Final Report

DOE Award DE-FG02-09ER25917
to Stanford University

Numerical Optimization Algorithms and Software for Systems Biology

PI: Michael Saunders

Date of report: February 3, 2013
Period covered by report: Sept 15, 2009 – Sept 14, 2012

Primary contractor

STANFORD UNIVERSITY
Stanford, CA 94305

Michael A. Saunders (PI)
Yinyu Ye (Co-PI)
Dept of Management Science and Engineering
475 Via Ortega, CA 94305-4121
650-723-1875 650-723-7262
saunders@stanford.edu yinyu-ye@stanford.edu

Subcontractor 1

UNIVERSITY OF CALIFORNIA, SAN DIEGO (UCSD)
9500 Gilman Drive, La Jolla, CA 92093

Bernhard Ø. Palsson (PI)
Department of Bioengineering
9500 Gilman Drive, La Jolla, CA 92093-0412
858-534-5668
palsson@ucsd.edu

Subcontractor 2

UNIVERSITY OF ICELAND
101 Reykjavik, Iceland

Ines Thiele (PI)
Center for Systems Biology
Sturlugata 8, IS-101 Reykjavik, Iceland
+354-618-6246
ithiele@hi.is

---

1 Progress during the 3-year funding period

Below we summarize the progress achieved on the specific aims stated in the proposal to DOE. A new website `http://www.stanford.edu/group/SOL/multiscale` records these activities. It will continue to document the Stanford/UCSD/Iceland collaboration now funded by NIH and DOE 2012–2017.

Aim 1 *Develop reliable algorithms for solving optimization problems involving large stoichiometric matrices.*

1(a) *Extend existing sparse linear programming algorithms to enable the solution of such systems, in which the matrix coefficients represent reactions at multiple timescales and thus vary over many orders of magnitude.*

Flux balance analysis (FBA) leads to large *multiscale* linear programs with constraints of the form $Sv = b$ and $l \le v \le u$, where the variables v are steady state reaction rates (fluxes). The matrix values S_{ij} and solution values v_j both vary over a wide range (hence multiscale). Large S_{ij} values arise when constraints of the form $\alpha \le v_1/v_2 \le \beta$ are included (with β large). Linear programming solvers typically scale the problem, solve, and then unscale the solution. Unfortunately the unscaled solution may not satisfy the constraints accurately.

Reformulation via lifting. To remove the need for scaling, PhD student Yuekai Sun developed a "chaining" reformulation in which a constraint $v_1/v_2 \le 10^6$ is replaced by a sequence of constraints $v_1 \le 100s_1$, $s_1 \le 100s_2$, $s_2 \le 100v_2$ with the help of auxiliary variables s_1, s_2. We now call this a "lifting technique" by analogy with the "lifted Newton method" of Albersmeyer and Diehl (2010), in which systems of nonlinear equations are more easily solved by introducing additional equations and variables (though the motivation is different).

For large FBA models, the lifting technique provides significant benefits for standard linear programming solvers such as CPLEX, Gurobi, and our own SQOPT. Our manuscript Sun *et al.* (2012) has been reviewed and revised for *BMC Systems Biology*. It gives numerical results for a genome-scale (77000×68000) reconstruction of *E. coli* metabolism and expression developed by Thiele *et al.* (2012); see Aim (2a) below.

Exact rational simplex method. Separately, PhD student Joshua Lerman at UCSD has explored the use of an "exact simplex solver" on an 18000×17500 reconstruction of *Thermotoga maritima* (Aim (2b) below). This requires many hours of computation (and would not be practical for the *E. coli* model), but it provides valuable benchmark information by which to judge our other approaches.

Quad precision simplex method. In collaboration with Prof Philip Gill and Dr Elizabeth Wong of UC San Diego (Mathematics Dept), we are developing a Fortran 2003 implementation of our sparse linear and nonlinear optimization solvers SQOPT and SNOPT. PI Saunders and PhD student Martina Ma have converted this version of SQOPT to use quadruple precision floating point, and have applied it to the *T.*

maritima and *E. coli* models and also to many standard linear programming test problems. Strangely, we have achieved far greater accuracy than expected.

Normally with double precision floating point (15-digit arithmetic) we ask for only 6-digit precision in the solution (i.e., satisfying the feasibility and optimality conditions to 6 digits), and this may not be achievable without the lifting reformulation. With quad precision (32-digit arithmetic) we asked for "only" 15-digit precision in the solution because the input data is less accurate than that. Unexpectedly in *all cases* we obtained *30-digit precision* in the solution (without the aid of lifting). We have yet to explain why, but this is certainly a welcome empirical result.

Although quad-SQOPT is significantly slower than standard SQOPT, it is much faster than the "exact simplex" solver. In practice we can run standard SQOPT first (ideally with lifting), save the solution, and then warm-start quad-SQOPT. The *T. maritima* and *E. coli* models solve to exceptional accuracy this way in only a few seconds after the warm start.

