurate, predicted growth rates by the ME-Model correlate with measured growth rates in batch culture better than standard M-Models, in which growth rate is maximized subject to a specified nutrient uptake, and the correlation increases when compared with laboratory evolved strains (M-Model Pearson’s r=0.49, ME-Model Pearson’s r=0.61) as opposed to wild-type strains (M-Model Pearson’s r=0.30, ME-Model Pearson’s r=0.39). Other methods that include various approximate constraints on the total flux through the metabolic network also show an increased performance in growth rate prediction, though all computational methods [177, 178] still correlate better with each other than with the experimental data (Supplementary Table S5B in [158]).

When the uptake of glucose is restricted below the amount required for optimal growth in batch culture, the cell’s growth is carbon-limited. Growth rate linearly increases with glucose uptake when glucose availability is low. In this region (termed as the Strictly Nutrient-Limited (SNL) region in Figure 4.2A), the capabilities of the proteome are not fully utilized as the proteome could process more incoming glucose if it was available (Figures 4.1E-G). By varying the glucose availability, we find that a region exists in which the cell is both nutrient- and proteome- limited; we refer to this transition region as the Janusian region [179].

ME-Model computations thus reveal three distinct regions of microbial growth (Figure 4.2A; see Supplementary information, Optimization procedure, Computational definition, and identification of growth regions in [158]).

When the uptake of non-carbon sources is restricted below the amount required for optimal growth in batch culture, the cell's growth is limited by that nutrient. Unlike carbon-source limitation, we find the nutrient- and proteome-limited regions to be distinct (Figure 4.2B). However, in the SNL region, growth is sometimes non-linear as a function of uptake rate, due to changing biomass requirements (e.g., Sulfur and Magnesium).

4.3.5 Effect of proteome limitation on secretion phenotypes

To understand the proteome-limited growth regions in the ME-Model, we first investigate trends in secretion phenotypes and biomass yield. Under glucose limitation, different metabolic pathways are utilized in the Janusian region than in the SNL region, resulting in acetate secretion (Figure 4.2C, red). This metabolic switch, combined with growth rate-dependent ATP requirements, results in a concave biomass yield as a function of growth rate (Figure 4.2D, red). Both the biomass yield and secretion trends have repeatedly been experimentally observed [166]. The example of glucose limitation provides an illustrative example for the general behavior in the Janusian growth region. In the Janusian region, the cell increases its growth rate through differential expression of pathways, as illustrated in Figure 4.2E. Due to proteome limitations, the cell switches to pathways that require less protein mass but are lower in nutrient yield (defined as energy and/or biomass precursors produced per molecule of limiting nutrient consumed). This behavior is in contrast to that in the SNL region, in which high-yield pathways are optimal (as in M-Models) and growth rate increases through changes in the effective catalytic rate of metabolic enzymes (Figure 4.1G). These results provide further support that 'overflow' metabolism can be understood in terms of proteomic constraints, as suggested with a small-scale model [174].

The ME-Model also predicts that acetate will be secreted at all growth rates when *E. coli* is Nitrogen (Ammonium)-limited (Figure 4.2C, blue). Experimentally, acetate is secreted under nitrogen limitation even at low growth rates [180]. This secretion phenotype is explained by the ME-Model as follows: protein 'saved' by utilizing low-yield carbon metabolism is diverted to synthesize other enzymes that are not operating at their maximal catalytic capacity.

No Janusian region is observed under non-carbon limitation. In the ME-Model, this is likely due to reaction network topology—while there are many alternative pathways for energy, redox, and biomass precursor generation in carbon metabolism, non-carbon nutrient assimilation is often achieved using more linear pathways. As a result, there are fewer opportunities for trade-offs between uptake rate and biomass yield. However, perhaps including variable substrate affinities for alternative pathways would reveal Janusian regions corresponding to non-carbon limitations.

4.3.6 Central carbon fluxes reflect growth optimization subject to catalytic constraints

Further supporting the importance of proteomic constraints on metabolic phenotypes is the prediction of central carbon fluxes by the ME-Model. When glucose availability is varied, the ME-Model predicts changes in central carbon metabolism consistent with the changes from ^{13}C fluxomic data sets (Figure 4.3; Supplementary Figure S2 in [158], Pearson's r=0.93, 0.90, 0.86) [181, 4, 5]. Importantly, the ME-Model predicts the dominant changes in pathway splits as the glucose availability is varied (Figure 4.3, insets).

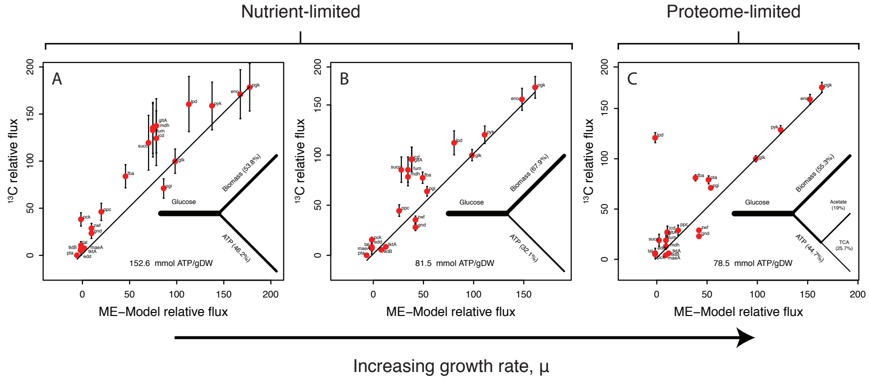


Figure 4.3: **Central carbon metabolic flux patterns under glucose-limited and glucose-excess conditions.** (A-C) Relative fluxes from ^{13}C experiments are plotted versus the fluxes predicted by the ME-Model. (A, B) Comparison of nutrient-limited model solutions with chemostat culture conditions and (C) comparison of the batch ME-Model solution with batch culture data. All simulations and experiments correspond to growth in glucose minimal media. Fluxes are normalized so that glucose uptake is 100. Insets show the main flux changes under increasing glucose concentrations. The only model parameter that is modulated is the glucose uptake rate bound. Data were obtained from Nanchen et al (2006) and Schuetz et al (2007). The ME-Model flux for the reaction 'pyk' is taken to include phosphoenolpyruvate (PEP) to pyruvate (PYR) conversion via the phosphotransferase system (PTS). Flux splits shown as insets were computed using the ME-Model. The percentages indicate the percent carbon (Glucose) converted to CO_2 (for branch labeled 'TCA'), acetate, and biomass. Both the TCA and acetate branches contribute to ATP production. The total mmol ATP per gDW biomass produced is indicated.

Previous studies have evaluated the ability of M-Models together with assumed optimality principles to predict metabolic fluxes [4, 5]. These studies concluded that no single objective function applied to M-Models can accurately represent fluxomic data from all environmental conditions studied [4]. Instead, metabolic fluxes can be understood as being Pareto optimal: multiple objectives are simultaneously optimized and their relative importance varies depending on the environmental condition [5]. The three objectives needed to explain most of the variations in the data from Schuetz et al were (1) maximum ATP yield, (2) maximum biomass yield, and (3) minimum sum of absolute fluxes (which is a proxy for minimum enzyme investment). These three objectives formed a Pareto optimal surface that was valuable for interpreting fluxomic data; however, the surface was large and it was not possible to predict the importance of each of the objectives a priori.

By explicitly accounting for variable growth demands, enzyme expression, and constraints on enzymatic activity, the ME-Model eliminates the need for multiple objectives; growth rate optimization alone is sufficient to predict the fluxes through central carbon metabolism (Figure 4.3; Supplementary Figure S2 in [158]; Supplementary Table S7 in [158]). The three original objectives chosen by Schuetz et al are biologically meaningful dimensions and required for interpreting fluxomic data when using an M-Model. In contrast, the ME-Model accounts for all three of these dimensions implicitly during growth rate maximization without adjusting any model parameters (see Supplementary information in [158] and Supplementary Table S7 in [158]). Accordingly, ME-Models can determine, at least qualitatively, the importance and weighting of the objectives for growth in a given environment. Ultimately, the primary changes in flux through central carbon metabolism can be understood as responses to the same constraints causing the observed relationship in biomass yield (Figure 4.2D): at low growth rates under carbon limitation, the dominant changes are due to a changing ATP demand, and in the transition from carbon-limited to carbon-excess (proteome-limited) conditions, the primary changes are due to the switch to lower yield carbon catabolism (Figure 4.3, insets).

4.3.7 *In silico* gene expression profiling from nutrient-limited to batch growth conditions

We now use the ME-Model to predict groups of proteins that change in expression under various degrees of glucose limitation. Under glucose limitation, the optimal proteome changes due to shifting growth demands and proteomic constraints. The groups of functionally related proteins that shift in our simulations match those previously reported experimentally [182, 183], but the model predictions of quantitative differential expression (at the level of single genes) are weak. We separate the analysis of the SNL region (Figure 4.4; Supplementary Table S8A in [158]) from the Janusian region (Figure 4.5; Supplementary Table S8B in [158]), due to the different dominant constraints and phenotypic responses specific to each region.

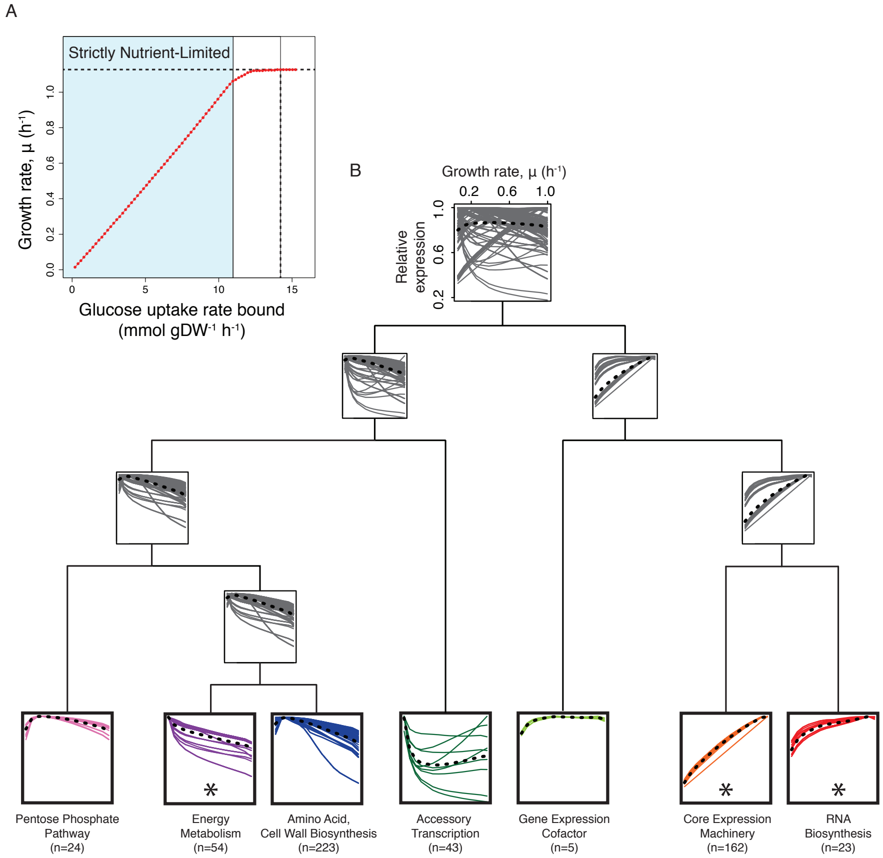


Figure 4.4: **Growth rate-dependent gene expression under glucose limitation.** (A) Gene expression changes predicted by the ME-Model to occur in the Strictly Nutrient-Limited (SNL) growth region indicated in light blue under glucose limitation in minimal media are analyzed. (B) ME-Model-computed relative gene-enzyme pair expression is plotted as a function of growth rate; the normalized *in silico* expression profiles are clustered hierarchically (see Materials and methods). Solid lines are expression profiles of individual gene-enzyme pairs and dotted black lines are the centroid of each cluster. Each leaf node is colored and qualitatively labeled by function. The number of genes in each leaf node is indicated and listed in Supplementary Table S8A in [158]. Asterisks indicate clusters with monotonic expression changes that significantly match the directionality observed in expression data (Wilcoxon signed-rank test, $P < 1 \times 10^{-4}$). Expression data were obtained from a previous study [183], in which *E. coli* was cultivated in a chemostat at dilution rates $0.3\,h^{-1}$ and $\sim 0.5\,h^{-1}$.

In the SNL region, the expression of most proteins decreases as growth rate increases (Figure 4.4B, left side of tree, Supplementary Figure S3 in [158]). The largest group of proteins includes those responsible for amino-acid and cell wall synthesis; the growth rate-dependent decrease in expression of these proteins is due to the combined effects of a decrease in cell wall and protein biomass (g/gDW) and an increase in the effective catalytic rate of enzymes (Figures 4.1E-G). Proteins involved in energy metabolism also decrease in expression with increasing growth rate due to changes in catalytic rate and growth rate-dependent demands. Surprisingly, the predicted expression levels of several accessory transcription proteins, including four stress-associated sigma factors (RpoS, RpoH, RpoE, and RpoN), are elevated at very low growth rates, reflecting an association with metabolic proteins needed for slow growth.

A smaller number of proteins show increases in their relative expression levels at higher growth rates (Figure 4.4B, right side of tree, Supplementary Figure S3 in [158]). These proteins include those responsible for protein synthesis (ribosome, RNAP, and accessory proteins such as elongation factors) and proteins involved in RNA biosynthesis. The increase in expression of RNA biosynthetic machinery is necessary for de novo synthesis of ribonucleotides and to ensure flux through nucleotide salvage pathways (mainly to support an increase in rRNA biomass). Finally, the expression profile of the pentose phosphate pathway reflects the interplay between the increasing demand for ribonucleotide precursors and the decreasing demand for amino-acid precursors.

To validate our predicted expression changes, we compared gene clusters with expression data from *E. coli* grown at $0.3\,h^{-1}$ and $\sim 0.5\,h^{-1}$ in a glucose-limited chemostat [183]. In this data set, genes in Energy Metabolism (purple), Core Expression Machinery (orange), and RNA Biosynthesis (red) all significantly change in the predicted direction (Wilcoxon signed-rank test, $P < 1 \times 10^{-4}$), supporting our predicted expression profiles. The other clusters showed no significant changes in the data set; these clusters are either small in size or do not change monotonically, hindering direct comparison with this data set. The ME-Model is not yet predictive of quantitative gene expression changes (at the level of single genes); the correlation over the entire data set is statistically significant ($P < 0.005$), but weak (Pearson's r=0.14). Our approach is at present limited to qualitative predictions of the direction of change of small groups of functionally related proteins.

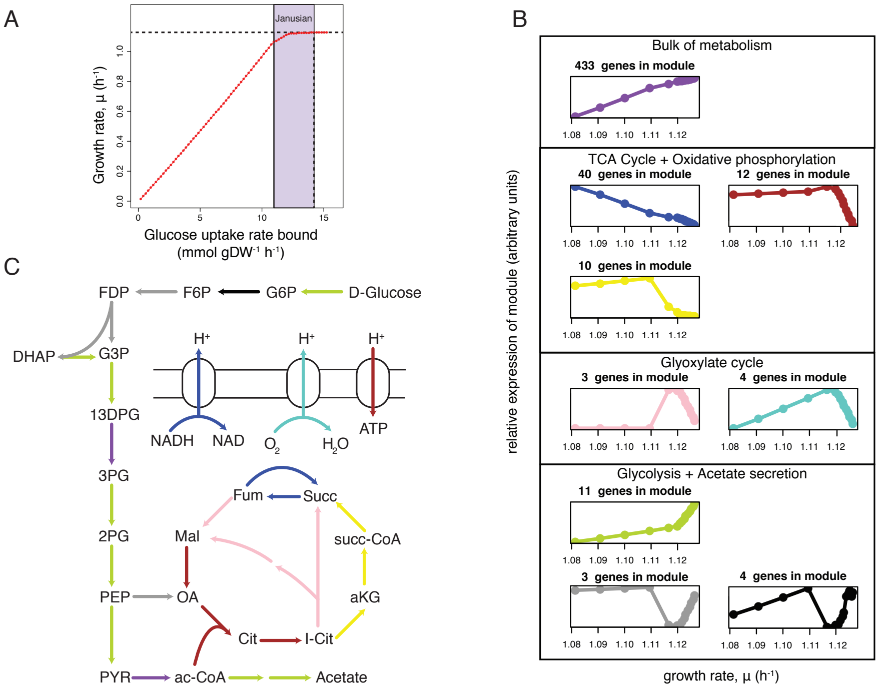


Figure 4.5: **Gene expression during the Janusian region.** (A) Gene expression changes predicted by the ME-Model to occur in the Janusian growth region indicated in purple under glucose limitation in minimal media are analyzed. (B) Simulated expression profiles are clustered using signed power ($\beta = 25$) correlation similarity and average agglomeration. A freely available R package was used (Langfelder and Horvath, 2008). Eleven clusters resulted. Two small clusters were removed because they represented stochastic expression of alternative isozymes. The first principal component of the remaining nine clusters is displayed and grouped qualitatively by function. (C) Many of the expression modules correspond to genes of central carbon energy metabolism. Reactions are colored according to the module color in (B).

In the Janusian region of growth (Figure 4.5), the cell transitions from carbon-limited to proteome-limited constraints, resulting in a distinct transcriptional response. At the beginning of this transition, the cell has reached a nutrient level where enzymes are saturated (Figure 4.1G); as growth rate increases, the total demand of anabolic processes increases, causing a global increase in the bulk of metabolism and gene expression machinery (Figure 4.5B). To meet these proteome demands, energy metabolism is altered to favor lower yield catabolic pathways that require less protein (so that the protein can instead be used for anabolic processes); this is accomplished through a decrease in TCA Cycle and Oxidative Phosphorylation expression in favor of a transient increase in the Glyoxylate Cycle followed by a large increase in Glycolysis and acetate secretion (Figures 4.5B and C), consistent with previously observed changes in gene expression in the transition to glucose-excess environments [182].

The ME-Model predicts intricate expression changes as glucose availability changes by employing relatively simple constraints on molecular catalysis and biomass composition. This study is the first to attempt genome-scale prediction of gene expression levels under changing growth rate and/or nutrient limitation from optimality principles alone. Systematic consideration of transcriptional regulation and inclusion of missing constraints and parameters impacting optimality (e.g., kinetic constraints and parameters) are future endeavors necessary to extend the predictive power to the level of single genes (see Discussion).

4.4 Discussion

The ME-Model is a microbial growth model that computes the optimal cellular state for growth in a given steady-state environment. It takes as input the availability of nutrients to the cell and produces experimentally testable predictions for: (1) the cell's maximum growth rate (μ*) in the specified environment, (2) substrate uptake/by-product secretion rates at μ*, (3) metabolic fluxes at μ*, and (4) gene product expression levels at μ*.