1(b) *Develop a convex optimization algorithm for computing thermodynamically feasible reaction fluxes in a general instance of a genome-scale integrated metabolic and macromolecular biosynthetic network.*

The Fleming *et al.* (2012) article "A variational principle for computing nonequilibrium fluxes and potentials in genome-scale biochemical networks" presents for the first time an eminently scalable convex optimization extension of Flux Balance Analysis. We derive an analytically elegant formulation of steady-state mass conservation, energy conservation, and the second law of thermodynamics. We also demonstrate for the first time the exact way in which reaction flux is a dual variable to the familiar notion of chemical potential. This foundational article provides a scalable method for predicting thermodynamically consistent steady-state fluxes and chemical potentials for an arbitrarily large biochemical network.

1(c) *Implement the convex optimizer in an embarrassingly parallel manner to enable massive sampling of the thermodynamically feasible set.*

In 2008 when this aim was written, we originally envisaged the development of a parallel algorithm for random sampling of the thermodynamically feasible set. During the effort to complete Aim 1(b), we observed that it might be possible develop a globally convergent algorithm to conduct gradient-based search within the thermodynamically feasible set to additionally satisfy reaction kinetic constraints, over and above thermodynamic and steady state constraints. We also observed that by adding reaction kinetic constraints, the absolute concentration of molecular species would be predicted, allowing us to satisfy Aim 3(a). In essence, thermodynamically constrained non-equilibrium steady states are only predictive of the "relative" concentration of molecular species within a reaction. For details, see the discussion in Fleming *et al.* (2012). As we know that all biochemical reactions must satisfy some form of reaction kinetic constraint, we decided it would not be fruitful to sample all thermodynamically feasible steady states, as a vanishing minority of such samples would also (by chance) satisfy reaction kinetic constraints. Therefore, our focus shifted toward gradient-based search within the thermodynamically feasible set to satisfy reaction kinetic constraints also. The successful outcome of this work is detailed under Aim 3(a).

1(d) *Disseminate platform-independent industrial-quality software to the systems biology community.*

COBRA Toolbox v2: Dr Ronan Fleming (Iceland) is co-administrator of the COnstraint-Based Reconsruction and Analysis (COBRA) toolbox, the second version of which was released as an open source project on 1 January 2011 (`opencobra.sourceforge.net`). This widely used open source toolbox is detailed in the *Nature Protocol* article by Schellenberger *et al.* (2011). Users of the toolbox are supported via the online forum `groups.google.com/group/cobra-toolbox`.

COBRApy: Joshua Lerman (UCSD) has submitted a manuscript entitled *COBRApy: COnstraints-Based Reconstruction and Analysis for Python.* This documents open source Python code for Constraint-Based Reconstruction and Analysis, complementing the MATLAB code that is already available. `opencobra.sourceforge.net`.

von Bertalanffy 1.0: We published an algorithmic pipeline for quantitative assignment of reaction directionality in multicompartmental genome scale models based on an application of the second law of thermodynamics to each reaction. Given experimental or computationally estimated standard metabolite species Gibbs energy and metabolite concentrations, the algorithm bounds reaction Gibbs energy, which is transformed to in vivo pH, temperature, ionic strength and electrical potential. This cross platform MATLAB extension to the COBRA toolbox is computationally efficient, extensively documented and open source: `opencobra.sourceforge.net`.

fastFVA: An open source implementation of flux variability analysis (`notendur.hi.is/ithiele/downloads.html`). This efficient implementation makes large-scale flux variability analysis (FVA) feasible and tractable allowing more complex biological questions regarding network flexibility and robustness to be addressed. Networks involving thousands of biochemical reactions can be analyzed within seconds, greatly expanding the utility of flux variability analysis in systems biology.

rBioNet: A COBRA toolbox extension for reconstructing high-quality biochemical networks. rBioNet enables the construction of publication-level biochemical networks while enforcing necessary quality control and assurance measures based on the Nature Protocol detailing the reconstruction process (Thiele and Palsson, 2010, *Nature Protocols*). rBioNet has an intuitive user interface facilitating the reconstruction process for novices and experts. rBioNet was developed to assist the most recent metabolic reconstruction of *Synechocystis* spp (Aim 3(c)). `doi:10.1093/bioinformatics/btr308`.

LUSOL interface: A MATLAB interface to our sparse LU factorization package LUSOL was developed by PhD student Nick Henderson. LUSOL is the only sparse-matrix package that can determine the rank and null space of a large stoichiometric matrix S in a numerically reliable and efficient way. The interface makes use of MATLAB's new class structures. We have used it to implement new MATLAB functions for representing a nullspace operator Z (such that $SZ = 0$) and computing products Zv and $Z^T w$ for arbitrary vectors v and w.
`http://www.stanford.edu/group/SOL/software/lusol.html`.