Important to the predictions of the ME-Model is the proper coupling between metabolism and gene product expression. Through comparison of model simulations with experimental data, we derived two general classes of molecular efficiencies that vary based on the growth rate and the degree of nutrient limitation. For ribosomes (and tRNA and mRNA), we propose a growth rate-dependent Michaelis-Menten-type model for polymerization speed, which has preliminary experimental evidence [169], though we have not seen it previously proposed. We furthermore show that two simple assumptions allow us to approximate the effect of nutrient limitation on metabolic enzyme catalysis. While enzyme-specific trends in catalytic rates depend on the limiting nutrient [184, 171], our formulation is a first step toward modeling genome-scale effects of nutrient limitation and suggests that simple principles may underlie these trends. Both of these molecular efficiency variables are essential for genome-scale modeling of gene expression and warrant future studies to validate and refine them further. Paired proteomic and metabolomic data sets under nutrient-limited conditions will allow for a deeper understanding of nutrient limitation-dependent effective catalytic rates, and new data sets [185] and models [186] on the processes of gene expression can help to refine model parameters and determine their genome-scale effects.

The proteomic constraints inherent to the ME-Model result in qualitatively different growth predictions compared with previous genome-scale models. In the ME-Model, growth rate is not a simple linear function of substrate uptake bounds; instead, the ME-Model predicts a maximal growth rate and optimal substrate uptake rates, which better reflects empirical growth models and better predicts experimentally measured growth rates and substrate uptake rates. The ME-Model reveals three distinct growth regions, which we term SNL, Janusian, and Batch; while nutrient-limited (chemostat culture) and nutrient-excess (batch culture) conditions are commonplace, the Janusian region (where the cell is limited by both nutrient availability and proteome capacity) is rarely considered in microbiology. Interestingly, we observe the Janusian region to occur under carbon limitation but not under various non-carbon limitations. We take this to mean that Janusian regions may exist for non-carbon limitations, but the constraints that may cause them to arise are outside the scope of the current ME-Model.

The proteomic constraints in the ME-Model also improve predictions of by-product secretion and metabolic flux under both nutrient-excess and nutrient-limited conditions. By accounting for the metabolic cost of proteins and limitations of protein production capacity, the ME-Model accurately decouples substrate uptake, growth rate, and growth yield, allowing for important rate-yield trade-offs to be predicted. In particular, we show that seemingly inefficient metabolism in batch culture and under nitrogen limitation (both when carbon is in excess), can be explained and predicted through proteomic trade-offs. This capability rectifies the dominant failure mode in predicting metabolic flux previously reported for M-Models [5], and suggests that a single objective of growth rate (if the proper constraints are included) may be able to predict metabolic fluxes. This result shows that proteomic constraints are necessary to accurately predict metabolic responses—optimal growth and metabolic phenotypes cannot be fully understood without taking gene expression into account. From a practical standpoint, the natural parsimony present in ME-Model simulations [93] strongly reduces the optimal solution space, allowing for more precise predictions, an important feature in diverse applications. The effect of proteomic constraints on secretion phenotypes is of particular importance for applications in systems metabolic engineering, and will be necessary for simulating behavior in complex media and predicting nutrient preferences.

At the level of gene expression, the ME-Model predicts detailed behavior in each growth region. In the SNL and Janusian growth regions, gene modules have distinct nutrient limitation-dependent profiles. A number of the gene modules change in the correctly predicted direction compared with expression data from *E. coli* in a chemostat at different growth rates [182, 183], supporting our predicted expression profiles. By predicting optimal gene expression profiles, the ME-Model aids in understanding the factors shaping the evolution of gene expression patterns (e.g., proteomic constraints and changing biomass composition).

Modeling optimal transcriptional responses is complementary to the elucidation and modeling of specific regulatory mechanisms [13, 187, 188]. It is tempting to relate the expression profiles predicted by the ME-Model to molecular mechanisms underlying the control gene expression *in vivo* [13, 188, 189]. For example, constitutively expressed genes display growth rate-dependent expression trends [12, 13], which might provide the cell with an economical way of responding to global changes in metabolic efficiency [170]. Also, PurR could be responsible for regulating the increase in expression of nucleotide biosynthesis genes at higher growth rates (as PurR is an autorepressor, this could be accomplished through mechanisms described in [13]. Finally, though the primary role of ArcA is to respond oxygen availability [190], it also represses many of the genes in the TCA cycle and Oxidative Phosphorylation that decrease during the glucose-limited to glucose-excess (Janusian) transition [182, 191]. However, as regulatory mechanisms are not explicitly considered in the ME-Model, the relation between regulatory mechanisms and simulated expression profiles is indirect; while this comparison can assist in explaining and expanding upon the functional roles of cellular regulators, much further work is required to validate the resulting hypotheses.

As it is an optimality model, the ME-Model is particularly suited for studies related to adaptive laboratory evolution (ALE). Recently, it was reported that it is not possible to predict some changes that occur during ALE in Batch culture using an M-Model [192]. This is because M-Models only take biomass yield optimization into account; these results are consistent with the rate-yield trade-offs present in the ME-Model under nutrient-excess conditions. In the ME-Model, a number of inherent factors can limit cellular growth (e.g., translation rate and metabolic catalysis); the ME-Model can thus provide alternative hypotheses for the mechanisms of growth increase and aid in understanding the results of ALE.

The ME-Model can simulate coarse- to fine-grained cellular and molecular phenotypes with an improved accuracy and scope compared with previous genome-scale models. The ME-Model shows complex behavior as a result of linear constraints applied to an integrated network. The ME-Model thus shows that intricate and seemingly unintuitive phenotypes can be modeled at a genome-scale with simple enough assumptions to understand their underlying cause. Due to the richness of the model simulations, we primarily focused on *E. coli* growing in glucose minimal media at different growth rates by modulating the availability of glucose; there are therefore many future opportunities to investigate model predictions under many environmental and genetic conditions.

A whole-cell *E. coli* model has been desired for some time [142] as such a model would have profound impacts for basic microbiology, the study of microbial communities, antibiotic discovery, the elucidation of regulatory networks, and systems metabolic engineering. We hope the ME-Model will serve as a scaffold for continued model development toward these practical applications.

4.5 Materials and methods

4.5.1 Network reconstruction

The two primary reaction networks used to create the ME-Model were the most recent metabolic reconstruction [157], and a network detailing the reactions of gene expression and functional enzyme synthesis [124]. The gene expression reconstruction is formalized as a set of 'template reactions' that can be applied to different components (e.g., gene, peptide, and set of peptides) to generate balanced reactions. Merging the *E. coli* metabolic network reconstruction with the gene expression reconstruction required a conversion of the Boolean Gene-Protein-Reaction associations (GPRs) into protein complexes. We utilized EcoCyc's annotation to map gene sets to functional enzyme complexes. The content of the final reconstruction is detailed in Supplementary Tables S1, S9, and S10 in [158].

4.5.2 Coupling constraint formulation and imposition

Coupling constraints provide a mechanism for linking the flux values of one or more reactions in the ME-Model. For example, they were used to bound the number of proteins that may be translated from an mRNA before the mRNA decays or is transmitted to a daughter cell. They are also the mechanism through which we related enzyme abundance and activity. Often, the coupling constraints are a function of the organism's growth rate (μ). The coupling constraints are a set of inequality constraints appended to the stoichiometric matrix as additional rows. Assumptions and literature citations for all parameters used can be found in Supplementary information in [158].

4.5.3 Optimization procedure

As the demand reactions and coupling constraints are functions of the organism's growth rate (μ), growth-rate optimization is not a linear program (LP) as in metabolic models, which rely on a linear biomass objective function. Instead, to optimize for growth rate, we solve a sequence of LPs to search for the maximum growth rate, μ^*, that still results in a feasible LP. This search for μ^* is accomplished through a binary search; the search procedure is slightly different depending on whether the cell is proteome-limited (Janusian and Batch growth modes) or SNL. Detailed traces of the execution of the optimization procedures can be found in Supplementary information in [158].

4.5.4 Hierarchical clustering

For Figure 4.4B, relative fractional proteome mass was calculated for each gene-enzyme pair. If a gene is present in multiple enzyme complexes, then it is represented twice, and all subunits of an enzyme complex are counted separately. To filter out the stochastic expression of alternative isozymes (to make the observed trends clear), we eliminated gene-enzyme pairs that were not expressed across all growth rates and filtered gene-enzyme pairs that changed in relative expression by >0.3 across more than one pair of consecutive growth rates. Hierarchical clustering was performed on the resulting expression profiles; we used a signed power ($\beta = 6$) correlation similarity (as in [193]) and average agglomeration.

4.5.5 File formats and accessibility

The model is freely available as part of the openCOBRA Project (http://opencobra.sourceforge.net).

4.6 Acknowledgements

We thank Thorsten Koch, Matthias Miltenberger, Ambros Gleixner, Michael Saunders, and Martina Ma for help solving linear programming problems. We thank Adam Feist, Harish Nagarajan, and Pep Charusanti for invigorating discussions. EJO and RLC were supported by NIH R01 GM057089. JAL was supported by NIH U01 GM102098. This work was supported in part by the US National Institute of Allergy and Infectious Diseases and the US Department of Health and Human Services through interagency agreement Y1-AI-8401-01. DRH is supported in part by Y1-AI-8401-01. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231.

Author Contributions: JAL, EJO, and BP conceived the study. JAL and EJO performed the computational analysis. JAL, EJO, and RLC performed model updates. RLC, DRH, and BP supervised the study. JAL and EJO wrote the manuscript. All authors helped edit the final manuscript.

Chapter 4 in full is a reprint of a published manuscript: O'Brien EJ*, Lerman JA*, Chang RL, Hyduke DR, Palsson BO. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol Syst Biol. 2013 Oct 1;9:693. doi: 10.1038/msb.2013.52. The dissertation author was the primary author of this paper responsible for the research. The other authors were Edward J. O'Brien (equal contributor), Roger L. Chang, Daniel R. Hyduke, and Bernhard Ø. Palsson.

Chapter 5

Reconciling a *Salmonella enterica* metabolic model with experimental data confirms that overexpression of the glyoxylate shunt can rescue a lethal *ppc* deletion mutant

5.1 Abstract

The *in silico* reconstruction of metabolic networks has become an effective and useful systems biology approach to predict and explain many different cellular phenotypes. When simulation outputs do not match experimental data, the source of the inconsistency can often be traced to incomplete biological information that is consequently not captured in the model. To address this problem, general approaches continue to be needed that can suggest experimentally testable hypotheses to reconcile inconsistencies between simulation and experimental data. Here, we present such an approach that focuses specifically on correcting cases in

which experimental data show a particular gene to be essential but model simulations do not. We use metabolic models to predict efficient compensatory pathways, after which cloning and overexpression of these pathways are performed to investigate whether they restore growth and to help determine why these compensatory pathways are not active in mutant cells. We demonstrate this technique for a *ppc* knockout of *Salmonella enterica* serovar Typhimurium; the inability of cells to route flux through the glyoxylate shunt when *ppc* is removed was correctly identified by our approach as the cause of the discrepancy. These results demonstrate the feasibility of our approach to drive biological discovery while simultaneously refining metabolic network reconstructions.

5.2 Introduction

The *in silico* reconstruction of metabolic networks is a systems biology framework that serves as a collection of highly curated genetic and biochemical information for a particular organism [19, 194, 195]. The subsequent conversion of this parts list to a mathematical format allows one to simulate phenotypic states and consequently to investigate different relationships between genotype and phenotype using a computational model. The ability to simulate different phenotypes is a notable strength of this modeling framework and distinguishes this approach from static maps of biochemical pathways. Static maps represent all known pathways in a network, whereas the simulation of metabolic network reconstructions provides additional information concerning which pathways are likely to carry flux - and are therefore active - vs. pathways that are present but not used. To date, the simulation of metabolic models has found applications in metabolic engineering [196, 197, 198], network analysis [199, 200, 201], biological discovery [202, 203, 204], and target identification for drug discovery research [205, 206]. Metabolic models have also been used to investigate drug off-target effects by incorporating structural data for proteins [207] and to provide context for the analysis of high-throughput omics data [208, 207, 209, 204].

Because these models are constructed from experimental data, attempts to reconcile inconsistencies between experimental data and model simulations often form the basis for hypothesis-driven biological discovery [210]. In turn, the new discoveries serve to refine the models. This continuous loop in which one performs simulations, carries out experiments to test simulation results, resolves inconsistencies, and then performs new simulations to start a new round of model reconciliation ultimately improves the predictive capabilities of the models and thereby increases their utility. In one example, the use of systems analysis coupled with high-throughput screening and follow-up genetic and biochemical work led to the functional assignment of eight ORFs in *Escherichia coli* that had two new enzymatic activities and four unidentified transport properties [203]. Other studies in *E. coli* identified a new mechanism that enables growth on myo-inositol [211] and the gene that encodes succinate semialdehyde dehydrogenase in this organism [212].

Several computational algorithms have been developed to close gaps in metabolic models and to reconcile inconsistencies between model simulations and experimental data [210, 213]; however, there is a constant need for new gap filling strategies that complement existing ones so that the accuracy of metabolic network reconstructions can continue to be refined and improved. Here, we present such a method based on an analysis of a metabolic model for a knockout mutant vs. the wild-type. Our method focuses specifically on formulating hypotheses that can correct metabolic models when a gene is essential experimentally but it is nonessential in the model. We demonstrate our approach by investigating a case in which a metabolic reconstruction for *Salmonella enterica* serovar Typhimurium (hereafter referred to as *S.* Typhimurium; [214] predicted *in silico* that a Δppc mutant would be viable in glucose M9 minimal medium but the actual mutant, when constructed and tested in the laboratory, was not. The absence of key regulatory information from the model was found to be the cause of the discrepancy.

5.3 Materials and methods

5.3.1 Bacterial strains

Bacterial strains used in this study are summarized in Table 5.1. *Salmonella enterica* serovar Typhimurium strain 14028s was a generous gift provided by Fred Heffron at Oregon Health & Science University.

Table 5.1: Strains and plasmids used in this study. Amp^R and Cam^R indicate genes that confer resistance to ampicillin and chloramphenicol, respectively

Strain or plasmid	Characteristics	Source
Salmonella enterica serovar Typhimurium strain 14028s	Wild-type	See Materials and methods
Δppc	*ppc* deletion mutant	This study
$\Delta ppc \Delta iclR$	*ppc* and *iclR* double deletion mutant	This study
ppc(pS7)	The Δppc mutant bearing plasmid pS7	This study
ppc(pS8)	The Δppc mutant bearing plasmid pS8	This study
ppc(pS10)	The Δppc mutant bearing plasmid pS10	This study
ppc(pS8 + pS10)	The Δppc mutant bearing the plasmids pS8 and pS10	This study
pKD13	PCR template used to generate *ppc* knockout cassette based on kanamycin resistance	[215]
pKD46	Encodes arabinose-inducible λ-Red recombination system	[215]
pCP20	Encodes FLP recombinase	[215]
pASK-IBA33+	Expression plasmid containing a tetracycline inducible promoter; Amp^R	IBA GmbH, Germany
pASK1988	Derivative of pASK-IBA33+ containing Cam^R in place of Amp^R	This study
pS7	pASK-IBA33+ containing *aceBAK*	This study
pS8	pASK-IBA33+ containing *aceBA*	This study
pS10	pASK1988 containing *aceK*	This study

5.3.2 Growth media

Strains of Salmonella Typhimurium 14028s were cultured at 37 °C with magnetic stir bars for aeration in either Luria-Bertani (LB) broth or M9 minimal medium. The M9 medium contained 2 g L^{-1} glucose, 100 μM CaCl2, 2 mM MgSO4, 6.8 g L^{-1} Na2HPO4, 3 g L^{-1} KH2PO4, 0.5 g L^{-1} NaCl, 1 g L^{-1} NH4Cl, and 250 μL L^{-1} trace elements. The trace element solution consisted of (L^{-1}) FeCl3∘6H2O (16.67 g), ZnSO4∘7H2O (0.18 g), CuCl2∘2H2O (0.12 g), MnSO4∘H2O (0.12 g), CoCl2∘6H2O (0.18 g) and Na2EDTA∘2H2O (22.25 g). Antibiotics were added as necessary at the following concentrations: ampicillin at 100 μg mL^{-1}, kanamycin at 50 μg mL^{-1}, and chloramphenicol at 25 μg mL^{-1}. LB powder was purchased from EMD Chemicals (Gibbstown, NJ) and used at the manufacturer's recommended concentration. All other chemicals were purchased from Sigma-Aldrich (St. Louis, MO).

5.3.3 Construction of the Δ*ppc* mutant

The Δ*ppc* knockout mutant in *S.* Typhimurium 14028s was created using the λ-Red recombination system [215]. A kanamycin resistance cassette containing flanking FRT sites was generated by PCR using pKD13 as the template. The ends of the cassette contained 50 nucleotides that are homologous to the 50 bp immediately upstream and downstream of *ppc*. The plasmids pKD46 and pCP20 were used to insert the kanamycin cassette into the genome via homologous recombination and to remove the kanamycin resistance marker, respectively. Correct insertion of the marker and subsequent removal from the chromosome were confirmed by PCR. All PCR products were purified with the QIAGEN PCR clean-up kit (Valencia, CA).

5.3.4 Growth rate and glucose uptake rate measurements

For growth rate measurements, strains were first cultured in LB media overnight, centrifuged at approximately 3000 g, washed twice with glucose M9, and an aliquot inoculated into 25-mL flasks containing 10 mL of glucose M9. Cultures were aerated at 37 °C in an air incubator using a magnetic stir bar that spun inside the flask. The next day, an aliquot was inoculated into 250-mL Erlenmeyer flasks containing 100 mL glucose M9 media in triplicate such that the initial OD600 nm was 0.05. The flasks were then transferred to a 37 °C water bath with continuous magnetic stir bar aeration as described above. Cell growth was monitored by measuring the OD600 nm every 30 min. The growth rate was then calculated by fitting an exponential curve to the time course OD600 nm measurements.

The glucose uptake rate was obtained by collecting supernatant at the same time we took each OD600 nm measurement. The supernatant was first filtered through 0.22-μM syringe-fitted membranes and then injected into a Waters HPLC system fitted with a Bio-Rad Aminex HPX-87H ion exclusion column (300 × 7.8 mm). The mobile phase was 5 mM H2SO4; the flow rate was 0.5 mL min^{-1}; and the area of the glucose peak in each sample was measured using refractive index detection. The glucose concentration in each sample was then obtained through comparison to a standard curve. Lastly, the glucose uptake rate was calculated from the glucose concentration at each time point, the growth rate, and an experimentally determined value of 0.41 g dry weight (gDW) per liter per unit OD ($R^2 = 0.94$).

5.3.5 *In silico* modeling

The genome-scale metabolic models for *Salmonella* Typhimurium LT2 [214] and *Escherichia coli* K-12 MG1655 [157] were implemented using the COBRA toolbox [216]. Growth simulations were performed by constraining the model such that model parameters representing the *in silico* growth environment mimicked glucose M9 media and then by maximizing the default objective function. In both models, the default objective function was an equation representing biomass production from the different cellular components (e.g. DNA, RNA, amino acids) in stoichiometric amounts [120]. To compute the minimized sum of fluxes, the growth rate was first fixed to 0.1 h^{-1}, while the sum of all fluxes in the model was minimized [127]. A priority score was defined for each simulated double knockout because some double knockouts do not show a decrease in growth rate despite an increase in the minimal sum of fluxes. This effect stems from the fact that metabolic models do not account for the total cost of protein synthesis. The priority score was defined for each pair of genes as follows: Minimum flux with knockout of gene 1 and gene 2 - Maximum (Minimum flux with knockout of gene 1, Minimum flux with knockout of gene 2). Synthetic lethality was determined by computationally removing all pairs of enzymes in the model. Gene pairs for which the *in silico* growth rate was $< 0.001\ h^{-1}$ were defined as synthetically lethal.

5.3.6 Construction of pASK1988

Plasmid pASK1988 was constructed by replacing the ampicillin resistance gene of pASK-IBA33+ (IBA GmbH, Goettingen, Germany) with the chloramphenicol resistance gene of pACBSR [217]. The CamR gene from pACBSR was PCR amplified with primers that included an AgeI restriction site on one end and a BlpI site on the other. The plasmid pASK-IBA33+, excluding the ampicillin resistance gene, was similarly PCR amplified using primers that introduced AgeI and BlpI restriction sites on each end. All PCR was carried out using Phusion DNA polymerase. Both PCR products were purified from a 1% agarose gel using the QIAGEN QIAquick Gel Extraction Kit, digested with AgeI and BlpI, and ligated with T4 DNA ligase at 16 °C overnight. The ligated products were then transformed into TOP10 cells (Life Technologies, Carlsbad, CA) by heat shock at 42 °C, recovered in SOC media, and plated on LB agar plates containing chloramphenicol. Successful transformants were cultured in LB media with chloramphenicol overnight for plasmid isolation the following day using the QIAprep Spin Miniprep Kit. We confirmed successful replacement of AmpR with CamR by PCR. The AgeI and BlpI restriction enzymes, Phusion DNA polymerase, and T4 DNA ligase were all purchased from New England BioLabs (Ipswich, MA).

5.3.7 Construction of pS7, pS8, pS10

The genes *aceBA*, *aceK*, and *aceBAK* were amplified by PCR using *S.* Typhimurium 14028s genomic DNA as a template. The full *aceBAK* operon and the genes *aceBA* were cloned into pASK-IBA33+ according to the manufacturer's directions, yielding pS7 and pS8, respectively. The gene *aceK* was cloned into pASK1988, yielding pS10. All plasmids were isolated using the QIAprep Spin Miniprep Kit.

5.3.8 Induction and protein overexpression

Strains bearing pS7, pS8, pS10, or both pS8 and pS10 were first cultured in LB medium overnight with the appropriate antibiotic. The next day, an aliquot was inoculated into two 250-mL Erlenmeyer flasks containing 20 mL of LB media such that the initial OD600 nm was 0.05. After allowing the cultures to grow at 37 °C until they reached mid-log phase (OD600 nm approximately 0.5), anhydrotetracycline (ATc) was added to one of the flasks to a final concentration of 100 ng mL^{-1}. Both flasks were then cultured for an additional 3 h, after which the cultures were spun down and washed twice with glucose M9 media. The washed cells were then inoculated into 250-mL Erlenmeyer flasks containing 100 mL of M9 medium with or without inducer in triplicate. The six flasks were then placed in a 37 °C water bath and their OD600 nm values measured periodically over the following several hours.

5.4 Results

5.4.1 In contrast to model simulations, a Salmonella Typhimurium Δ*ppc* mutant is nonviable in glucose M9 medium

We implemented the consensus *S.* Typhimurium metabolic reconstruction [214] and model simulations suggested that a Δ*ppc* knockout mutant would be viable in glucose M9 medium. When the Δ*ppc* knockout mutant was experimentally constructed, it was found to be viable in LB medium but not in glucose M9. Supplementation of the glucose M9 medium with 5 mM succinate restored growth of the Δ*ppc* mutant to a rate of 0.87 ± 0.010 h^{-1}, similar to that of the wild-type

$0.86 \pm 0.021\ h^{-1}$. The glucose uptake rate calculated for the Δ*ppc* mutant on the supplemented M9 medium was 14 ± 6.2 mmol $gDW^{-1}\ h^{-1}$, which was also similar to the value for the wild-type (15 ± 0.41 mmol $gDW^{-1}\ h^{-1}$).

5.4.2 Comparing efficient flux states enables a hypothesis-driven approach to reconcile metabolic models with experimental data

We hypothesized that overexpressing one or more key genes in the Δ*ppc* background might restore growth and thereby reconcile simulation results with the experimental data. Furthermore, if the reactions that correspond to these key genes are indeed compensatory, then they likely form a bottleneck - and therefore become critical reactions - when the first gene is deleted. We have developed a procedure to identify such bottlenecks by first calculating the minimized sum of all fluxes for a reference metabolic model, a value that represents an optimal flux state for the entire network. The reference here is the model for Δ*ppc*. Next, one constructs models of all possible double knockouts such that *ppc* is one of the two deleted genes, and the minimized sum of fluxes is likewise calculated for each member of this set. The reaction corresponding to the second gene in the pair is a potential bottleneck if the double mutant shows a significant increase in the minimized total flux over the single Δ*ppc* knockout. This second gene thereby becomes a potential candidate for overexpression.

We performed this analysis for Δ*ppc* using both the *S.* Typhimurium [214] and *E. coli* [157] metabolic reconstructions. Even though all experimental work performed here was carried out in *S.* Typhimurium, we utilized the *E. coli* reconstruction as well because it is the most extensively curated reconstruction for a microbe and because of the close phylogenetic relationship between the two organisms. Both models pointed to isocitrate lyase, encoded by *aceA* , as the key compensatory enzyme, but in different ways. In the *E. coli* model, the minimized sum of fluxes increases 1.3-fold for the Δ*ppc*Δ*aceA* double mutant and has the highest priority score of 23.4 (see Supporting Information, Table S1 in [218]). In the *S.* Typhimurium model, the two genes are synthetically lethal (Table S2 in [218]). Isocitrate lyase is part of the *aceBAK* operon, which encodes genes for the glyoxylate shunt (Figure 5.1). Supplementation of glucose M9 media with intermediates of the glyoxylate shunt (glyoxylate, malate, and succinate) restored growth in the Δ*ppc* mutant (Table S3 in [218]). In *E. coli*, a prior study noted increased flux through the glyoxylate shunt in an adaptively evolved Δ*ppc* mutant; however, no direct causal link between the two was conclusively proven [219].

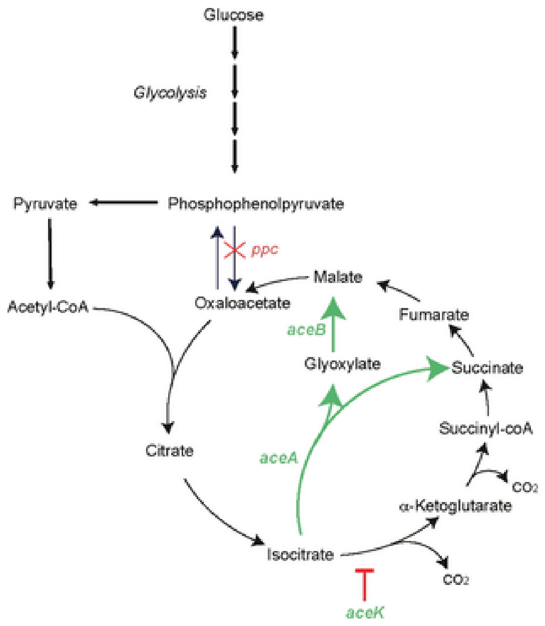


Figure 5.1: **The location of the *ppc* knockout and the glyoxylate shunt within glycolysis and the TCA cycle.**

5.4.3 Deleting *iclR* from the Δ*ppc* mutant restores viability

In *E. coli*, the Δppc mutant is unable to convert phosphoenolpyruvate (PEP) into oxaloacetate, which diverts PEP toward pyruvate biosynthesis [220]. Excess pyruvate, in turn, can activate IclR, a transcription factor that regulates transcription of genes involved in the glyoxylate shunt [221]. We therefore created a $\Delta ppc\Delta iclR$ double mutant in *S.* Typhimurium to investigate this possible mechanism linking deletion of *ppc* to the glyoxylate shunt and viability. Growth was restored in the $\Delta ppc\Delta iclR$ double mutant at a rate of $0.45 \pm 0.01\ h^{-1}$ in glucose M9 medium, which is approximately 60% of the wild-type growth rate (Figure 5.2).

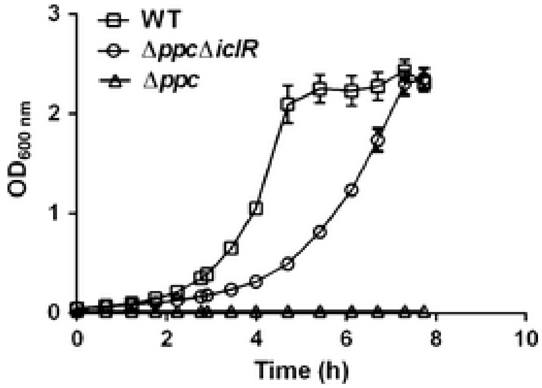


Figure 5.2: **Growth curves for wild-type *Salmonella* Typhimurium and the Δ*ppc* and Δ*ppc*Δ*iclR* mutants in glucose M9 media.** Error bars represent the standard deviation from three independent replicates.

5.4.4 Simultaneous expression of *aceBA* and *aceK* from two separate plasmids can rescue growth in the Δ*ppc* mutant, but overexpression of *aceBA*, *aceK*, or *aceBAK* individually from a single plasmid cannot

We next overexpressed *aceBAK* using the pASK-IBA33+ inducible expression vector to confirm directly whether this operon by itself can rescue growth in the Δ*ppc* mutant. The inducer was anhydrotetracycline (ATc), which can be toxic at high concentrations [222]. We performed a dose-response study to assess its toxicity to *S.* Typhimurium and found that ATc did not inhibit growth at concentrations up to 100 ng mL^{-1} (Fig. S1 in [218]). Both *aceBA* and *aceK* were then cloned into pASK-IBA33+ and pASK1988 (Fig. S2 in [218]), yielding pS8 and pS10, respectively, and both transformed into the Δ*ppc* mutant, yielding strain *ppc*(pS8+pS10). Induction and simultaneous expression of *aceBA* and *aceK* rescued growth in the Δ*ppc* mutant (Figure 5.3). The *ppc*(pS8+pS10) mutant was no longer viable in glucose M9 when it was cured of one of the two plasmids (data not shown). Consistent with this observation, transformants bearing either pS8 or pS10 were also not viable in glucose M9 (data not shown).

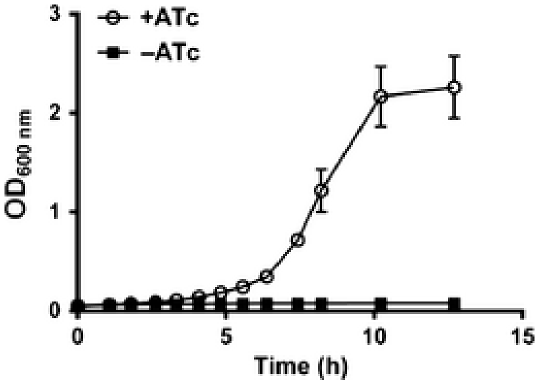


Figure 5.3: **Growth curves for *ppc*(pS8 + pS10) in glucose M9 media in the presence and absence of the inducer anhydrotetracycline (ATc).** Error bars represent the standard deviation from three independent replicates.

5.5 Discussion

The *in silico* reconstruction of metabolic networks provides a computational framework with which to organize genomic, transcriptomic, proteomic, and metabolomic data, allowing one to compute phenotypic states from genome-scale information. Continual refinement of the models to ensure consistency with experimental data serves to improve their accuracy and predictive ability. We present a method here for model refinement that focuses on reconciling inconsistencies between simulated vs. experimental gene essentiality data that is based on an analysis of synthetic lethality and the minimized sum of fluxes in the models. We demonstrate our approach with *ppc*/*aceBAK* by showing that overexpression of *aceBA* and *aceK* is sufficient to rescue the Δ*ppc* mutant. Viewed another way, the *S.* Typhimurium metabolic model is incorrect because it erroneously allows flux to flow through the glyoxylate shunt when *ppc* is deleted due to the absence of regulatory information (i.e. *iclR* and *aceK*). We also provide indirect evidence that the key regulatory feature is a 184 intergenic region between *aceA* and *aceK*. This regulatory information would now ideally be incorporated into the *S.* Typhimurium model to update and refine it, but doing so is challenging computationally as it would require implementing conditional enzyme expression rules into a model that is based primarily on metabolism. However, these data can be more easily incorporated into expanded models that account for both metabolism and gene expression [93].

Expression of the *aceBAK* operon must occur on two separate plasmids to rescue the Δ*ppc* mutant. The proteins AceB and AceA catalyze the two reactions of the glyoxylate shunt, whereas *aceK* is a regulator that controls the branch point between the TCA cycle and the glyoxylate bypass [223]. The intergenic region between *aceB* and *aceA* is only 32 bp long, but between *aceA* and *aceK* it is 184 bp [224, 225]. The long space between the latter two genes is palindromic and therefore is capable of forming a stable stem-loop structure, which might play a significant role in regulating transcription or *aceK* [226, 227]. Data presented here support this hypothesis: growth could be restored only when *aceBA* and *aceK* were cloned separately into two different expression plasmids such that the 184 intergenic region was removed. We did not observe restoration of growth in the Δppc mutant when we cloned the *aceBAK* operon in its entirety and attempted to express it from a single expression vector (pS7).

We resequenced the genome of the *ppc*(pS8+pS10) mutant using Illumina technology to confirm that there were no additional mutations in the genome that might have contributed to restored growth. No mutations were detected, providing further evidence that overexpression of *aceBA* and *aceK* by itself is sufficient to rescue growth.

5.6 Acknowledgements

This work was funded by grant 1R01GM098105 from NIH/NIGMS to B.O.P. The authors thank Ramy Aziz, Kathy Andrews, Howard Li, Mallory Embree, Haythem Latif, Harish Nagarajan, and Edward O'Brien for helpful discussions. The authors declare no conflict of interest.

Chapter 5 in full is a reprint of a published manuscript: Fong, N. L., Lerman, J. A., Lam, I., Palsson, B. O. and Charusanti, P. (2013), Reconciling a *Salmonella enterica* metabolic model with experimental data confirms that overexpression of the glyoxylate shunt can rescue a lethal *ppc* deletion mutant. FEMS Microbiology Letters, 342: 6269. doi:10.1111/1574-6968.12109. The dissertation author was the second author of this paper, responsible for the computational analysis that inspired the research. The other authors were Nicole L. Fong, Irene Lam, Bernhard Ø. Palsson, and Pep Charusanti.

Chapter 6

ME-Models as a conduit for integration of systems and synthetic biology

6.1 Introduction

Synthetic biology approaches are rapidly maturing, making it possible to engineer genes with predictable mRNA and protein expression levels in model organisms such as *E. coli.* Tools developed by synthetic biologists can now accurately compute sequence-dependent binding energies for interactions between: (1) RNA polymerase and an arbitrary DNA promoter sequence [228], and (2) the ribosome and an arbitrary Shine-Dalgarno sequence upstream of a coding sequence on an mRNA [229]. Promoter strength can be tuned to yield gene expression over approximately 3 orders of magnitude, while the 'Ribosome Binding Site Calculator' out of the lab of Howard Salis at Penn State University allows for tuning protein levels over a range of 100,000-fold. The latest version of the RBS Calculator takes RNA secondary structure around the Shine-Dalgarno sequence into account, since this has been shown to influence the ability of the ribosome to bind these elements. This refinement is just one in a long series of refinements that has led to the genesis of predictive tools in this space. It's efforts like this that are closing the gap between genotype (the actual sequence at base-pair resolution) and phenotype.

Recently, these tools were put to the test on a massive scale. A library composed of thousands of combinations of promoters and ribosome binding sites (RBSs) was constructed by George Church and colleagues at Harvard. Each promoter and RBS combination drives expression of GFP (or the green fluorescent protein). Real-time PCR (polymerase chain reaction) and relative fluorescence intensity measurements were applied to determine relative RNA and protein levels for each construct in the library. The data from this screen (presented in [230]) indicates that 92.5% of the variance in RNA expression levels can be explained by the promoter sequence. An additional 3.8% can be explained by the sequence of the RBS, which clearly exerts its influence on RNA expression levels post-transcriptionally. One hypothesis as to why there is an influence is that a strong RBS leads to a higher density of ribosomes on the transcript, which in turn may sterically hinder the RNA degradation machinery. Astonishingly, only 3.7% of the variance remains unexplained. Protein expression levels could not be predicted as well, but they were still predicted decently: 53.8% of the variance in protein expression levels could be explained by the promoter sequence, while an additional 29.6% could be explained by the RBS sequence. 16.7% of the variance in protein levels remains unexplained.

Synthetic approaches have advanced to the point that it is now clearly easier to predict heterologous gene expression levels vs. predicting the expression levels of native genes on the chromosome! When the synthetic biology tools mentioned above are applied to natural promoter and RBS sequences, they don't perform nearly as well [personal communication with Ali Ebrahim and my own experience with the data for *Thermotoga maritima*]. In natural systems, the expression of a gene is tuned in many ways (probably tens to hundreds) during the course of evolution. As François Jacob stated [8], evolution is a "tinkerer, not an engineer." Evolution has many ways to tune expression beyond modulating the initiation rates of transcription and translation (synthetic biologists have focused almost exclusively here). One example is the control of steady-state mRNA levels through tweaks in degradation rates. Gene expression is also influenced by supercoiling and the 3D structure of the chromosome. Ultimately, evolution randomly interleaves the genetic code with many other codes (many of which we have not discovered or are unsure how to read yet). The specific factors underlying the expression levels of native genes are therefore indecipherable to us; however, in synthetic biology the determinants of expression can be engineered to be completely ‘orthogonal’ to the native regulatory circuits by ‘refactoring’ the operons. This means that the genetic elements in the operon are re-organized and re-coded (sometimes randomly) to limit the influence of native regulation as much as possible. A good example of refactoring was recently put forth by Chris Voigt and colleagues [231]. There, they refactored the nitrogen fixation gene cluster in *Klebsiella oxytoca.* The refactored cluster has almost the same activity as the native nitrogen fixation system, but the various cistrons are highly organized because they were systematically designed. Later, they built a version that works in *E. coli*, whereas transplanting the original system from *Klebsiella oxytoca* leads to little or no activity (probably due to differences in regulation). Similar approaches have also been applied to gene clusters important for antibiotic production in *Streptomyces orinoci* [232].