LSMR: For sparse least-squares problems $\min \|Ax - b\|$, where A is a sparse matrix or a linear operator. MATLAB, Fortran 90, and python implementations.
`http://www.stanford.edu/group/SOL/software/lsmr.html`.

LSRN: Another least-squares solver for the case where A is a dense or sparse matrix or a linear operator and is extremely large in one dimension (but the other dimension is only a few thousand). A may also be rank-deficient. Parallel python and C++ implementations.
`http://www.stanford.edu/group/SOL/software/lsrn.html`.

PDCO: (Primal-Dual Convex Optimizer) A MATLAB package for large-scale convex optimization with linear constraints (where the constraint matrix may be a linear operator). Incorporates LSMR.
`http://www.stanford.edu/group/SOL/software/pdco.html`.

PNOPT: (Proximal Newton Optimizer) A MATLAB package for minimizing composite functions using proximal Newton-type methods.
`http://www.stanford.edu/group/SOL/software/pnopt.html`.

Aim 2 *Investigate cyclic dependency between metabolic and macromolecular biosynthetic networks.*

2(a) *Predict the material and energy cost of macromolecular synthesis in an integrated metabolic, transcriptional and translational model of Escherichia coli.*

The integrated metabolic, transcriptional and translational model of *Escherichia coli* (Thiele, 2009, Thesis) was employed to investigate the connection between genotype-environment-phenotype (Thiele *et al.*, *PLoS ONE*, 2012). This stoichiometric metabolic, transcriptional and translational model of *E. coli* MG1655 integrates its metabolic (Feist *et al.*, 2008, *MSB*) and macromolecular synthesis machinery networks (Thiele *et al.*, 2009, *PLoS Comp Biol*). It accounts for the synthesis, assembly, and function of 1,827 protein coding and 110 RNA coding genes corresponding to over 40% of the entire genome at single nucleotide resolution. This model of metabolism and gene expression (ME-matrix) encompasses many cellular functions detailed in 76,589 reactions and 62,212 components. Many of the ME-matrix genes are highly conserved in enterobacter and non-enterobacter species. This conservation is particularly interesting as the ME-matrix accounts for all major antibiotic targets except DNA gyrase. It could be exploited for functionally assessing lethal or sub-lethal antibiotic doses and combination therapies of novel antibiotic substances.

Metabolism provides precursors for the macromolecular synthesis machinery, which in turn synthesizes the metabolic enzymes that catalyze biochemical reactions. An approximation of this interdependency results in linear inequalities that tightly constrain quasi-steady state reaction rates over many orders of magnitude (i.e., $mmol.g_{DW}^{-1}.hr^{-1}$ vs. $nmol.g_{DW}^{-1}.hr^{-1}$). The stoichiometric coefficients in the ME-matrix are distributed over four orders of magnitude, because many precursors are required to form one macromolecule. We explicitly included enzymes into metabolic reactions; thus, additional constraints were needed to ensure that the catalyzing enzyme is synthesized within the network when its metabolic reaction is used (as described in Thiele *et al.*, 2010, *Biophys J*). We added linear coupling constraints to ensure that if an enzyme's biosynthetic flux is zero then its utilization flux is zero. Additionally, if a utilization reaction carries a high flux then the biosynthetic flux also needs to be higher (Thiele *et al.*, 2010, *Biophys J*). The ME-matrix model is amenable to flux balance analysis (FBA), the linear optimization of a biologically motivated objective (e.g., growth rate) subject to constraints (e.g., environmental conditions).

A key interest of systems biology is to develop a mechanistic basis for the genotype-phenotype relationship. The ME-matrix explicitly captures the nucleotide sequence for over 1900 genes and stoichiometrically represents their cellular functions, so we asked if and how codon usage bias (CUB) (genotype) evolves to maximize growth rate (phenotype) in different growth environments. We generated a range of perturbed ME-matrices differing only in codon usage from the wildtype ME-matrix. We found that *in silico* changes in CUB can alter a strain's ability to grow in certain environments and affect the growth rate. Observed changes in growth rate of most strains were caused by an unattainable proteome to meet metabolic requirements arising from tRNA supply shortage. This shortage could be adjusted by expansion of tRNA content or reading. Our results also strengthened previous observations that synonymous codon usage significantly impacts achievable growth phenotypes as we identified reduced and increased maximal growth rates of the *in silico* strains depending on environmental niche. Using a genome-scale analysis framework that is novel to molecular systems biology, we provided an explanation of how expansion of tRNA content and/or reading could be used as an evolutionary mechanism to deal with mismatches between CUB (genotype) and environment to maximize growth rate (phenotype). This work will have great impact for recombinant protein expression, protein and metabolic engineering, and minimal genome design. The integrated metabolic, transcriptional, and translational model can also be employed to rate metabolic engineering designs in terms of metabolic and energetic costs.

Modification was completed of the script to generate, from the necessary parts list, a reconstruction of macromolecular synthesis for a generic prokaryote. This provided the basis for reconstructing the macromolecular synthesis network of *Thermotoga maritima* (Lerman *et al.*, *Nature Communications*, 2012).

2(b) *Reconstruct and analyze the macromolecular synthesis network of Thermotoga maritima.*

The article by Lerman *et al.* (2012) entitled "*In silico method for modelling metabolism and gene product expression at genome scale*" describes the reconstruction and analysis of an integrated model of metabolism and macromolecular expression for *Thermotoga maritima*. The final model accounts for about 60% of the organism's annotated open reading frames. We assessed the model's ability to predict byproduct secretion and systems-level molecular phenotypes by comparing our model's predictions to byproduct secretion, amino acid (AA) composition, transcriptome, and proteome measurements. With the only external constraints for our model being the measured sugar uptake rate and doubling time for *T. maritima* during log-phase growth in minimal maltose medium at 80 degrees C, our model accurately predicted acetate secretion. Predicted AA incorporation was linearly correlated (0.8 Pearson correlation coefficient (PCC); $p < 2.2 \times 10^{-5}$ t-test) with measured AA composition. Additionally, there were statistically significant positive correlations between transcriptome (0.57 PCC; $p < 1.5 \times 10^{-8}$ t-test) and proteome (0.57 PCC; $p < 1.1 \times 10^{-10}$ t-test) measurements and the relevant synthesis fluxes in our integrated model. The strong concordance is interesting because the model does not account for transcriptional regulation, and may be due to *T. maritima*'s genome possessing few regulatory states and gene-products for basic cellular functions being over-represented in the cell's transcriptome and proteome. These results, and follow-up work exceeding the original aims (H. Latif *et*

al., submitted) illustrate that it is possible to sketch a molecular description of an organism solely from simple chemical equations.

As with the analogous *E. coli* model (Aim 2(a)), completion of this model represents a major landmark in quantitative systems biology because it is an integrated mathematical representation of two major subsystems in a living organism. We demonstrate the ability to compute reliably with the integrated stoichiometric matrix. A provisional patent application that includes portions of the research described in the published manuscript was filed by the University of California, San Diego Technology Transfer Office on May 9, 2012 entitled *Method for in silico modeling of gene product expression and metabolism.*

Aim 3 *Quantify the significance of thermodynamic constraints on prokaryotic metabolism.*

3(a) *Refine convex flux balance analysis for simultaneous prediction of metabolic fluxes and concentrations in Escherichia coli.*

As discussed under Aim 1(c), in order to make predictions of absolute steady state molecular species concentration, we found that it is necessary to enforce reaction kinetic constraints, in addition to thermodynamic and steady state constraints. When a steady state satisfies some form of reaction kinetic constraint for each reaction in a network, we say it is kinetically feasible. With this definition in place, the following two questions arise: (1) Is it possible to define easily testable conditions that guarantee the existence of a kinetically and thermodynamically feasible non-equilibrium state?, and (2) Is it possible to develop an efficient globally convergent algorithm to compute such steady states? After a prolonged and intensive collaborative effort between researchers at Iceland and Stanford, we were able to answer both questions in the affirmative.

In a 2012 *Journal of Theoretical Biology* article entitled "Mass conserved elementary kinetics is sufficient for the existence of a non-equilibrium steady state", Fleming and established easily testable conditions for existence of at least one kinetically and thermodynamically feasible non-equilibrium state, for an arbitrary biochemical reaction network, assuming continuous reaction rate laws were such that molecular species concentration could never be a negative quantity. This result does not omit the possibility that a molecular species could have a concentration of zero at steady state. The Akle *et al.* (2012) manuscript "Existence of positive steady states for mass conserving and mass-action chemical reaction networks with a single terminal-linkage class" deals with the fact that several systems biology models assume that chemical reaction networks governed by mass action kinetics admit strictly positive steady states (no molecular species with a concentration of zero), yet the literature supports this assumption for only some particular classes of networks. We established that weakly reversible networks formed by a single linkage class with no material exchange across the boundary always admit at least one strictly positive steady state. We also extended this result to systems where the network is formed by a weakly connected single linkage class with a prescribed rate of material exchange across the boundary.

At the 2012 Multiscale Modeling Consortium Meeting in Bethesda, Fleming, Sun, Thiele, and Saunders presented a poster entitled "A globally convergent algorithm for computing non-equilibrium steady state concentrations in genome-scale biochemical networks". This gave a mathematical proof of global convergence for an algorithm for simultaneously computing stable non-equilibrium steady state absolute molecular

concentrations and reaction rates for multi-scale biochemical reaction networks. Any steady state predicted with this algorithm satisfies thermodynamic and elementary reaction kinetic constraints. The results of numerical experiments with a genome-scale kinetic model of *E. coli* metabolism (iAF1260) were also presented.