The effect of adding expression vectors (plasmids) to model organisms such as *E. coli* can be quantitatively modeled now that synthetic biology approaches are reaching a level of maturity. Models can be used to compute the systems-wide effects of heterologous gene expression, whether these effects are metabolic burdens due to expression of the vector and its products, or as a result of new metabolic pathways siphoning resources away from growth. A few key parameters such as promoter and RBS strength can be translated directly into model constraints on mRNA and protein production rate, but other key parameters such as the k_{eff} parameters for newly introduced enzyme-reaction pairs remain unknown and must be sampled or approximated using *in vitro* approaches.

I modified the code used for the project presented in Chapter 4 to allow for heterologous gene expression from a plasmid (or plasmids if multiple expression vectors are desired). In the sections that follow, I provide results showing that ME-Models provide a basic conduit for integration of systems and synthetic biology. In a few years time, hopefully we’ll be more routinely writing genetic code from scratch and simulating the functions of these sequences in models such as the *E. coli* ME-Model. Synthesizing whole genes is gaining traction now that costs have dropped significantly (from more than \$10 per base prior to the year 2000 to less than \$1 per base pair around 2005 [233]). My hope is that when it becomes even more economical to order entire vectors/designs, they can also be easily simulated in the context of the larger biological system. When one can test an ordered construct *in silico* by simply uploading the order file to a web-based modeling application, we'll know we're on the right track to closing the gap between systems and synthetic biology approaches to engineering life. What follows are a few basic illustrative examples.

6.2 pUC19 cloning vector

Here, I perform basic simulations of an *E. coli* cell carying the pUC19 cloning vector. The pUC19 cloning vector was created by Messing and co-workers [234], and is one of the most popular vectors for heterologous gene expression.

To simulate the effect it has on *E. coli*, I added reactions to express pUC19 *in silico*. I left the copy number (the number of plasmids per cell) as a free variable. Typically, the copy number relates to the strength of the origin(s) of replication on the plasmid, but this is not precisely known for pUC19 (and many other vectors). Additionally, as in [235], I added a constraint to enforce a stoichiometric relationship between plasmid production and production of each open reading frame on the plasmid (the number of proteins produced per plasmid). This parameter lumps the transcriptional and translational efficiency. Looking at the data from George Church's screen, many values for this parameter are achievable experimentally (up to a maximum of about 2000 proteins per plasmid, considering a value of 1838 corresponds to fully activated LacZ production from the *lac* operon, which has very strong binding sites [235]).

The first simulation performed relates to expression of *bla*, or the beta-lactamase enzyme (EC 3.5.2.6) that is present on the plasmid to confer resistance to Ampicillin and aid in selection of cells harboring pUC19. In deciding how to couple *bla* production to plasmid production, I consulted literature and found that it is not uncommon for antibiotic resistance markers to take up approximately 20% of the proteome of the cell [235, 236]! A typical *E. coli* proteome is composed of approximately 3 million proteins. Assuming about equal molecular weights, there would therefore be approximately 600,000 copies of *bla* per cell. A conservative (quite high) estimate for the pUC19 plasmid copy number (per cell) is 700. Taking 600,000 *bla* copies per cell / 700 pUC19 copies per cell, I reasoned an appropriate coupling value would be approximately 857 *bla* copies per copy of pUC19. Figure 6.1 shows the predicted impact on growth rate for maintenance of pUC19 at various copy numbers per cell. The simulation demonstrates that the burden of plasmid carriage mostly arises due to the expression of plasmid-encoded protein (vs. the expression of the vector itself).

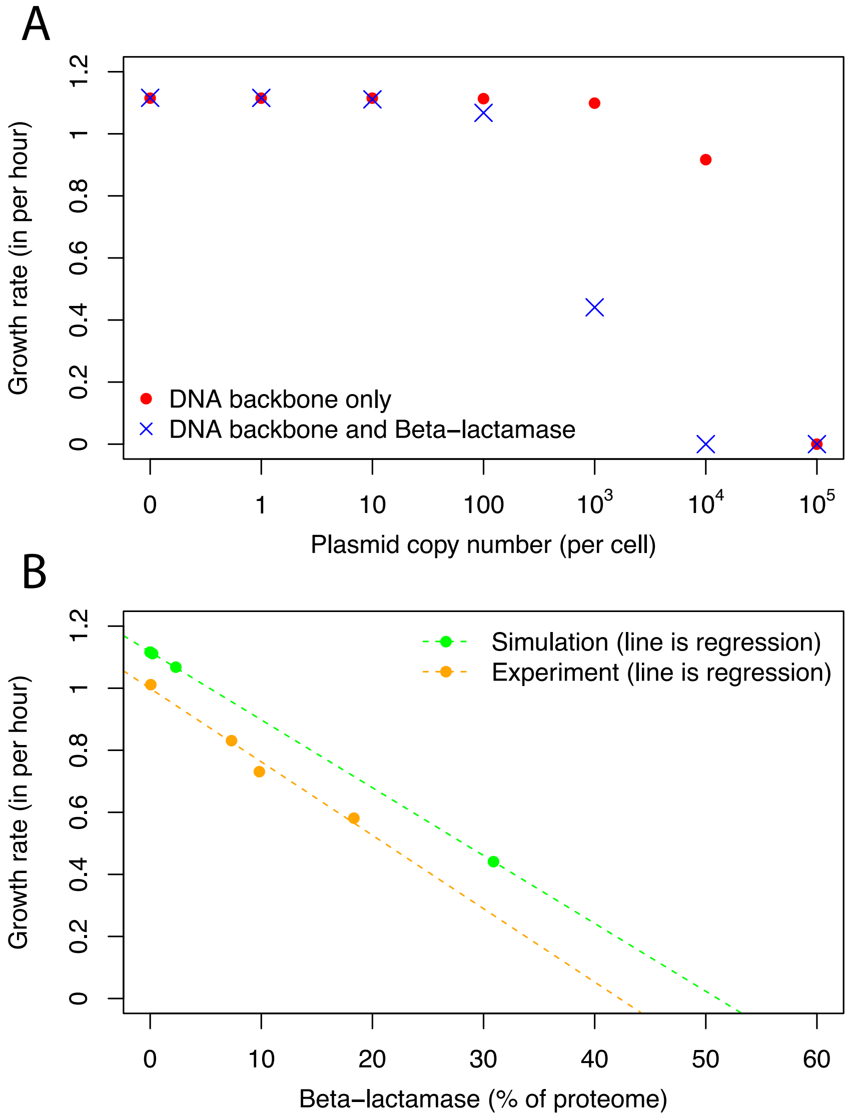


Figure 6.1: **Metabolic burden of plasmid maintenance.** A) The predicted impact on growth rate for maintenance of pUC19 at various copy numbers per cell (x-axis). The red dots indicate the predicted impact if only the maintenance of the DNA backbone is considered. The blue 'X' marks show the predicted response when both the maintenance of the DNA backbone and expression of beta-lactamase are enforced. B) Percent of the proteome occupied by beta-lactamase and its impact on growth rate. Orange dots are data points from [237] as in [11], whereas green points derive from ME-Model simulations. Note: Plasmid vector pBR329 was used for the experiments, though the exact vector probably has little impact on the qualitative shapes of these relationships.

A study in the late 1980s [238] demonstrated that common vectors (pUC19 probably included) lead to unnecessarily high levels of antibiotic resistance gene expression. Weakening the promoter driving expression frees up the proteome for expression of the other genes on the plasmid. Interestingly, the development of vectors without (or with minimal) antibiotic-based selection remains an active area of research today, but it has probably been overlooked somewhat considering pUC19 is still used in our lab for metabolic engineering projects.

6.3 Production of spider silk proteins

ME-Models are well-suited for modeling costs associated with specific gene expression, so I thought it would be interesting to use the model to probe the cost of expressing specific proteins; however, it is somewhat uninteresting to model overexpression of most proteins. Although costs of synthesis vary significantly among amino acids, and every protein has a unique sequence of amino acids, most proteins have a characteristic amino acid composition (approximately 7.4% Alanine, approximately 3.3% Cysteine, and so on). For a protein at or near the typical amino acid composition, the simulation will almost certainly produce a result like that shown in Figure 6.1B (even the slope and intercepts will be approximately the same!).

The spider silk protein is a special case because it has atypical amino acid composition. Spider silk is very rich in glycine (44.9%), and requires metabolic engineering for high levels of production in hosts such as *E. coli*. Recently, Sang Yup Lee and colleagues achieved recombinant spider silk protein production [239]. Their study provides the basis for the analysis that follows, especially for comparison of ME-Model simulations to their actual experience engineering *E. coli* for spider silk production. To simulate spider silk production *in silico*, I replaced the *bla* gene on pUC19 with the spider silk gene and simulated overexpression in the ME-Model. As a control, I compare to simulations expressing *bla*. Figure 6.2 shows the amino acid and codon usage for production of *bla* vs. spider silk. Glycine and glycine codons are expected outliers. Interestingly, one strategy (predictable from the simulation) for producing spider silk experimentally is elevation of the glycyl-tRNA pool.

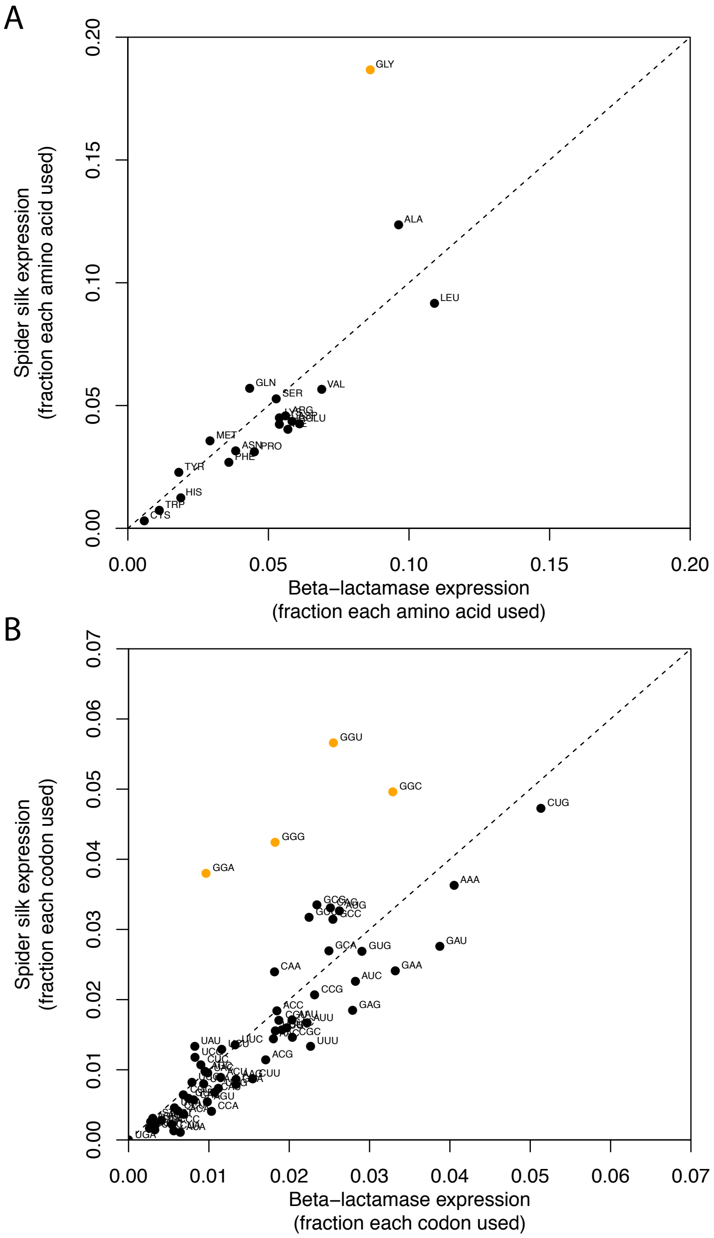


Figure 6.2: **Amino acid and codon usage for production of *bla* vs. spider silk.** A) Amino acid usage (Glycine highlighted in orange). B) Codon usage (with codons coding for Glycine highlighted in orange). Note that there are 61 codons (stop codon usage is not plotted), but 63 data points on the plot. This is because the start codon is plotted separately as 'START,' and UGA is used infrequently to code for L-Selenocysteine. Differential use of Glycine is discussed in the text.

The model also predicts genes that would likely need to change expression in order to support high levels of spider silk production (Figure 6.3). Such changes include the elimination of the glycine cleavage system (*gcvT*, *gcvH*, and *gcvP*), and overexpression of *glyA*, *glyS*, *glyQ*, and *serA-C*, which are needed in higher abundance to support production of spider silk. The effects of these changes were verified experimentally [239], and found to match the model's predictions. Interestingly, adding exogenous glycine to the media did not lead to increased production because it led to allosteric inhibition of *serA*. This could not be predicted by the ME-Model (though the need for *serA* was correctly predicted). The model also suggests *folD* overexpression could be beneficial, though this remains to be tested experimentally.

These computations were provided as an illustrative example. In this particular case, many of the predictions are intuitive, and a detailed model is not necessarily needed. But that's not always going to be the case, especially in cases where proteins either require unique prosthetic groups or lead to new pathways that draw further resources away from growth. Also, it's likely that additional constraints relating specific codon usage to translation efficiency (discussed in Chapter 7) will make things more interesting. The present example also illustrates that some biology important for redesign (e.g. allosteric regulation) is not yet captured in the model.

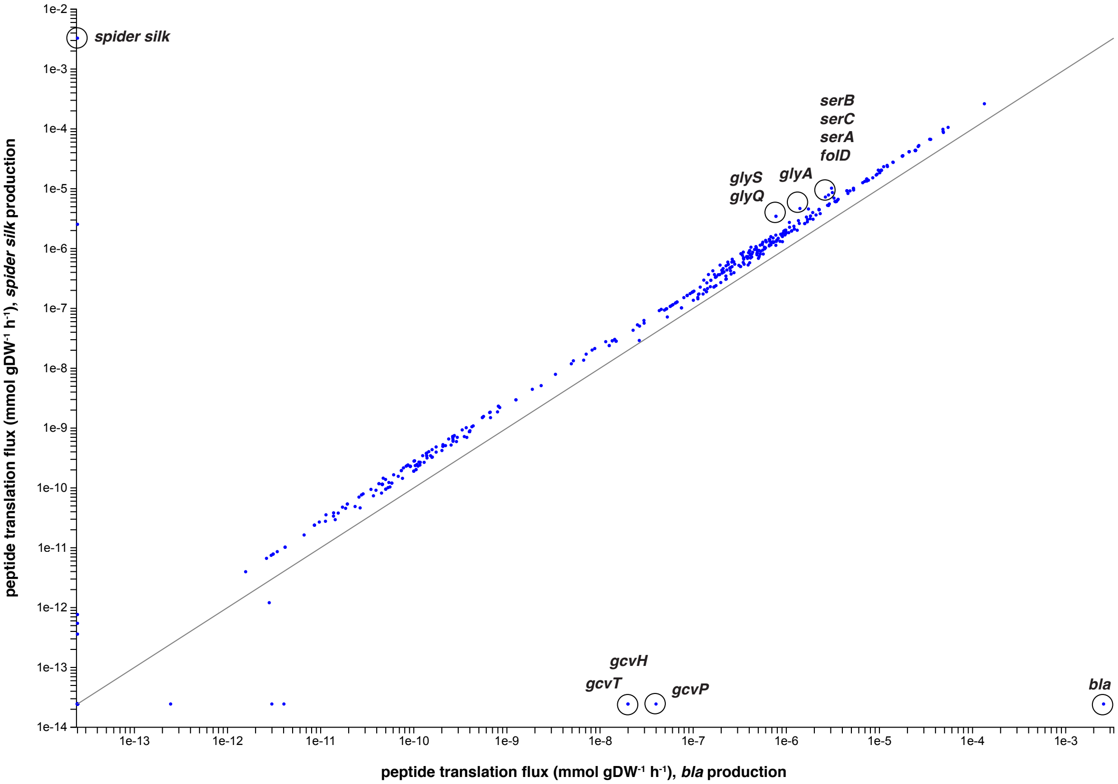


Figure 6.3: **Differential gene expression for production of *bla* vs. spider silk.** The x-axis corresponds to overexpression of *bla* with a plasmid copy number of 1000. *bla* expression is coupled to plasmid copy number as previously described. The y-axis corresponds to overexpression of the spider silk protein with a plasmid copy number of 1000. The spider silk protein is coupled with the same parameter (857 protein copies per plasmid) as with forced *bla* expression. Genes of interest are circled and discussed further in the text.

6.4 Introduction of a 2-step heterologous pathway to produce indole-3-acetaldehyde

This last specific example concerns adding a plasmid coding for a new 2-step pathway for production of indole-3-acetaldehyde (a nonnative target metabolite for *E. coli* K-12 MG1655). This compound can be produced by some *E. coli* strains, and interestingly it has been found to inhibit *E. coli* O157:H7 biofilm formation [240]. Two reactions (and their catalysts) were added to the model: (1) 2-oxoglutarate + L-trpyophan → L-glutamate + indole-3-pyruvate, and (2) indole-3-pyruvate + $H^+ \rightarrow CO_2$ + indole-3-acetaldehyde. A demand reaction was also added for indole-3-acetaldehyde so that it could leave the cell. A k_{cat} of 65 reactions per second was arbitrarily set for the first step, while a k_{cat} of 10 reactions per second was arbitrarily set for the second step. Note that it is not uncommon for one reaction in a design to have a low k_{cat} value compared to the rest of the reactions in the design (e.g. a design that depends on a promiscuous enzyme activity for one of the reactions, or a highly unoptimized synthetically designed protein). These considerations often determine whether the genes in the design should be split across multiple plasmids, since it may be beneficial for some genes to be expressed from low or high copy number plasmids. One important unknown is the amount of flux that will be siphoned away from growth by the design. For this example, I assumed the proteins would act at rates corresponding to their k_{cat} values. I imagine that much more work will go into formulating more appropriate constraints in the future. They will probably be non-linear constraints related to expression ratios (of enzymes consuming metabolites at the branch point into the nonnative pathway), and take into account as many kinetic and thermodynamic considerations as possible.