3(b) *Use existing data to validate and interpret flux and concentration prediction in Escherichia coli and Thermotoga maritima.*

The work envisaged within this aim has not yet been completed as it relies on completion of Aim 3(a), which was only completed near the end of the funding period, after 4 years of effort. In 2013 we shall publish the manuscript detailing the globally convergent algorithm, with a biochemical application to predict flux and concentrations in an *E. coli* genome-scale model. We could have published the algorithm on its own in 2012 but elected not to as we believe its utility will be better understood when applied biochemically.

3(c) *Predict thermodynamically favorable pathways for hydrogen production by Thermotoga maritima on a range of substrates.*

The article by J. Nogales *et al.* (2012) investigates how hydrogen production in *T. maritima* can be increased using inexpensive substrates and a limited number of genetic modifications (given the lack of published genetic modification protocols available for *T. maritima*). Flux balance analysis (FBA) was carried out on all possible strains with three or fewer knockouts, where each strain was tested on several different carbon sources. The results of the computational study are being analyzed, and already there are some interesting findings. In agreement with previous experimental reports we found that limiting the amount of sulfur in the medium increases the overall yield of hydrogen. This result strongly highlights the potential of FBA in the optimization of the culture medium for optimal H_2 production. Furthermore, single knockout mutants (acetate kinase) growing on cellulose or glucose achieve up to 12% increase in hydrogen production, while the yield increase for xylan was 33%. After *in silico* double knockout mutant analysis, using glucose as carbon source, several combinations of gene deletions yielded a significant increment in H_2 production, although at the expense of slower growth rates. For example, when the genes encoding for glucose-6-phosphate isomerase and xylose isomerase were knocked out and the amount of sulfur present in the medium was limited, the hydrogen yield could be increased by 41% compared to the wild type yield.

Interestingly, the double knockout mutant analysis suggested that by redirecting the carbon flux through the oxidative pentose phosphate pathway, hydrogen yield could be increased. A further investigation revealed that utilization of the oxidative pentose phosphate pathway could create an imbalance of redox cofactors because *T. maritima* lacks an efficient reaction for NADP+ regeneration, yielding the lower growth rate computed in the double mutants. This result showed the role of the oxidative pentose phosphate pathway as an anaplerotic pathway in *T. maritima* as well as in other anaerobic microorganisms and strongly suggested that the NADPH and NADH pools in *T. maritima* are not connected. Further *in silico* analysis suggested that by inclusion of a transhydrogenase and/or ferredoxin-NADPH oxidoreductase reactions into the model, one could retain high growth rate and a high yield of hydrogen.

Finally, we tested our strategy by using cheap carbon sources such as glycerol, a waste product from the biodiesel industry. Interestingly, and unlike that in the wild-type *in silico* strain, glycerol could be metabolised by an *in silico* engineered strain with a thermostable ferredoxin-NADPH oxidoreductase as NADP+-regenerating enzyme, computing a significant H_2 yield. In summary, our *in silico* analysis allowed us to predict the optimal culture conditions for H_2 production in *T. maritima*. Moreover, it suggested a novel strategy for increasing the H_2 production based on the inclusion of an oxidative module involving redirecting the carbon flux through the oxidative pentose phosphate pathway (by blocking glycolysis) and regenerating NADP+ (by the inclusion of a thermostable transhydrogenase and/or ferredoxin-NADPH oxidoreductase). Furthermore, the oxidative module design allowed both growth and H_2 production in cheap and (until now) unavailable carbon sources for *T. maritima*, such as glycerol.

In 2010, we began a project to calculate the standard Gibbs energies of formation for *T. maritima* metabolites. Unfortunately, due to the lack of experimental data on heat capacities and standard enthalpies of formation, the Gibbs free energy of formation at a temperature of 353 Kelvin could be estimated for only one fifth (101/503) of the metabolites. This allowed us to estimate the standard Gibbs energies for only 17% of the reactions in the *T. maritima* model at 353K. It is difficult to state with confidence that such estimates, mainly derived from semi-empirical models of sparse experimental data, are correct. Most of the experimental data on heat capacities is for hydrocarbons of interest to chemists rather than biochemists. There have been insufficient measurements of thermochemical properties of organic molecules of biochemical interest to thermodynamically constrain *T. maritima* metabolism.

In collaboration with the group of Ron Milo, Weizmann Institute, Israel, we developed a method for combining the thermodynamic estimations of group contribution methods with the more accurate reactant contribution method (derived without estimation from thermochemical data), by decomposing each reaction into two parts and applying one of the methods for each part. This method gives priority to the reactant contributions over group contributions while guaranteeing that all estimations will be consistent, i.e., will not violate the first law of thermodynamics. We show that there is a significant increase in the accuracy of our estimations compared to standard group contribution. Specifically, we find an 80% reduction in the median cross-validation error for reactions that can be derived by reactant contributions only. We provide the full framework and open source code for deriving the standard Gibbs energy estimations with uncertainties, and believe this will facilitate the wide use of thermodynamic data for a better understanding of metabolism. A manuscript entitled "Consistent Estimation of Gibbs Energy using Component Contributions" has been submitted detailing this work. While this work allows more accurate use of existing experimental thermochemical data, further progress is limited by the lack of biochemically relevant thermochemical data.

We are investigating the feasibility of *ab initio* calculation of standard Gibbs energies of formation for *T. maritima* metabolites in collaboration with Dr William Cannon, Pacific Northwest National Laboratory (see, e.g., Cannon, Pettitt, and McCammon, 1994. Sulfate anion in water: Model structural, thermodynamic and dynamic properties. *J. Phys. Chem.* 98:6225). Even if *ab initio* calculation is computationally tractable, there is a need for more experimental measurement of thermochemical prop-

erties of metabolites in order to test the quality of such calculations. Dr Robert Goldberg, National Institute of Standards and Technology, is currently seeking support for such an effort (personal communication).

Thermodynamically realistic modeling of *T. maritima* metabolism will only become possible once sufficient experimental or *ab initio* data becomes available. We recognized that this was unlikely to occur within the period of funding for this grant and looked to engage in modeling the metabolism of another organism of strategic interest to the DOE. Prof Ines Thiele therefore directed Dr Juan Nogales (Iceland 2010, UCSD 2011–2012) and Dr Steinn Gudmundsson (Iceland) to develop a high-quality metabolic network reconstruction of the cyanobacterium, Synechocystis sp. PCC6803, capturing key photosynthetic processes for the first time at a mechanistic level (Nogales *et al., PNAS* (2012) and J. Nogales *et al., Bioengineered* (2013)). The key results include:

1. A low robustness of the metabolic network, as up to 38% of its metabolic genes are essential.
2. Two main states of the photosynthetic apparatus were identified: Ci-limited state and light-limited state.
3. Nine alternative electron flow pathways exist assisting the photosynthetic linear electron flow with optimal photosynthesis performance.
4. High photosynthetic-yield pathways exist to optimize growth under suboptimal light condition, while the joint work of low photosynthetic-yield pathways guarantees cell viability under excessive Ci or light limitation.
5. Photorespiration was found to be essential for optimal photosynthetic performance, clarifying its role in high-light acclimation.

The *in silico* results unraveled an extremely high photosynthetic robustness driving optimal autotrophic metabolism at the expense of metabolic robustness. Further, a high grade of cooperativity between alternative pathways was found to be critical for optimal photosynthetic performance under perturbation. This modeling effort gave rise to a better understanding of the photosynthetic process underlying bioengineering projects for optimal biofuel production using photosynthetic organisms.

3(d) *Numerically sample mass conserved, thermodynamically feasible steady state fluxes and concentrations in Escherichia coli growing on various media.*

Under Aim 1(c) we discussed the rationale for refocussing effort from sampling thermodynamically feasible steady state fluxes, toward searching for thermodynamically feasible fluxes that also satisfy reaction kinetic constraints. The successful work to also satisfy reaction kinetic constraints is detailed in Aim 3(a). As discussed, we envisage biochemical application of this algorithm to Aim 3(b), including the modeling of *E. coli* growing on various media, thus obviating Aim 3(d).

2 Publications

Each section below is in chronological order.

Journal articles

1. S. Gudmunsson and I. Thiele. Computationally efficient flux variability analysis, *BMC Bioinformatics* 11 (2010), 489, doi:10.1186/1471-2105-11-489.

2. I. Thiele, R. M. T. Fleming, A. Bordbar, J. Schellenberger, and B. Ø. Palsson. Functional characterization of alternate optimal solutions of Escherichia coli's transcriptional and translational machinery, *Biophys. J.* 98:10 (2010) 2072–2081, doi:10.1016/j.bpj.2010.01.060.

3. R. M. T. Fleming and I. Thiele. von Bertalanffy 1.0: A COBRA toolbox extension to thermodynamically constrain metabolic models, *Bioinformatics* 27:1 (2011) 142–143, doi:10.1093/bioinformatics/btq607.

4. S. G. Thorleifsson and I. Thiele. rBioNet: A COBRA toolbox extension for reconstructing high-quality biochemical networks, *Bioinformatics* 27:14 (2011) 2009–2010, doi:10.1093/bioinformatics/btr308.

5. J. Schellenberger, R. Que, R. M. T. Fleming, I. Thiele, J. D. Orth, A. M. Feist, D. C. Zielinski, A. Bordbar, N. E. R. Lewis, J. Kang, D. Hyduke, and B. Ø. Palsson. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox v2.0, *Nature Protocols*, 6 (2011) 1290–1307, doi:10.1038/nprot.2011.308.