The purpose of this highly simplified example is to illustrate that the ME-Model can be used to compute the best plasmid in the design space for a given objective. To do so, the parameters of the plasmid are sampled. These parameters include: (1) the number of plasmids per cell, and (2) the number of proteins produced per plasmid copy number. As discussed previously, both of these parameters are experimentally tunable over broad ranges. As these simulations are computationally demanding for the time being, I limited the search space to the following: (1) 1 plasmid with potential copy numbers from the list [1, 5, 10, 50, 100, and 500], and (2) the number of proteins per plasmid was constrained to be an integer on the range [1, 1000]. The feasible results are shown in Figure 6.4 (note that much of the design space is infeasible).

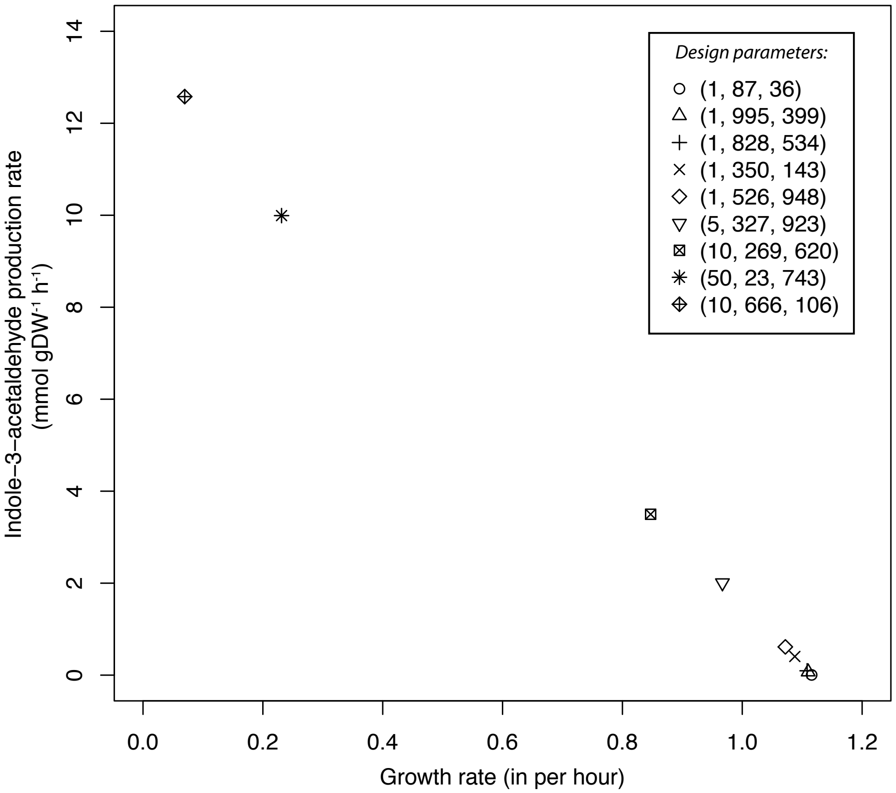


Figure 6.4: **Results from sampling the design space for production of indole-3-acetaldehyde.** The tuples in the legend correspond to (plasmid copy number, number of proteins per plasmid for the enzyme catalyzing design reaction 1, number of proteins per plasmid for the enzyme catalyzing design reaction 2). Note that 20 total designs were sampled, but only 9 produced a feasible result. No simulations with plasmid copy numbers above 50 were feasible. The space could be more deeply sampled to mine for interesting trends (e.g. the ratio of copy numbers for the enzymes catalyzing new reactions 1 and 2), but that was not done here.

Chapter 7

Conclusions and Outlook

7.1 Conclusions

In this dissertation, I develop ME-Models. ME-Models are microbial growth models that compute the optimal cellular state for growth in a given steady-state environment. They take as input the availability of nutrients to the cell and produce experimentally testable predictions for: (1) the cell's maximum growth rate (μ*) in the specified environment, (2) substrate uptake/by-product secretion rates at μ*, (3) metabolic fluxes at μ*, and (4) gene product expression levels at μ*.

ME-Models explicitly account for the production of all RNAs and proteins. In the first part of this dissertation (Chapters 2 and 3), I prototyped my approach using the simple microorganism *Thermotoga maritima*. The *T. maritima* genome had been sequenced in 1999 (one of the first!), and needed correction and complete re-annotation. We developed a framework for multi-omic data analysis that annotates genomic features involved in transcription, translation, and regulation. The genome organization of *T. maritima* displayed many distinctive properties (quantitatively) compared to other organisms. *T. maritima* has very strong promoters and RBSs (perhaps evidence of sequence-level adaptation given that recognizing and retaining contacts at the initiation sites for both transcription and translation is difficult at 80 °C?). We also hypothesized that growth at this temperature places constraints on regulatory flexibility. Importantly, the information generated was used to build the *T. maritima* ME-Model in Chapter 3. Once all the RNAs and proteins are produced, metabolism was linked to gene expression through additional constraints called 'coupling constraints.' I show these constraints extend and refine growth phenotype prediction for this organism. Specifically, I show the ME-Model for *T. maritima*: (1) has increased scope (75% of 206 functions estimated essential), (2) has a more realistic solution space in that the composition of the cell is variable, the cost of enzymes are accounted for, and detailed enzyme properties are accounted for, (3) provides more opportunity for data integration and analysis, and (4) is a useful for discovery (the examples I show are related to transcriptional regulation and gene function annotation).

In the second part of this dissertation (Chapter 4), I extended the core methods developed for *T. maritima* to *E. coli.* This was not a trivial endeavor, since there is much more information available for *E. coli* (much harder to build a complete model!). Backed by the wealth of phenotypic information available for *E. coli*, I was able to better support the statement that ME-Models provide a fundamental advance in the evolution of genome-scale biochemical models of life. I showed exactly why and how ME-Models extend and refine growth phenotype prediction.

Below, I explicitly state the improvements ME-Models bring to the table with respect to the 8 biggest weaknesses of M-Models (I stated these in Chapter 1). The biggest changes are as follows: (1) the cell composition is now a free variable, (2) energy requirements (both growth and non-growth associated) are reduced since the energy required for macromolecular synthesis is accounted for directly (though the energy requirements that remain are still fixed), (3) absolute rates (such as growth rates) can be predicted even when substrate uptake and by-product secretion rates are not specified, (4) GPRs were removed and replaced with explicit enzymes and coupling constraints, and (5) more predictions can be directly experimentally validated since transcript and protein levels can be directly compared to simulations. In addition, ME-Models incorporate some aspects of enzyme kinetics (see Chapter 4). No progress was made toward incorporating regulation (transcriptional or metabolic); however, the ME-Model can compute the need for some form of regulation by simulating 2 or more conditions. With my help, the ME-Model is now being extended to include spatial informaiton at the level of compartmentalizing network components to 10 cellular locations (mostly involving the inner and outer membranes). It will be fascinating to see if topobiological constraints (such as cellular crowding/diffusion) become relevant for constraint-based models in the years to come. The ME-Models do not provide temporal resolution, though I have been thinking about an interesting extension involving dynamic flux balance analysis. I am thinking that this will be especially interesting when modeling a diauxic growth shift (e.g. growth on glucose to growth on xylose). When growth conditions change, some proteins will no longer be useful, or even harmful for the new growth condition. The cell can only rid itself of these proteins at the rate of its growth rate. It will be interesting to see the intersection of this approach with knowledge of regulated post-translational modifications and targeted degradation of proteins (these circuits may have evolved to accelerate various shifts). Finally, ME-Models do nothing to address missing information (metabolite damage, enzyme promiscuity, and spontaneous side reactions), although I am guessing that they could be used as tools when one attempts to fill these gaps.

In the third part of this dissertation, I provide an example why it is important to consider protein and pathway cost. I demonstrate this for a *ppc* knockout of *Salmonella enterica* serovar Typhimurium. The M-Model said the organism should grow, but it did not experimentally. The *Salmonella enterica* serovar Typhimurium ME-Model was not ready at the time, so I instead relied on 'ME-style thinking' to get this project done and resolve the discrepancy (see Chapter 5 for details). The results held up in the ME-Model when I tested them later. Ultimately, the inability of cells to route flux through the glyoxylate shunt when *ppc* is removed was correctly identified by our approach as the cause of the discrepancy. Unaccounted-for regulation was at the root of the discrepancy, and so I got a glimpse of what is to come once regulation is explicitly included in these models (though to be honest I know of no good way to do this).

Finally, I discuss (and begin to demonstrate) how I believe ME-Models will serve as a bridge between systems and synthetic biology approaches for metabolic engineering.

Throughout, I tackled many behind the scenes barriers that derailed and discouraged me throughout the years. Chief among these barriers were:

1. Detailing the biochemistry of macromolecular synthesis. This is a manual process. Unlike for metabolism, the information is not sitting in a database waiting to be queried. I benefited greatly from the work of Ines Thiele [113], who came before me and detailed much of the basic reactions I relied on.

2. Omics integration is required to define various additional cellular parts and network interactions. There are no shortcuts to doing these experiments.

3. Limited software. I had to write a custom database schema, help with COBRApy [241], and create my own versions of BiGG [194] and SimPheny^{TM}, the latter of which Genomatica created to help standardize processes for M-Model creation. Both tools were insufficient for my project.

4. Representation of machinery usage constraints. It was difficult to find general constraints so as not to overtfit.

5. Commercially available linear algebra software was not good enough for my ill-conditioned constraint matrices and crazy optimization routines. This one took so long to resolve! The solutions are detailed in the supplementary material associated with Chapters 3 and 4. Luckily, I received great help from some brilliant American and German mathematicians. I also had to learn how to use resources at NERSC, supercomputers at the National Energy Research Scientific Computing Center. It took awhile to find these resources and people, and speak the highly specialized language required to get the help needed.

Ultimately, my dissertation not only contributes knowledge, but a very general approach to generate additional knowledge through computation. The entire phylogeny of approaches operating on a stoichiometric matrix (S) of an organism benefits from expansions to S (see in [242]). The methods I helped develop allow researchers of today and tomorrow to ask systems-level questions *in silico* beyond metabolism and quantitatively analyze, in a bottom-up and mechanistic manner, a variety of omics data in the context of a growing organism. As a result, I have no doubts that ME-Models will impact (at least in some small measurable way, but hopefully much more significantly) basic microbiology, the study of microbial communities, antibiotic discovery, the elucidation of regulatory networks, and systems metabolic engineering through existing and emerging synthetic biology approaches. Read the next and final section of my dissertation for my detailed 5-10 year vision of what's on the horizon.

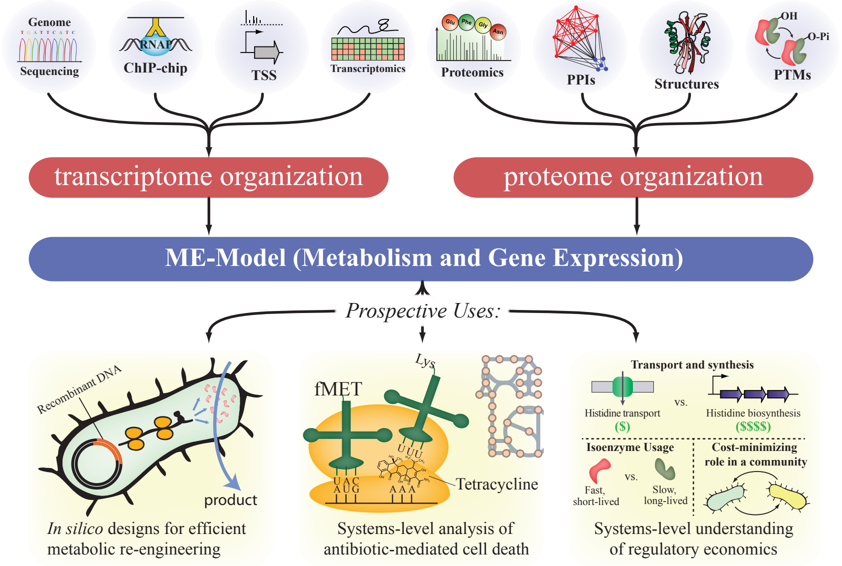


Figure 7.1: **ME-Models enable new applications of constraint-based modeling.** ME-Models afford direct integration of knowledge of organizational structures underlying the transcriptome and proteome. The transcriptome is organized into transcription units, which are determined through Genomic DNA sequencing, ChIP-chip data for the RNA polymerase, knowledge of transcription start sites (TSSs), and transcriptomics data. Transcription units are nodes in the ME-network and serve as constraints for expression of specific proteins. Knowledge of the proteome is also easily integrated. 3D structures and protein-protein interactions (PPIs) can be used to determine properties of protein complexes, which lead to constrains for production of catalysts. Post-translational modifications (PTMs), including phosphorylation, methylation, and prosthetic groups can be used to increase the resolution of the in silico proteome. Example applications enabled by our ME-Modeling approach: (1) modeling recombinant protein or metabolite overproduction, (2) modeling processes underlying antibiotic-mediated cell death, since the integrated model accounts for the majority of antibiotic targets, and (3) interpreting regulatory circuits in terms of economic efficiency. The ME-Model approximates the content of the transcriptome and proteome in the absence of regulatory constraints with failures indicative of regulatory constraints.

7.2 Outlook

7.2.1 The most promising basic uses of the *E. coli* ME-Model

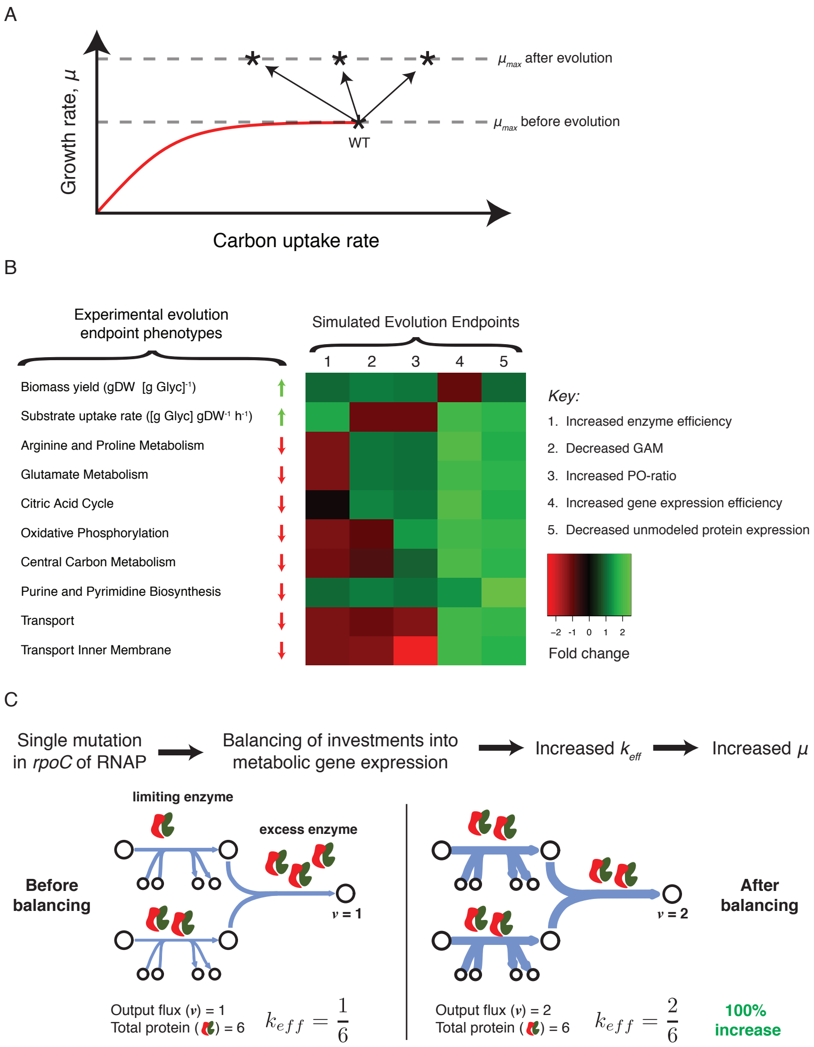
As an analysis tool for adaptive laboratory evolution (ALE) data

What follows is an excerpt from Chapter 4, where I foreshadowed the ME-Model's ability to interpret ALE data: "As it is an optimality model, the ME-Model is particularly suited for studies related to adaptive laboratory evolution (ALE). Recently, it was reported that it is not possible to predict some changes that occur during ALE in Batch culture using an M-Model [192]. This is because M-Models only take biomass yield optimization into account; these results are consistent with the rate-yield trade-offs present in the ME-Model under nutrient-excess conditions. In the ME-Model, a number of inherent factors can limit cellular growth (e.g., translation rate and metabolic catalysis); the ME-Model can thus provide alternative hypotheses for the mechanisms of growth increase and aid in understanding the results of ALE."

Here, I reduce this to practice and show you exactly what I mean. I removed this material from Chapter 4 because the analysis I'm about to present is severely underdeveloped. I considered not including it in my dissertation at all, but I think the concept the analysis exposes is an important one to keep in mind when analyzing ALE data with the ME-Model, and so I have this material here in this outlook section.

The ME-Model can simulate various mechanisms of growth increase through evolution. A few key parameters determine the maximum growth rate in the ME-Model, so *in silico* and *in vivo* phenotypic changes can be compared to understand systems-level mechanisms of growth increase (Figure 7.2A). I use *E. coli* grown in glycerol batch culture as an illustrative example. When evolved in excess glycerol, mutations in *rpoC* lead to large changes in gene expression [243, 244]. We compare *in silico* changes in substrate uptake rate, biomass yield, and expression of cellular subsystems to measurements from evolved strains. We find that increasing the effective catalytic rate of enzymes in the ME-Model results in phenotypic changes that closely match with experiments (Figure 7.2B). Increasing the average catalytic rate of metabolic enzymes results in increases in glucose uptake and growth yield, and decreases in expression for a number of subsystems (Figure 7.2B). The ME-Model thus provides a systems-level hypothesis for the mechanism of evolution in glycerol: The altered gene expression caused by the mutated RNA polymerase results in a rebalancing of the proteome that increases the average flux per enzyme (Figure 7.2C). Other simulated mechanisms do not have such a close agreement with the data (Figure 7.2B, all other columns). Interestingly, there is experimental evidence that the global k_{eff} increases as a function of growth rate, but this data is limited to transitions between nutrient-limited and batch growth conditions [170]. It remains to be seen whether this is a general strategy that can be used during evolution in batch growth conditions.

Figure 7.2: **Evolution to higher growth rate in Batch culture by increase in whole-cell enzyme effective catalytic rate (k_{eff}).** A) The ME-Model was used to analyze cells evolving in glycerol in batch culture conditions. Many *in silico* evolutionary trajectories are possible. B) Evolution results in changes in biomass yield, substrate uptake rate, and the differential expression of genes in the subsystems listed (see Experimental Methods in [158]) [244]. The directionality of the change during evolution is shown with arrows. We simulated five different global parameters that affect the maximum growth rate achievable in ME-Model simulations. For each parameter, changes in the identified phenotypes are calculated after a change in the parameter that would increase the maximum growth rate in the ME-Model. The fold change of subsystems in the ME-Model is calculated based on the change in the fractional proteome mass of all genes in that subsystem. Increasing k_{eff} produces results most consistent with experimental data. C) Simulation results combined with gene expression and physiological data from wild-type and evolved strains support an increase in whole-cell k_{eff}. *In vivo*, the increase in k_{eff} is likely achieved by balancing investments into metabolic gene expression to achieve the maximal growth rate.


A
Growth rate, μ
Carbon uptake rate
μmax after evolution
μmax before evolution
WT
B
Experimental evolution endpoint phenotypes
Simulated Evolution Endpoints
1
2
3
4
5
Biomass yield (gDW [g Glyc]-1)
Substrate uptake rate ([g Glyc] gDW-1 h-1)
Arginine and Proline Metabolism
Glutamate Metabolism
Citric Acid Cycle
Oxidative Phosphorylation
Central Carbon Metabolism
Purine and Pyrimidine Biosynthesis
Transport
Transport Inner Membrane
Key:
1. Increased enzyme efficiency
2. Decreased GAM
3. Increased PO-ratio
4. Increased gene expression efficiency
5. Decreased unmodeled protein expression
−2 -1 0 1 2
Fold change
C
Single mutation in rpoC of RNAP
Balancing of investments into metabolic gene expression
Increased keff
Increased μ
limiting enzyme
excess enzyme
Before balancing
v = 1
Output flux (v) = 1
Total protein ( ) = 6
keff = 1/6
After balancing
v = 2
Output flux (v) = 2
Total protein ( ) = 6
keff = 2/6
100% increase

This type of analysis was previously infeasible (or indirect at best) as genome-scale metabolic model. M-Models cannot predict changes in nutrient uptake and gene expression, and do not include the proteomic parameters of the ME-Model. Much more work is needed to verify the predictions the ME-Model makes. One issue confounding the analysis of ALE data is that abundance of a molecule does not always correlate with its activity in the cell. It is highly likely that the data fitting procedure described below can be used to circumvent these problems.

As a tool to probe new regulatory functions

Regulatory constraints and interactions are beyond the scope of the ME-Model. It may be fruitful to determine if the patterns the ME-Model predicts are consistent with our knowledge of transcriptional regulation through the action of transcription factors (TFs). One could compare hundreds of *in silico* expression shifts and check the predictions against the analogous *in vivo* shifts. It should be kept in mind that many incorrect predictions will be as a result of the fact that suboptimal control of gene expression is widespread in bacteria [245]. To rectify the failure modes (where learning might take place), I would stray away from adding regulators and regulatory rules to fix the failure modes (as this is just fitting, and I feel an entirely new modeling paradigm is required to describe and understand regulatory processes). Instead, I'd focus on identifying the trade-offs the regulatory network evolved to help the cell optimize itself for growth. ME-Models will only be able to make statements about steady-state growth conditions, so regulation related to dynamics cannot be considered using this approach. Detailed knowledge of the regulatory network topology can be gained by integrating ChIP-exo binding data with RNA-seq expression profiles from transcription factor deletion strains. Such transcriptional regulatory network reconstructions serve as one starting point for finding missing constraints.

As a tool to back-calculate hard-to-measure parameters and model non-optimal growth

As they are, ME-Models contain many unknown parameters. The most important unknowns relate to translation efficiency (# of proteins per mRNA) and effective catalytic rates, or k_{eff} parameters. Data sources are rapidly coming online to determine translation efficiency (e.g., ribosome profiling, and quantitative transcriptomics paired with quantitative proteomics). Upon first inspection, it is clear that the # of proteins per mRNA seems to be conserved across conditions (Pearson's r between 0.9 and 0.97 for growth on different carbon sources in minimal media) [personal communication with Ali Ebrahim]. I therefore assume these parameters can one day soon be imposed to yield fairly general predictions for new conditions of interest. That leaves the unknown effective catalytic rate parameters as the biggest and most important unknown parameter set. Interestingly, most enzymes seem to operate 'moderately efficient' in the cell; the distribution of k_{cat} parameters is log-normal and centered on approximately 10 s^{-1} [246]. This perhaps underlies the success of the ME-Model at predicting basic microbial growth phenotypes despite setting k_{eff} parameters using crude assumptions. But nailing down these values is particularly important for predicting absolute gene expression levels, something the ME-Model does a pretty poor job of.

A promising emerging approach is to fit the data to the model and back-calculate a k_{eff} distribution most consistent or compatible with the data. The challenges here are that the data are noisy/incomplete, and the distribution of k_{eff} values may not be unique. That said, preliminary results are highly promising [personal communication with Edward O'Brien]. Once refined, this method will represent a big step up from approaches such as GIMME [109]. Our refined method will exploit the structure of the ME-Model to constrain the range of parameter values. This procedure can be repeated for many different experimental conditions, and the parameter set will become more predictive over time (for example k_{cat} can be taken as the highest k_{eff} value observed for any fit growth condition). Getting this approach to work will be a highly rewarding line of basic research. It should also have big applied implications. For example, it would be very useful to metabolic engineers to be able to compute the effect of a small perturbation away from a very well-characterized and parameterized base condition. Stay tuned for major updates along these lines.

As a tool for modeling simple microbial communities

M- and ME-Models now exist for multiple organisms, so an intriguing potential application is using the models for analysis of relationships between different organisms. Although much is known about these relationships, the use of M- and ME-Models will allow a much more systematic, detailed, and complete description and will give researchers the ability to model and perturb the relationships *in silico*. Numerous challenges confront investigators of the interface and interactions between two metabolic networks. Exchanges of metabolites between the networks can be difficult to model since they often must pass through an unpredictable external environment. Often, the networks are under separate systems of regulatory control. Generally, both networks can exist outside of the interaction under study and thus must maintain some independence of the other network. Therefore, many of the assumptions used in the reconstruction of metabolic networks of individual cells do not necessarily hold for interactions between two networks. For this reason, models describing complex communities have failed to appear.

A potential research avenue that escapes many of these concerns but still provides insight into network interactions and interfaces is the analysis of interactions between different strains of the same species, many of which are known to produce or consume distinct metabolites. ME-Models can help model both competitive or syntrophic relationships among these strains, natural phenomena particularly well-suited for constraint-based modeling.

In the future, I imagine high-throughput expression profiling technologies (combined with the data fitting procedure previously described) can help constrain individual genome-scale models to capture growth parameters and community composition of simple bacterial communities over time. ME-Models are particularly well-suited to capture the metabolic load of communication (signaling) and transfer of metabolites (and even electrons) between species.

As a tool for studying translation (especially tRNA properties)

ME-Models will likely prove useful for investigating the impact of cellular tRNA pool perturbations on global protein expression. As a result of the immense cost of protein synthesis, the process has been highly regulated and tuned during the course of evolution. Regulation occurs at the level of translation initiation, translation elongation, and translation termination.

Translation initiation rates are a function of many physical features. These features include the number of available free ribosomes, the folding energy of the 5' UTR, the beginning of the coding sequence, and the base pairing potential between the 5' UTR and the 16S ribosomal rRNA. These features are extremely hard to quantify, so the constraints underlying translation elongation have been the traditional focus. One argument for focusing on translation elongation is that relative initiation rates can perhaps be inferred based on elongation pressures (a cell would unlikely make a transcript that it would not subsequently translate). Elongation rates have been shown to be highly dependent on codon order and composition of the particular mRNA being translated [186]. The degeneracy of the genetic code (61 codons coding for just 20 amino acids) combined with the relative efficiency of anticodons in reading certain codons during protein synthesis, leads to an enormous amount of flexibility in the process. Organisms have optimized the throughput of the process due to selective pressures to enhance translation efficiency, especially in highly expressed proteins such as the ribosomal proteins.

ME-Models have the ability to incorporate factors determining translation efficiency at the systems-level, and costs of maintaining translational throughput are explicitly modeled. Currently, the constraints bounding translation efficiency in the ME-Model are biologically unrealistic. For example, one bound the burst size (proteins per mRNA), can be computed using: (1) the maximum amino acid incorporation rate, (2) the size of the ribosome footprint in nucleotides, and (3) the mean lifetime of an mRNA molecule. The mRNA can then be assumed to be completely saturated with ribosomes. In reality, this situation cannot occur due to collisions between ribosomes, so we set the burst size to an arbitrary number that was lower, but more realistic.

In the future, a ribosome flow model can be used to more tightly bound the burst size. A ribosome flow model is a simple, probabilistic, physically plausible computational model of ribosome progression across an mRNA that is solely based on the coding sequence [247]. Ribosomes advance probabilistically according to cellular tRNA concentrations and cannot pass each other (i.e. a downstream ribosome can prevent an upstream ribosome from progressing if it does not clear a portion of the mRNA fast enough). These models are likely not ready for integration with the ME-Model at this time, but one day soon they will be. Experimental methods to quantify the cellular tRNA concentrations must be improved.

Although many of the molecular details of translation are part of the ME-Model, constraints of this sort have not been previously integrated into stoichiometric models for use with flux balance analysis. The integration is expected to define a much more biologically relevant reduced solution space for ME-Models, given the enormous importance of translation with respect to the cellular economy. A model with an understanding of constraints underlying translation would allow one to pre-compute the effects of targeted changes to the cellular tRNA pool. These perturbations could take on many forms. The model can first be prototyped using simple perturbations such as the complete removal of a certain tRNA gene or the addition of a new tRNA gene. The model could then be used to help formulate arguments for the presence of tRNA genes in phage genomes, which are under evolutionary pressures to be small and compact [248]. The presence of tRNA genes in phage genomes is further evidence of the importance of translation elongation constraints. As phages and metabolic engineers face similar challenges [249], I imagine the constraints revealed by such an analysis could potentially be fundamental with respect to metabolic engineering, which usually involves expressing a pathway for metabolite overproduction and tinkering the cell to withstand heterologous protein expression. This possibility can be explored using expression of a peptide (from a plasmid) designed to deplete one or more tRNA types. This depletion phenomenon has been observed in industrial applications [250]. For example, spider silk is rich in glycine, so it imposes unique metabolic and macromolecular synthesis constraints on the cell to support the overproduction. With respect to metabolite overproduction, tRNA pool perturbations could be used to shift the state of the cell's proteome to favor the pathways relevant to the overproduction. It is likely the specific applications will be refined as the modeling progresses.

7.2.2 The most promising applied uses of the *E. coli* ME-Model

As a tool to power the *in silico* portion of the iterative metabolic engineering loop

ME-Models will be used to close the design-build-test loop that is currently taking the metabolic engineering industry by storm. I purposefully omit many details here due to intellectual property and licensing concerns. I feel ME-Models will be particularly useful as *context for content*, in that data can be mapped and analyzed in the context of the model. The data fitting techniques described above will be key. The trick will be making the data actionable. Metabolic engineers care about better understanding the system, but at the end of the day a decision needs to be made about how to best improve the current strain design, leading to the next design. Once methods are built up more, I believe ME-Models will become a tool of choice for bridging synthetic and systems biology approaches to metabolic engineering.

As a tool to probe or design antibiotic functions in the context of the larger system

ME-Models could improve network-based drug target identification, a prerequisite for rational drug development, in two key ways.

The first key improvement: As ME-Models explicitly account for the costs of enzyme expression and dilution to daughter cells, the most efficient growth simulations will try to minimize the materials required to assemble the cell; i.e., ME-Models will efficiently use enzymes when simulating growth at a specified rate. After genetic or chemical perturbations, a cell may lose its ability to satisfy growth demands efficiently. As we will see, this possibility can only be directly assessed using a ME-Model.

It has been previously noted that organisms make efficient use of their enzymes to maintain a minimal total flux through their biochemical network. In M-Models, compensatory pathways, regardless of complexity, can be used to support growth. This is mirrored in the results seen in Chapter 5. In M-Models, an enzyme may carry infinite fluxes, unless v_{max} constraints are imposed, and pathways carried out by simple monomeric enzymes are equivalent to longer pathways supported by complex multimeric proteins with expensive cofactor requirements. ME-Models rectify these problems, allowing for cost-conscious, semi-quantitative forecasts of the ability to grow after genetic and/or chemical perturbations. If compensatory pathways are more expensive than the pathways lost, the ME-Model prediction will indicate slower growth.

The second key improvement: ME-Models contain a majority (about 80% for the *E. coli* ME-Model) of the 206 functions estimated as essential for a minimal organism, whereas M-Models contain approximately 30% of these core functions. With the ME-Model, many of these functions are essential for growth and ribosome production. This broader coverage of cellular functions inherently increases the ability to ME-Models to predict and interpret phenotypic states over M-Models. This increased scope allows for insight into many more potential targets. Interestingly, many antibiotics target proteins involved in macromolecular synthesis. Of particular interest now is the ME-Model's capability to predict synergy between an existing target and all metabolic enzymes. The ME-Model can simulate complete removal of the protein, or the coupling constraint controlling the protein's efficiency can be tuned to simulate partial inhibition. This ability was successfully demonstrated with the simulation of different bulk translational efficiencies for the ribosome in the *T. maritima* ME-Model (see Chapter 3).

7.2.3 Automating the construction of a ME-Model for a bacteria of your choice

I'm sure this is going to happen. In this dissertation, I focused on *T. maritima* and *E. coli*. These microbes are both gram-negative, but besides that, they are as different as two microbes get. A similar approach to Model SEED [251] can be taken building off of my work, and I think in this case it will be more successful since the macromolecular synthesis machinery and reactions are more conserved than your average metabolic machinery and functions. The specifics of the constraints will be tricky, particularly when considering that ribosomal capacity and protein synthetic efficiency must be approximated, but are probably highly variable among microbes. A model for *Staphylococcus aureus* would be of particular interest due to its relevance to human health.

7.2.4 ME-Models for Yeast and Humans

Best of luck with this. Detailed versions of these models analogous to the *E. coli* ME-Model will require sustained efforts lasting many years, but perhaps coarse-grained versions could be built much faster.

7.2.5 Roadmap to a steady-state whole-cell *E. coli* model

There's a lot to do with ME-Models as they are today (see above sections). In my opinion, its more pressing to explore these applications than to continue expanding the model. But we are rapidly moving towards a whole-cell stoichiometric model for *E. coli*, and so here I lay out my vision for how this is realized. First, I'd start with the 40 functions estimated as essential for a minimal organism that are not present in the current model (note: integrating protein translocation and secretion is underway at the time of writing). This list can be found by sorting the 'in ME-Model?' column of Supplementary Table S3B in [158].

Next, I'd move on to integrating the functions of the proteins that compose the highest unaccounted-for fraction of the wild-type proteome (by mass) when growing in M9 glucose medium. Finally, it will be a major challenge to mechanistically incorporate stress responses and maintenance energy expenditures. This will move us one step closer to being able to model stationary (no growth) phase, during which expression of mRNAs ceases (mostly), and the cell acts as a 'sac of enzymes.' Life is complex, but it is possible to sketch a molecular description from simple, but stoichiometrically accurate, chemical equations represented on a genome scale.

Bibliography

[1] Famili I, Förster J, Nielsen J, Palsson BO (2003) Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. Proceedings of the National Academy of Sciences 100: 13134-13139.

[2] Thiele I, Palsson BO (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protocols 5: 93–121.

[3] McCloskey D, Palsson BO, Feist AM (2013) Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. Mol Syst Biol 9.

[4] Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. Mol Syst Biol 3: 119.

[5] Schuetz R, Zamboni N, Zampieri M, Heinemann M, Sauer U (2012) Multidimensional optimality of microbial metabolism. Science 336: 601–604.

[6] Joyce AR, Palsson BO (2006) The model organism as a system: integrating 'omics' data sets. Nat Rev Mol Cell Biol 7: 198–210.

[7] Palsson B (2000) The challenges of in silico biology. Nat Biotech 18: 1147–1150.

[8] Jacob F (1977) Evolution and tinkering. Science 196: 1161–1166.

[9] Pirt SJ (1965) The maintenance energy of bacteria in growing cultures. Proc R Soc Lond B Biol Sci 163: 224–231.

[10] Bailey JE, Ollis DF (1976) Biochemical engineering fundamentals. Chemical Engineering Education .

[11] Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T (2010) Interdependence of cell growth and gene expression: origins and consequences. Science 330: 1099–1102.

[12] Klumpp S, Hwa T (2008) Growth-rate-dependent partitioning of RNA polymerases in bacteria. Proc Natl Acad Sci USA 105: 20245–20250.

[13] Klumpp S, Zhang Z, Hwa T (2009) Growth rate-dependent global effects on gene expression in bacteria. Cell 139: 1366–1375.

[14] Kim M, Zhang Z, Okano H, Yan D, Groisman A, et al. (2012) Need-based activation of ammonium uptake in Escherichia coli. Mol Syst Biol 8: 616.

[15] Klumpp S, Scott M, Pedersen S, Hwa T (2013) Molecular crowding limits translation and cell growth. Proceedings of the National Academy of Sciences .

[16] You C, Okano H, Hui S, Zhang Z, Kim M, et al. (2013) Coordination of bacterial proteome with metabolism by cyclic AMP signalling. Nature advance online publication.

[17] Price ND, Reed JL, Palsson BO (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Micro 2: 886–897.

[18] Kitano H (2002) Systems biology: a brief overview. Science 295: 1662–1664.

[19] Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO (2009) Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol 7: 129–143.

[20] Reed JL, Famili I, Thiele I, Palsson BO (2006) Towards multidimensional genome annotation. Nat Rev Genet 7: 130–141.

[21] Overbeek R, Bartels D, Vonstein V, Meyer F (2007) Annotation of bacterial and archaeal genomes: improving accuracy and consistency. Chem Rev 107: 3431–3447.

[22] Guell M, van Noort V, Yus E, Chen WH, Leigh-Bell J, et al. (2009) Transcriptome complexity in a genome-reduced bacterium. Science 326: 1268–1271.

[23] Kuhner S, van Noort V, Betts MJ, Leo-Macias A, Batisse C, et al. (2009) Proteome organization in a genome-reduced bacterium. Science 326: 1235–1240.

[24] Qiu Y, Cho BK, Park YS, Lovley D, Palsson BO, et al. (2010) Structural and operational complexity of the Geobacter sulfurreducens genome. Genome Res 20: 1304–1311.

[25] Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, et al. (2010) The primary transcriptome of the major human pathogen Helicobacter pylori. Nature 464: 250–255.

[26] Yoon SH, Reiss DJ, Bare JC, Tenenbaum D, Pan M, et al. (2011) Parallel evolution of transcriptome architecture during genome reorganization. Genome Res 21: 1892–1904.