6. R. M. T. Fleming, C. M. Maes, M. A. Saunders, Y. Ye, and B. Ø. Palsson. A variational principle for computing nonequilibrium fluxes and potentials in genome-scale biochemical networks, *J. Theor. Biol.* 292 (2012) 71–77, doi:10.1016/j.jtbi.2011.09.029.

7. I. Thiele, R. M. T. Fleming, R. Que, A. Bordbar, and B. Ø. Palsson. Multiscale modeling of metabolism and macromolecular synthesis in E. coli and its application to the evolution of codon usage, *PLoS One* 7(9):e45635 (2012), doi:10.1371/journal.pone.0045635.

8. J. A. Lerman, D. R. Hyduke, H. Latif, V. A. Portnoy, N. E. Lewis, J. D. Orth, A. C. Schrimpe-Rutledge, R. D. Smith, J. N. Adkins, K. Zengler, and B. Ø. Palsson. *In silico* method for modelling metabolism and gene product expression at genome scale, *Nat. Commun.* 3 (2012) 929, doi:10.1038/ncomms1928.

9. J. Nogales, S. Gudmunsson, and I. Thiele. An *in silico* re-design of the metabolism in *Thermotoga maritima* for increased biohydrogen production, *Int J. Hydrogen Energy* 37:17 (2012) 12205–12218, doi:10.1016/j.ijhydene.2012.06.032.

10. J. Nogales, S. Gudmundsson, E. M. Knight, B. Ø. Palsson, and I. Thiele. Detailing the optimality of photosynthesis in cyanobacteria through systems biology analysis, *Proc. Natl. Acad. Sci. USA* 109:7 (2012) 2678–2683, http://www.pnas.org/content/109/7/2678.long.

11. R. M. T. Fleming and I. Thiele. Mass conserved elementary kinetics is sufficient for the existence of a non-equilibrium steady state concentration, *J. Theor. Biol.* 314 (2012) 173–181, doi:10.1016/j.jtbi.2012.08.021.

12. J. Nogales, S. Gudmundsson, and I. Thiele. Towards systems metabolic engineering in cyanobacteria: Opportunities and bottlenecks, *Bioengineered* 4:3 (2013) 1–6, http://www.ncbi.nlm.nih.gov/pubmed/23138691.

In revision

1. S. Akle, O. A. Dalal, R. M. T. Fleming, M. A. Saunders, N. A. Taheri, and Y. Ye. Existence of positive steady states for mass conserving and mass-action chemical reaction networks with a single terminal-linkage class. *J. of Mathematical Biology*, in revision, 2013.
2. Y. Sun, R. M. T. Fleming, I. Thiele, and M. A. Saunders. Robust flux balance analysis of multiscale biochemical reaction networks, *BMC Systems Biology*, in revision, 2013.
3. X. Meng, M. A. Saunders, and M. W. Mahoney. LSRN: a parallel iterative solver for strongly over- or under-determined systems, *SIAM J. Sci. Comp.*, `http://arxiv.org/abs/1109.5981`, in revision, 2013.

Submitted

1. H. Latif, J. A. Lerman, V. A. Portnoy, Y. Tarasova, H. Nagarajan, A. C. Schrimpe-Rutledge, R. D. Smith, J. N. Adkins, D.-H. Lee, Y. Qiu, B. Ø. Palsson, and K. Zengler. The genome organization of *Thermotoga maritima* reflects its lifestyle, *PLoS Genetics*, submitted Jun 14, 2012.
2. A. Ebrahim, J. A. Lerman, B. Ø. Palsson, and D. R. Hyduke. COBRApy: COnstraints-Based Reconstruction and Analysis for Python, *BMC Systems Biology*, submitted Aug 7, 2012.
3. E. Noor, H. S. Haraldsdottir, R. Milo, and R. M. T. Fleming. Consistent estimation of Gibbs energy using component contributions, *PLoS Comp Bio*, submitted Dec 17, 2012.