[27] Buescher JM, Liebermeister W, Jules M, Uhr M, Muntel J, et al. (2012) Global network reorganization during dynamic adaptations of Bacillus subtilis metabolism. Science 335: 1099–1103.

[28] Nicolas P, Mader U, Dervyn E, Rochat T, Leduc A, et al. (2012) Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. Science 335: 1103–1106.

[29] Sorek R, Cossart P (2010) Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. Nat Rev Genet 11: 9–16.

[30] Palsson B, Zengler K (2010) The challenges of integrating multi-omic data sets. Nature Chemical Biology 6: 787–789.

[31] Huber R, Langworthy TA, Konig H, Thomm M, Woese CR, et al. (1986) Thermotoga maritima sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90 C. Archives of Microbiology 144: 324–333.

[32] Dipippo JL, Nesbo CL, Dahle H, Doolittle WF, Birkland NK, et al. (2009) Kosmotoga olearia gen. nov., sp. nov., a thermophilic, anaerobic heterotroph isolated from an oil production fluid. Int J Syst Evol Microbiol 59: 2991–3000.

[33] Nesbo CL, Dlutek M, Zhaxybayeva O, Doolittle WF (2006) Evidence for existence of "mesotogas," members of the order Thermotogales adapted to low-temperature environments. Appl Environ Microbiol 72: 5061–5068.

[34] Nesbo CL, Kumaraswamy R, Dlutek M, Doolittle WF, Foght J (2010) Searching for mesophilic Thermotogales bacteria: "mesotogas" in the wild. Appl Environ Microbiol 76: 4896–4900.

[35] Nesbo CL, Bradnan DM, Adebusuyi A, Dlutek M, Petrus AK, et al. (2012) Mesotoga prima gen. nov., sp. nov., the first described mesophilic species of the Thermotogales. Extremophiles 16: 387–393.

[36] Zhaxybayeva O, Swithers KS, Foght J, Green AG, Bruce D, et al. (2012) Genome Sequence of the Mesophilic Thermotogales Bacterium Mesotoga prima MesG1.Ag.4.2 Reveals the Largest Thermotogales Genome To Date. Genome Biol Evol 4: 700–708.

[37] Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, et al. (2005) Genome streamlining in a cosmopolitan oceanic bacterium. Science 309: 1242–1245.

[38] Nelson KE (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature 399: 323–329.

[39] Conners SB, Mongodin EF, Johnson MR, Montero CI, Nelson KE, et al. (2006) Microbial biochemistry, physiology, and biotechnology of hyperthermophilic Thermotoga species. FEMS Microbiol Rev 30: 872–905.

[40] Mongodin EF, Hance IR, Deboy RT, Gill SR, Daugherty S, et al. (2005) Gene transfer and genome plasticity in Thermotoga maritima, a model hyperthermophilic species. Journal of bacteriology 187: 4935–4944.

[41] Zhaxybayeva O, Swithers KS, Lapierre P, Fournier GP, Bickhart DM, et al. (2009) On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. Proceedings of the National Academy of Sciences of the United States of America 106: 5865–5870.

[42] Martin W, Baross J, Kelley D, Russell MJ (2008) Hydrothermal vents and the origin of life. Nature reviews Microbiology 6: 805–814.

[43] Achenbach-Richter L, Gupta R, Stetter KO, Woese CR (1987) Were the original eubacteria thermophiles? Systematic and applied microbiology 9: 34–39.

[44] Munoz R, Yarza P, Ludwig W, Euzeby J, Amann R, et al. (2011) Release LTPs104 of the All-Species Living Tree. Systematic and applied microbiology 34: 169–170.

[45] Fields PA (2001) Review: Protein function at thermal extremes: balancing stability and flexibility. Comp Biochem Physiol A Mol Integr Physiol 129: 417–431.

[46] Kumar S, Nussinov R (2001) How do thermophilic proteins deal with heat? Cell Mol Life Sci 58: 1216–1233.

[47] Gerday C, Glansdorff N, American Society for Microbiology CN - Jefferson or Adams Building Reading Rooms QR1009; P59 2007 Reference - Science Reading Room (Adams tFQP (2007) Physiology and biochemistry of extremophiles. Washington, D.C.: ASM Press, xvi, 429 p. pp.

[48] Robb FTCNJoABRRQT (2008) Thermophiles : biology and technology at high temperatures. Boca Raton, FL: CRC Press, xiii, 353 p. pp.

[49] Boucher N, Noll KM (2011) Ligands of thermophilic ABC transporters encoded in a newly sequenced genomic region of Thermotoga maritima MSB8 screened by differential scanning fluorimetry. Appl Environ Microbiol 77: 6395–6399.

[50] Aziz RK (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genom 9: 75.

[51] Latif H, Lerman JA, Portnoy VA, Tarasova Y, Nagarajan H, et al. (2013) The Genome Organization of Thermotoga maritima Reflects Its Lifestyle. PLoS Genet 9: e1003485.

[52] Cho BK, Zengler K, Qiu Y, Park YS, Knight EM, et al. (2009) The transcription unit architecture of the Escherichia coli genome. Nat Biotechnol 27: 1043–1049.

[53] Vijayan V, Jain IH, O'Shea EK (2011) A high resolution map of a cyanobacterial transcriptome. Genome Biol 12: R47.

[54] Koide T, Reiss DJ, Bare JC, Pang WL, Facciotti MT, et al. (2009) Prevalence of transcription promoters within archaeal operons and coding sequences. Molecular systems biology 5: 285.

[55] Gruber AR, Lorenz R, Bernhart SH, Neubock R, Hofacker IL (2008) The Vienna RNA websuite. Nucleic Acids Res 36: W70–4.

[56] Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. Bioinformatics 25: 1335–1337.

[57] Ross W, Gosink KK, Salomon J, Igarashi K, Zou C, et al. (1993) A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. Science 262: 1407–1413.

[58] Blatter EE, Ross W, Tang H, Gourse RL, Ebright RH (1994) Domain organization of RNA polymerase alpha subunit: C-terminal 85 amino acids constitute a domain capable of dimerization and DNA binding. Cell 78: 889–896.

[59] Schneider TD (1996). New Approaches in Mathematical Biology: Information Theory and Molecular Machines.

[60] Schneider TD (1997) Information content of individual genetic sequences. J Theor Biol 189: 427–441.

[61] D'Haeseleer P (2006) What are DNA sequence motifs? Nat Biotechnol 24: 423–425.

[62] Schneider TD (1991) Theory of molecular machines. II. Energy dissipation from molecular machines. J Theor Biol 148: 125–137.

[63] Shultzaberger RK, Roberts LR, Lyakhov IG, Sidorov IA, Stephen AG, et al. (2007) Correlation between binding rate constants and individual information of E. coli Fis binding sites. Nucleic Acids Res 35: 5275–5283.

[64] Rhodius VA, Mutalik VK (2010) Predicting strength and function for promoters of the Escherichia coli alternative sigma factor, sigmaE. Proc Natl Acad Sci U S A 107: 2854–2859.

[65] Keseler IM, Collado-Vides J, Santos-Zavaleta A, Peralta-Gil M, Gama-Castro S, et al. (2011) EcoCyc: a comprehensive database of Escherichia coli biology. Nucleic Acids Res 39: D583–90.

[66] Kroger C, Dillon SC, Cameron AD, Papenfort K, Sivasankaran SK, et al. (2012) The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. Proc Natl Acad Sci U S A 109: E1277–86.

[67] Albrecht M, Sharma CM, Dittrich MT, Muller T, Reinhardt R, et al. (2011) The transcriptional landscape of Chlamydia pneumoniae. Genome Biol 12: R98.

[68] Mitschke J, Georg J, Scholz I, Sharma CM, Dienst D, et al. (2011) An experimentally anchored map of transcriptional start sites in the model cyanobacterium Synechocystis sp. PCC6803. Proc Natl Acad Sci U S A 108: 2124–2129.

[69] Sierro N, Makita Y, de Hoon M, Nakai K (2008) DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. Nucleic Acids Res 36: D93–6.

[70] Chen H, Bjerknes M, Kumar R, Jay E (1994) Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of Escherichia coli mRNAs. Nucleic Acids Res 22: 4953–4957.

[71] Molina N, van Nimwegen E (2008) Universal patterns of purifying selection at noncoding positions in bacteria. Genome Res 18: 148–160.

[72] Nelson CE, Hersh BM, Carroll SB (2004) The regulatory content of intergenic DNA shapes genome architecture. Genome Biol 5: R25.

[73] Novichkov PS, Laikova ON, Novichkova ES, Gelfand MS, Arkin AP, et al. (2010) RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes. Nucleic Acids Res 38: D111–8.

[74] Maier T, Schmidt A, Guell M, Kuhner S, Gavin AC, et al. (2011) Quantification of mRNA and protein and integration with protein turnover in a bacterium. Mol Syst Biol 7: 511.

[75] Nie L, Wu G, Zhang W (2006) Correlation between mRNA and protein abundance in Desulfovibrio vulgaris: a multiple regression to identify sources of variations. Biochemical and biophysical research communications 339: 603–610.

[76] Towsey M, Hogan JM, Mathews S, Timms P (2007) The in silico prediction of promoters in bacterial genomes. Genome Inform 19: 178–189.

[77] Rangannan V, Bansal M (2011) PromBase: a web resource for various genomic features and predicted promoters in prokaryotic genomes. BMC Res Notes 4: 257.

[78] Gerland U, Moroz JD, Hwa T (2002) Physical constraints and functional characteristics of transcription factor-DNA interaction. Proceedings of the National Academy of Sciences of the United States of America 99: 12015–12020.

[79] Galtier N, Lobry JR (1997) Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J Mol Evol 44: 632–636.

[80] Frock AD, Gray SR, Kelly RM (2012) Hyperthermophilic Thermotoga species differ with respect to specific carbohydrate transporters and glycoside hydrolases. Appl Environ Microbiol 78: 1978–1986.

[81] Darfeuille F, Unoson C, Vogel J, Wagner EG (2007) An antisense RNA inhibits translation by competing with standby ribosomes. Molecular cell 26: 381–392.

[82] Waters LS, Storz G (2009) Regulatory RNAs in bacteria. Cell 136: 615–628.

[83] Rinker KD, Kelly RM (1996) Growth physiology of the hyperthermophilic Archaeon Thermococcus litoralis: development of a sulfur-free defined medium, characterization of an exopolysaccharide, and evidence of biofilm formation. Appl Environ Microbiol 62: 4478–4485.

[84] Portnoy VA, Herrgard MJ, Palsson BO (2008) Aerobic fermentation of D-glucose by an evolved cytochrome oxidase-deficient Escherichia coli strain. Applied and environmental microbiology 74: 7561–7569.

[85] Pysz MA, Ward DE, Shockley KR, Montero CI, Conners SB, et al. (2004) Transcriptional analysis of dynamic heat-shock response by the hyperthermophilic bacterium Thermotoga maritima. Extremophiles 8: 209–217.

[86] Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, et al. (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. Nat Methods 6: 550–551.

[87] Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821–829.

[88] Halasz G, van Batenburg MF, Perusse J, Hua S, Lu XJ, et al. (2006) Detecting transcriptionally active regions using genomic tiling arrays. Genome Biol 7: R59.

[89] Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, et al. (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nature methods 7: 709–715.

[90] Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.

[91] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511–515.

[92] Schrimpe-Rutledge AC, Jones MB, Chauhan S, Purvine SO, Sanford JA, et al. (2012) Comparative Omics-Driven Genome Annotation Refinement: Application across Yersiniae. PLoS One 7: e33903.

[93] Lerman JA, Hyduke DR, Latif H, Portnoy VA, Lewis NE, et al. (2012) In silico method for modelling metabolism and gene product expression at genome scale. Nat Commun 3: 929.

[94] Liu X, Brutlag DL, Liu JS (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing : 127–138.

[95] Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology 2: 28–36.

[96] Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14: 1188–1190.

[97] Markham NR, Zuker M (2008) UNAFold: software for nucleic acid folding and hybridization. Methods in molecular biology 453: 3–31.

[98] Takemoto K, Nacher JC, Akutsu T (2007) Correlation between structure and temperature in prokaryotic metabolic networks. BMC Bioinformatics 8: 303.

[99] Kingsford CL, Ayanbule K, Salzberg SL (2007) Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. Genome Biol 8: R22.

[100] Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, et al. (2011) Rfam: Wikipedia, clans and the ”decimal” release. Nucleic Acids Res 39: D141–5.

[101] Brenner S (2010) Sequences and consequences. Philosophical transactions of the Royal Society of London Series B, Biological sciences 365: 207–12.

[102] Otero JM, Nielsen J (2010) Industrial systems biology. Biotechnology and Bioengineering 105: 439–460.

[103] Mahadevan R, Palsson BO, Lovley DR (2011) In situ to in silico and back: elucidating the physiology and ecology of Geobacter spp. using genome-scale modelling. Nature Reviews Microbiology 9: 39–50.

[104] Feist AM, Palsson BO (2008) The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. Nature Biotechnology 26: 659–67.

[105] Oberhardt MA, Palsson BO, Papin JA (2009) Applications of genome-scale metabolic reconstructions. Mol Syst Biol 5: 320.

[106] Reed JL, Palsson BO (2004) Genome-scale in silico models of E. coli have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. Genome Res 14: 1797–1805.

[107] Schellenberger J, Lewis NE, Palsson BO (2011) Elimination of thermodynamically infeasible loops in steady-state metabolic models. Biophysical Journal 100: 544–553.

[108] Akesson M, Forster J, Nielsen J (2004) Integration of gene expression data into genome-scale metabolic models. Metab Eng 6: 285–293.

[109] Becker SA, Palsson BO (2008) Context-specific metabolic networks are consistent with experiments. PLoS Comput Biol 4: e1000082.

[110] Colijn C (2009) Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. PLoS Comput Biol 5: e1000489.

[111] Shlomi T, Cabili MN, Herrgard MJ, Palsson BO, Ruppin E (2008) Network-based prediction of human tissue-specific metabolism. Nat Biotechnol 26: 1003–1010.

[112] Allen TE, Palsson BO (2003) Sequence-based analysis of metabolic demands for protein synthesis in prokaryotes. J Theor Biol 220: 1–18.

[113] Thiele I (2008) No Title. Dissertation: A Stoichiometric Model of Escherichia coli's Macromolecular Synthesis Machinery and its Integration with Metabolism .

[114] Schröder C, Selig M, Schönheit P (1994) Glucose fermentation to acetate, CO&lt;sub&gt;2&lt;/sub&gt; and H&lt;sub&gt;2&lt;/sub&gt; in the anaerobic hyperthermophilic eubacterium &lt;i&gt;Thermotoga maritima&lt;/i&gt;: involvement of the Embden-Meyerhof pathway. Archives of Microbiology 161: 460–470.

[115] Zhang Y (2009) Three-dimensional structural view of the central metabolic network of Thermotoga maritima. Science 325: 1544–1549.

[116] Kummerfeld SK, Teichmann SA (2006) DBD: a transcription factor prediction database. Nucleic Acids Res 34: D74–81.

[117] Andrianantoandro E, Basu S, Karig DK, Weiss R (2006) Synthetic biology: new engineering rules for an emerging discipline. Mol Syst Biol 2: 2006.0028.

[118] Vickers CE, Blank LM, Kromer JO (2010) Grand challenge commentary: Chassis cells for industrial biochemical production. Nat Chem Biol 6: 875–877.

[119] Schaechter M, Maaloe O, Kjeldgaard NO (1958) Dependency on medium and temperature of cell size and chemical composition during balanced grown of Salmonella typhimurium. J Gen Microbiol 19: 592–606.

[120] Feist AM, Palsson BO (2010) The biomass objective function. Curr Opin Microbiol 13: 344–349.

[121] Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? Nat Biotechnol 28: 245–248.

[122] Applegate DL, Cook W, Dash S, Espinoza DG (2007) Exact solutions to linear programming problems. Operations Res Lett 35: 693–699.

[123] Gupta RS, Schlessinger D (1976) Coupling of rates of transcription, translation, and messenger ribonucleic acid degradation in streptomycin-dependent mutants of Escherichia coli. J Bacteriol 125: 84–93.

[124] Thiele I, Jamshidi N, Fleming RM, Palsson BO (2009) Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. PLoS Comput Biol 5: e1000312.

[125] Holzhutter HG (2004) The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. Eur J Biochem 271: 2905–2922.

[126] Pramanik J, Keasling JD (1997) Stoichiometric model of Escherichia coli metabolism: incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. Biotechnol Bioeng 56: 398–421.

[127] Lewis NE (2010) Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Mol Syst Biol 6: 390.

[128] Gil R, Silva FJ, Pereto J, Moya A (2004) Determination of the core of a minimal bacterial gene set. Microbiol Mol Biol Rev 68: 518–537.

[129] Browning DF, Busby SJ (2004) The regulation of bacterial transcription initiation. Nat Rev Microbiol 2: 57–65.

[130] Bailey TL (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37: W202–8.

[131] Franco IS, Mota LJ, Soares CM, de Sa-Nogueira I (2007) Probing key DNA contacts in AraR-mediated transcriptional repression of the Bacillus subtilis arabinose regulon. Nucleic Acids Res 35: 4755–4766.

[132] Miwa Y, Nakata A, Ogiwara A, Yamamoto M, Fujita Y (2000) Evaluation and characterization of catabolite-responsive elements (cre) of Bacillus subtilis. Nucleic Acids Res 28: 1206–1210.

[133] Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30.

[134] Dennis PP (1974) In vivo stability, maturation and relative differential synthesis rates of individual ribosomal proteins in Escherichia coli B/r. J Mol Biol 88: 25–41.

[135] Singer P, Nomura M (1985) Stability of ribosomal protein mRNA and translational feedback regulation in Escherichia coli. Mol Gen Genet 199: 543–546.

[136] Ji H, Liu XS (2010) Analyzing 'omics data using hierarchical models. Nat Biotechnol 28: 337–340.

[137] Canales RD (2006) Evaluation of DNA microarray results with quantitative gene expression platforms. Nat Biotechnol 24: 1115–1122.

[138] Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270: 467–470.

[139] Kharchenko P, Vitkup D, Church GM (2004) Filling gaps in a metabolic network using expression information. Bioinformatics 20: i178–85.

[140] Sabatti C, Rohlin L, Oh MK, Liao JC (2002) Co-expression pattern from DNA microarray experiments as a tool for operon prediction. Nucleic Acids Res 30: 2886–2893.

[141] Rhodius VA, LaRossa RA (2003) Uses and pitfalls of microarrays for studying transcriptional regulation. Curr Opin Microbiol 6: 114–119.

[142] Crick F (1973) Project K: The Complete Solution of E. coli. Perspect Biol Med 17: 67–70.

[143] Overbeek R (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 33: 5691–5702.

[144] Wu CH (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 34: D187.

[145] Rose PW (2011) The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res 39: D392–401.

[146] Juhling F (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. Nucleic Acids Res 37: D159.

[147] Tong KL, Wong JT (2004) Anticodon and wobble evolution. Gene 333: 169–177.

[148] Mandal N, Mangroo D, Dalluge JJ, McCloskey JA, Rajbhandary UL (1996) Role of the three consecutive G:C base pairs conserved in the anticodon stem of initiator tRNAs in initiation of protein synthesis in Escherichia coli. RNA 2: 473–482.

[149] Guymon R, Pomerantz SC, Ison JN, Crain PF, McCloskey JA (2007) Post-transcriptional modifications in the small subunit ribosomal RNA from Thermotoga maritima, including presence of a novel modified cytidine. RNA 13: 396–403.

[150] Szymanski M, Barciszewska MZ, Erdmann VA, Barciszewski J (2002) 5S Ribosomal RNA Database. Nucleic Acids Res 30: 176–178.

[151] Larkin MA (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948.

[152] Gustilo EM, Vendeix FA, Agris PF (2008) tRNA's modifications bring order to gene expression. Curr Opin Microbiol 11: 134–140.

[153] Selinger DW, Wright MA, Church GM (2003) On the complete determination of biological systems. Trends Biotechnol 21: 251–254.

[154] Machado D, Costa R, Rocha M, Ferreira E, Tidor B, et al. (2011) Modeling formalisms in Systems Biology. AMB Expr 1: 1–14.

[155] Thiele I, Fleming RM, Que R, Bordbar A, Diep D, et al. (2012) Multiscale modeling of metabolism and macromolecular synthesis in E. coli and its application to the evolution of codon usage. PLoS ONE 7: e45635.

[156] Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, et al. (2012) A whole-cell computational model predicts phenotype from genotype. Cell 150: 389–401.

[157] Orth JD, Conrad TM, Na J, Lerman JA, Nam H, et al. (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism2011. Molecular Systems Biology 7: 535.

[158] O/'Brien EJ, Lerman JA, Chang RL, Hyduke DR, Palsson BO (2013) Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol Syst Biol 9.

[159] Riley M, Abe T, Arnaud MB, Berlyn MK, Blattner FR, et al. (2006) Escherichia coli K-12: a cooperatively developed annotation snapshot–2005. Nucleic Acids Res 34: 1–9.

[160] Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, et al. (2013) EcoCyc: fusing model organism databases with systems biology. Nucleic Acids Res 41: D605–D612.

[161] Kato J, Hashimoto M (2007) Construction of consecutive deletions of the Escherichia coli chromosome. Mol Syst Biol 3: 132.

[162] Donachie WD, Robinson AC (1987) Cell division: parameter values and the process. Escherichia coli and Salmonella typhimurium Cellular and Molecular Biology Vol. I and II: 1578–1593.

[163] Meyenburg KV, Hansen FG (1987) Regulation of chromosome replication. Escherichia coli and Salmonella typhimurium Cellular and Molecular Biology Vol. I and II: 1555–1577.

[164] Bremer H, Dennis PP (1996) Modulation of chemical composition and other parameters of the cell by growth rate. Escherichia coli and Salmonella : 1553–1569.

[165] Neijssel OM, Mattos M, Tempest DW (1996) Growth Yield and Energy Distribution. Escherichia coli and Salmonella : 1683–1692.

[166] Zhuang K, Vemuri GN, Mahadevan R (2011) Economics of membrane occupancy and respiro-fermentation. Mol Syst Biol 7: 500.

[167] Thiele I, Fleming RM, Bordbar A, Schellenberger J, Palsson BO (2010) Functional characterization of alternate optimal solutions of Escherichia coli/'s transcriptional and translational machinery. Biophys J 98: 2072–2081.

[168] Young R, Bremer H (1976) Polypeptide-chain-elongation rate in Escherichia coli B/r as a function of growth rate. Biochem J 160: 185.

[169] Proshkin S, Rahmouni AR, Mironov A, Nudler E (2010) Cooperation between translating ribosomes and RNA polymerase in transcription elongation. Science 328: 504–508.

[170] Valgepea K, Adamberg K, Seiman A, Vilu R (2013) Escherichia coli achieves faster growth by increasing catalytic and translation rates of proteins. Mol Biosyst 9: 2344–2358.

[171] Boer VM, Crutchfield CA, Bradley PH, Botstein D, Rabinowitz JD (2010) Growth-limiting intracellular metabolites in yeast growing under diverse nutrient limitations. Mol Biol Cell 21: 198–211.

[172] O/'Brien RW, Neijssel OM, Tempest DW (1980) Glucose phosphoenolpyruvate phosphotransferase activity and glucose uptake rate of Klebsiella aerogenes growing in chemostat culture. J Gen Microbiol 116: 305–314.

[173] Smith RW, Dean AC (1972) Beta-galactosidase synthesis in Klebsiella aerogenes growing in continuous culture. J Gen Microbiol 72: 37–47.

[174] Molenaar D, van Berlo R, de Ridder D, Teusink B (2009) Shifts in growth strategies reflect tradeoffs in cellular economics. Mol Syst Biol 5: 323.

[175] Monod J (1949) The growth of bacterial cultures. Annu Rev Microbiol 3: 371–394.

[176] Koch AL (1997) Microbial physiology and ecology of slow growth. Microbiol Mol Biol Rev 61: 305–318.

[177] Beg QK, Vazquez A, Ernst J, de Menezes MA, Bar-Joseph Z, et al. (2007) Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity. Proc Natl Acad Sci USA 104: 12663–12668.

[178] Adadi R, Volkmer B, Milo R, Heinemann M, Shlomi T (2012) Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. PLoS Comput Biol 8: e1002575.

[179] Button DK (1991) Biochemical basis for whole-cell uptake kinetics: specific affinity, oligotrophic capacity, and the meaning of the Michaelis constant. Appl Environ Microbiol 57: 2033–2038.

[180] Hua Q, Yang C, Oshima T, Mori H, Shimizu K (2004) Analysis of gene expression in escherichia coli in response to changes of growth-limiting nutrient in chemostat cultures. Applied and environmental microbiology 70: 2354–2366.

[181] Nanchen A, Schicker A, Sauer U (2006) Nonlinear dependency of intracellular fluxes on growth rate in miniaturized continuous cultures of Escherichia coli. Appl Environ Microbiol 72: 1164–1172.

[182] Vemuri GN, Altman E, Sangurdekar DP, Khodursky AB, Eiteman MA (2006) Overflow metabolism in Escherichia coli during steady-state growth: transcriptional regulation and effect of the redox ratio. Appl Environ Microbiol 72: 3653–3661.

[183] Nahku R, Valgepea K, Lahtvee PJ, Erm S, Abner K, et al. (2010) Specific growth rate dependent transcriptome profiling of Escherichia coli [bsol]{K12[bsol]} [bsol]{MG1655[bsol]} in accelerostat cultures. J Biotechnol 145: 60–65.

[184] Bennett BD, Kimball EH, Gao M, Osterhout R, Van Dien SJ, et al. (2009) Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. Nat Chem Biol 5: 593–599.

[185] Li GW, Oh E, Weissman JS (2012) The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature 484: 538–541.

[186] Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, et al. (2010) An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell 141: 344–354.

[187] Cho BK, Federowicz SA, Embree M, Park YS, Kim D, et al. (2011) The PurR regulon in Escherichia coli K-12 MG1655. Nucleic Acids Res 39: 6456–6464.

[188] Berthoumieux S, de Jong H, Baptist G, Pinel C, Ranquet C, et al. (2013) Shared control of gene expression in bacteria by transcription factors and global physiology of the cell. Mol Syst Biol 9: 634.

[189] Gerosa L, Kochanowski K, Heinemann M, Sauer U (2013) Dissecting specific and global transcriptional regulation of bacterial gene expression. Mol Syst Biol 9: 658.

[190] Cho BK, Knight EM, Palsson BO (2006) Transcriptional regulation of the fad regulon genes of Escherichia coli by ArcA. Microbiology 152: 2207–2219.

[191] Haverkorn van Rijsewijk BR, Nanchen A, Nallet S, Kleijn RJ, Sauer U (2011) Large-scale 13C-flux analysis reveals distinct transcriptional control of respiratory and fermentative metabolism in Escherichia coli. Mol Syst Biol 7: 477.

[192] Harcombe WR, Delaney NF, Leiby N, Klitgord N, Marx CJ (2013) The ability of flux balance analysis to predict evolution of central metabolism scales with the initial distance to the optimum. PLoS Comput Biol 9: e1003091.

[193] Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9: 559.

[194] Schellenberger J, Park JO, Conrad TM, Palsson BO (2010). BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions.

[195] Kim TY, Sohn SB, Kim YB, Kim WJ, Lee SY (2012). Recent advances in reconstruction and applications of genome-scale metabolic models.

[196] Perez Pulido R, Ben Omar N, Abriouel H, Lucas Lopez R, Martinez Canamero M, et al. (2005). Microbiological study of lactic acid fermentation of Caper berries by molecular and culture-dependent methods.

[197] Park JM, Kim TY, Lee SY (2011). Genome-scale reconstruction and in silico analysis of the Ralstonia eutropha H16 for polyhydroxyalkanoate synthesis, lithoautotrophic growth, and 2-methyl citric acid production.

[198] Licona-Cassani C, Marcellin E, Quek LE, Jacob S, Nielsen LK (2012). Reconstruction of the Saccharopolyspora erythraea genome-scale model and its use for enhancing erythromycin production.

[199] Almaas E, Oltvai ZN, Barabási AL (2005) The Activity Reaction Core and Plasticity of Metabolic Networks. PLoS Computational Biology 1: 7.

[200] Lee SJ, Lee DY, Kim TY, Kim BH, Lee J, et al. (2005). Metabolic engineering of Escherichia coli for enhanced production of succinic acid, based on genome comparison and in silico gene knockout simulation.

[201] Nam H, Lewis NE, Lerman JA, Lee DH, Chang RL, et al. (2012). Network context and selection in the evolution to enzyme specificity.

[202] Chen L, Vitkup D (2006) Predicting genes for orphan metabolic activities using phylogenetic profiles. Genome Biology 7: R17.

[203] Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, et al. (2006). Systems approach to refining genome annotation.

[204] Frezza C, Zheng L, Folger O, Rajagopalan KN, MacKenzie ED, et al. (2011). Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase.

[205] Becker D, Selbach M, Rollenhagen C, Ballmaier M, Meyer TF, et al. (2006). Robust Salmonella metabolism limits possibilities for new antimicrobials.

[206] Brynildsen MP, Winkler JA, Spina CS, MacDonald IC, Collins JJ (2013). Potentiating antibacterial activity by predictably enhancing endogenous microbial ROS production.

[207] Chang RL, Xie L, Xie L, Bourne PE, Palsson BO (2010) Drug Off-Target Effects Predicted Using Structural Analysis in the Context of a Metabolic Network Model. PLoS Computational Biology 6: 18.

[208] Chandrasekaran S, Price ND (2010). Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis.

[209] van Berlo RJ, de Ridder D, Daran JM, Daran-Lapujade PA, Teusink B, et al. (2011). Predicting metabolic fluxes using gene expression differences as constraints.

[210] Orth JD, Palsson BO (2010). Systematizing the generation of missing metabolic knowledge.

[211] Orth JD, Palsson B (2012). Gap-filling analysis of the iJO1366 Escherichia coli metabolic network reconstruction for discovery of metabolic functions.

[212] Fuhrer T, Chen L, Sauer U, Vitkup D (2007). Computational prediction and experimental verification of the gene encoding the NAD+/NADP+-dependent succinate semialdehyde dehydrogenase in Escherichia coli.

[213] Vitkin E, Shlomi T (2012). MIRAGE: a functional genomics-based approach for metabolic network model reconstruction and its application to cyanobacteria networks.

[214] Thiele I, Hyduke DR, Steeb B, Fankam G, Allen DK, et al. (2011). A community effort towards a knowledge-base and mathematical model of the human pathogen Salmonella Typhimurium LT2.

[215] Datsenko KA, Wanner BL (2000). One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products.

[216] Schellenberger J, Que R, Fleming RM, Thiele I, Orth JD, et al. (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0.

[217] Herring CD, Glasner JD, Blattner FR (2003). Gene replacement without selection: regulated suppression of amber mutations in Escherichia coli.

[218] Fong NL, Lerman JA, Lam I, Palsson BO, Charusanti P (2013) Reconciling a salmonella enterica metabolic model with experimental data confirms that overexpression of the glyoxylate shunt can rescue a lethal ppc deletion mutant. FEMS Microbiology Letters 342: 62–69.

[219] Fong SS, Nanchen A, Palsson BO, Sauer U (2006). Latent pathway activation and increased pathway capacity enable Escherichia coli adaptation to loss of key metabolic enzymes.

[220] Peng L, Arauzo-Bravo MJ, Shimizu K (2004). Metabolic flux analysis for a ppc mutant Escherichia coli based on 13C-labelling experiments together with enzyme activity assays and intracellular metabolite measurements.

[221] Lorca GL, Ezersky A, Lunin VV, Walker JR, Altamentova S, et al. (2007). Glyoxylate and pyruvate are antagonistic effectors of the Escherichia coli IclR transcriptional regulator.

[222] Ehrt S, Guo XV, Hickey CM, Ryou M, Monteleone M, et al. (2005). Controlling gene expression in mycobacteria with anhydrotetracycline and Tet repressor.

[223] Cortay JC, Bleicher F, Rieul C, Reeves HC, Cozzone AJ (1988). Nucleotide sequence and expression of the aceK gene coding for isocitrate dehydrogenase kinase/phosphatase in Escherichia coli.

[224] Chung T, Resnik E, Stueland C, LaPorte DC (1993). Relative expression of the products of glyoxylate bypass operon: contributions of transcription and translation.

[225] Cozzone AJ, El-Mansi M (2005). Control of isocitrate dehydrogenase catalytic activity by protein phosphorylation in Escherichia coli.

[226] Higgins CF, Ames GF, Barnes WM, Clement JM, Hofnung M (1982). A novel intercistronic regulatory element of prokaryotic operons.

[227] Chung T, Klumpp DJ, LaPorte DC (1988). Glyoxylate bypass operon of Escherichia coli: cloning and determination of the functional map.

[228] Brewster RC, Jones DL, Phillips R (2012) Tuning promoter strength through rna polymerase binding site design in ¡italic¿escherichia coli¡/italic¿. PLoS Comput Biol 8: e1002811.

[229] Salis HM, Mirsky EA, Voigt CA (2009) Automated design of synthetic ribosome binding sites to control protein expression. Nat Biotech 27: 946–950.

[230] Kosuri S, Goodman DB, Cambray G, Mutalik VK, Gao Y, et al. (2013) Composability of regulatory sequences controlling transcription and translation in escherichia coli. Proceedings of the National Academy of Sciences .

[231] Temme K, Zhao D, Voigt CA (2012) Refactoring the nitrogen fixation gene cluster from klebsiella oxytoca. Proceedings of the National Academy of Sciences 109: 7085-7090.

[232] Shao Z, Rao G, Li C, Abil Z, Luo Y, et al. (0) Refactoring the silent spectinabilin gene cluster using a plug-and-play scaffold. ACS Synthetic Biology 0: null.

[233] Carlson R (2009) The changing economics of DNA synthesis. Nat Biotech 27: 1091–1094.

[234] Yanisch-Perron C, Vieira J, Messing J (1985) Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. Gene 33: 103–119.

[235] Cunningham D, Koepsel R, Ataai M, Domach M (2009) Factors affecting plasmid production in escherichia coli from a resource allocation standpoint. Microbial Cell Factories 8: 27.

[236] Rozkov A, Avignone-Rossa C, Ertl P, Jones P, O'Kennedy R, et al. (2004) Characterization of the metabolic burden on escherichia coli dh1 cells imposed by the presence of a plasmid containing a gene therapy sequence. Biotechnology and Bioengineering 88: 909–915.

[237] Bentley WE, Mirjalili N, Andersen DC, Davis RH, Kompala DS (1990) Plasmid-encoded protein: The principal factor in the metabolic burden associated with recombinant bacteria. Biotechnology and Bioengineering 35: 668–681.

[238] Panayotatos N (1988) Recombinant protein production with minimal-antibiotic-resistance vectors. Gene 74: 357–363.

[239] Xia XX, Qian ZG, Ki CS, Park YH, Kaplan DL, et al. (2010) Native-sized recombinant spider silk protein produced in metabolically engineered escherichia coli results in a strong fiber. Proceedings of the National Academy of Sciences 107: 14059-14063.

[240] Lee JH, Kim YG, Kim CJ, Lee JC, Cho M, et al. (2012) Indole-3-acetaldehyde from rhodococcus sp. bfi 332 inhibits escherichia coli o157:h7 biofilm formation. Applied Microbiology and Biotechnology 96: 1071-1078.

[241] Ebrahim A, Lerman Ja, Palsson BO, Hyduke DR (2013) COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC systems biology 7: 74.

[242] Lewis NE, Nagarajan H, Palsson BO (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. Nat Rev Micro 10: 291–305.

[243] Herring CD, Raghunathan A, Honisch C, Patel T, Applebee MK, et al. (2006) Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. Nat Genet 38: 1406–1412.

[244] Conrad TM, Frazier M, Joyce AR, Cho BK, Knight EM, et al. (2010) Rna polymerase mutants found through adaptive evolution reprogram escherichia coli for optimal growth in minimal media. Proceedings of the National Academy of Sciences 107: 20500-20505.

[245] Price MN, Deutschbauer AM, Skerker JM, Wetmore KM, Ruths T, et al. (2013) Indirect and suboptimal control of gene expression is widespread in bacteria. Molecular systems biology 9: 660.

[246] Bar-Even A, Noor E, Savir Y, Liebermeister W, Davidi D, et al. (2011) The moderately efficient enzyme: Evolutionary and physicochemical trends shaping enzyme parameters. Biochemistry 50: 4402-4410.

[247] Reuveni S, Meilijson I, Kupiec M, Ruppin E, Tuller T (2011) Genome-scale analysis of translation elongation with a ribosome flow model. PLoS Comput Biol 7: e1002127.

[248] Bailly-Bechet M, Vergassola M, Rocha E (2007) Causes for the intriguing presence of trnas in phages. Genome Research 17: 1486-1495.

[249] Maynard ND, Gutschow MV, Birch EW, Covert MW (2010) The virus as metabolic engineer. Biotechnology Journal 5: 686–694.

[250] Widmaier DM, Tullman-Ercek D, Mirsky EA, Hill R, Govindarajan S, et al. (2009) Engineering the Salmonella type III secretion system to export spider silk monomers. Mol Syst Biol 5.

[251] Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, et al. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. Nature biotechnology 28: 977–982.