Poster presentations

1. S. Akle, O. A. Dalal, R. M. T. Fleming, M. A. Saunders, N. A. Taheri, and Y. Ye. Numerical optimization algorithms and software for systems biology: Existence of positive equilibria for mass conserving and mass-action biochemical reaction networks with a single terminal-linkage class, DOE Genomic Science Awardee Meeting IX, Crystal City, VA, Apr 10–13, 2011.
2. R. M. T. Fleming, I. Thiele, and M. A. Saunders. Numerical optimization algorithms and software for systems biology: von Bertalanffy 1.0: A COBRA toolbox extension to thermodynamically constrain metabolic models, DOE Genomic Science Awardee Meeting IX, Crystal City, VA, Apr 10–13, 2011.
3. J. A. Lerman, D. R. Hyduke, H. Latif, V. A. Portnoy, A. C. Schrimpe-Rutledge, J. N. Adkins, R. D. Smith, I. Thiele, M. A. Saunders, K. Zengler, and B. Ø. Palsson. *Thermotoga maritima* systems biology knowledgebase: A computational platform for multisubsystem reconstruction and multiscale stoichiometric modeling, DOE Genomic Science Awardee Meeting IX, Crystal City, VA, Apr 10–13, 2011.
4. I. Thiele, R. M. T. Fleming, S. G. Thorleifsson, A. Bordbar, R. Que, and B. Ø. Palsson. Numerical optimization algorithms and software for systems biology: An integrated model of macromolecular synthesis and metabolism of *Escherichia coli*, DOE Genomic Science Awardee Meeting IX, Crystal City, VA, Apr 10–13, 2011.
5. R. M. T. Fleming, Y. Sun, I. Thiele, and M. A. Saunders. A globally convergent algorithm for computing non-equilibrium steady state concentrations in genome-scale biochemical networks, MSM Consortium Meeting, Bethesda, MD, Oct 22–23, 2012.
6. Y. Sun, R. M. T. Fleming, I. Thiele, and M. A. Saunders. Flux balance analysis of multiscale biochemical reaction networks, MSM Consortium Meeting, Bethesda, MD, Oct 22–23, 2012.
7. I. Thiele and R. M. T. Fleming. Multiscale molecular systems biology: Reconstruction and model optimization, MSM Consortium Meeting, Bethesda, MD, Oct 22–23, 2012.

Oral presentations

1. I. Thiele. An integrated model of macromolecular synthesis and metabolism in *E. coli*, Wellcome Trust Functional Genomics and Systems Biology Workshop, Cambridge, England, Nov 2009.
2. I. Thiele. Introduction to metabolic systems biology, Seminar at The Netherlands Metabolomics Centre, Leiden University, The Netherlands, Dec 2009.
3. R. M. T. Fleming. Physico-chemical integration of metabolomic data with stoichiometric models of metabolism, Seminar at The Netherlands Metabolomics Centre, Leiden University, The Netherlands, Dec 2009.
4. R. M. T. Fleming, I. Thiele, C. M. Maes, M. A. Saunders, Y. Ye, and B. Ø. Palsson. Optimality principles in nonequilibrium biochemical networks, Wellcome Trust Functional Genomics and Systems Biology Workshop, Cambridge, England, Dec 2009.
5. R. M. T. Fleming, I. Thiele, C. M. Maes, M. A. Saunders, Y. Ye, and B. Ø. Palsson. Optimality principles in nonequilibrium biochemical networks, DOE Genomic Science Awardee Meeting VIII, Crystal City, VA, Feb 2010.
6. I. Thiele, R. M. T. Fleming, A. Bordbar, R. Que, and B. Ø. Palsson. An integrated model of macromolecular synthesis and metabolism of *E. coli*, DOE Genomic Science Awardee Meeting VIII, Crystal City, VA, Feb 2010.
7. R. M. T. Fleming. A variational principle for computing nonequilibrium fluxes and potentials in genome-scale biochemical networks, Para 2010: State of the Art in Scientific and Parallel Computing, Reykjavik, Jun 2010.
8. I. Thiele. Expanding genome-scale metabolic models, Workshop IOMPA, 11th International Conference on Systems Biology, Edinburgh, Oct 2010.
9. R. M. T. Fleming. On optimality principles in nonequilibrium biochemical networks, 11th International Conference on Systems Biology, Edinburgh, Oct 2010.
10. R. M. T. Fleming. Stoichiometric modelling of biochemistry: A variational approach, Biomedical Engineering Department seminar, University of Texas at Dallas, Mar 11, 2011.
11. S. Akle, O. A. Dalal, R. M. T. Fleming, M. A. Saunders, N. A. Taheri, and Y. Ye. Existence of positive equilibria for mass conserving and mass-action biochemical reaction networks with a single terminal-linkage class, DOE Genomic Science Awardee Meeting IX, Crystal City, VA, Apr 10–13, 2011.
12. M. A. Saunders, I. Thiele, R. M. T. Fleming, B. Ø. Palsson, Y. Ye, S. Akle, O. A. Dalal, J. A. Lerman, Y. Sun, and N. A. Taheri. Satisfying flux balance and mass-action kinetics in a network of biochemical reactions, DOE Genomic Science Awardee Meeting IX, Crystal City, VA, Apr 10–13, 2011. `http://www.stanford.edu/group/SOL/talks/11doe.pdf`.
13. R. M. T. Fleming. A variational approach to the solution of stochiometric polynomial systems, 1st International Conference on Constraint-Based Reconstruction and Analysis, Reykjavik, 2011.

3 Unexpended funds

At the time of close-out at Stanford, unexpended funds totalled approximately $100 (to be returned to DOE).