Analysis of TX-TL Synthetic Circuits using Sequence Specific Constraints Based Modeling

Jeffrey D. Varner*

School of Chemical and Biomolecular Engineering

Cornell University, Ithaca NY 14853

Running Title: Constraints based models of synthetic circuits

To be submitted: *Scientific Reports*

*Corresponding author:

Jeffrey D. Varner,

Professor, School of Chemical and Biomolecular Engineering,

244 Olin Hall, Cornell University, Ithaca NY, 14853

Email: jdv27@cornell.edu

Phone: (607) 255 - 4258

Fax: (607) 255 - 9166

Abstract

In this study, we used sequence specific constraints based modeling to evaluate the performance of synthetic circuits in an *E. coli* TX-TL system. A core *E. coli* metabolic model, consisting of XX metabolites and YY reactions, was developed from literature [REF]. This model, which described glycolysis, pentose phosphate pathway, amino acid biosynthesis and degradation and energy metabolism, was then augmented with sequence specific descriptions of genetic circuits which included mechanistic models of promoter function, transcription and translation. Thus, unlike other synthetic biology modeling efforts, sequence specific constraints based modeling explicitly couples the transcription and translation of circuit components with the availability of metabolic resources. Model parameters were largely taken from literature; our approach had very few adjustable parameters thereby allowing the a first principles prediction of circuit performance. We tested this approach by first simulating σ_{70}-induced deGFP expression and then expanded these studies to more complex multicomponent circuits. First principles predictions of circuit performance were consistent with measurements for a variety of cases. Further, global sensitivity analysis identified the key metabolic processes that controlled circuit performance. Taken together, sequence specific constraints based modeling offers a novel means to *a priori* estimate the performance of cell free synthetic circuits.

Keywords: Synthetic biology, Constraints based modeling, Biochemical modeling

Introduction

Cell free systems offer many advantages for the study, manipulation and modeling of metabolism compared to *in vivo* processes. Central amongst these advantages is direct access to metabolites and the microbial biosynthetic machinery without the interference of a cell wall. This allows us to control as well as interrogate the chemical environment while the biosynthetic machinery is operating, potentially at a fine time resolution. Second, cell-free systems also allow us to study biological processes without the complications associated with cell growth. Cell-free protein synthesis (CFPS) systems are arguably the most prominent examples of cell-free systems used today [**?**]. However, CFPS is not new; CFPS in crude *E. coli* extracts has been used since the 1960s to explore fundamentally important biological mechanisms [**?** **?**]. Today, cell-free systems are used in a variety of applications ranging from therapeutic protein production [**?**] to synthetic biology [**?**]. Interestingly, many of the challenges confronting in-vivo genome-scale kinetic modeling can potentially be overcome in a cell-free system. For example, there is no complex transcriptional regulation to consider, transient metabolic measurements are easier to obtain, and we no longer have to consider cell growth. Thus, cell-free operation holds several significant advantages for model development, identification and validation. Theoretically, genome-scale cell-free kinetic models may be possible for industrially important organisms, such as *E. coli* or *B. subtilis*, if a simple, tractable framework for integrating allosteric regulation with enzyme kinetics can be formulated.

Stoichiometric reconstructions of microbial metabolism popularized by constraint based modeling techniques such as flux balance analysis (FBA) have become standard tools to interrogate biological networks [**?**]. Since the first genome-scale stoichiometric model of *E. coli*, developed by Edwards and Palsson [**?**], stoichiometric reconstructions of hundreds of organisms, including industrially important prokaryotes such as *E. coli* [**?**] or *B. subtilis* [**?**], are now available [**?**]. Stoichiometric models rely on a pseudo-steady-state

assumption to reduce unidentifiable genome-scale kinetic models to an underdetermined linear algebraic system, which can be solved efficiently even for large systems using linear programming. Traditionally, stoichiometric models have also neglected explicit descriptions of metabolic regulation and control mechanisms, instead opting to describe the choice of pathways by prescribing an objective function on metabolism. Interestingly, similar to early cybernetic models, the most common metabolic objective function has been the optimization of biomass formation [**?**], although other metabolic objectives have also been estimated [**?**]. Recent advances in constraint-based modeling have overcome the early shortcomings of the platform, including capturing metabolic regulation and control [**?**]. Thus, modern constraint-based approaches are extremely useful for the discovery of metabolic engineering strategies and represent the state of the art in metabolic modeling [**? ?**].

In this study, we used sequence specific constraints based modeling to evaluate the performance of synthetic circuits in an *E. coli* TX-TL system. A core *E. coli* cell free metabolic model, consisting of XX metabolites and YY reactions, was developed from literature [REF]. This model, which described glycolysis, pentose phosphate pathway, amino acid biosynthesis and degradation and energy metabolism, was then augmented with sequence specific descriptions of genetic circuits which included mechanistic models of promoter function, transcription and translation. Thus, sequence specific constraints based modeling explicitly couples the transcription and translation of circuit components with the availability of metabolic resources. Model parameters were largely taken from literature; our approach had very few adjustable parameters thereby allowing the a first principles prediction of circuit performance. We tested this approach by first simulating σ_{70}-induced deGFP expression and then expanded these studies to more complex multi-component circuits. First principles predictions of circuit performance were consistent with measurements for a variety of cases. Further, global sensitivity analysis identified the key

metabolic processes that controlled circuit performance. Taken together, sequence specific constraints based modeling offers a novel means to *a priori* estimate the performance of cell free synthetic circuits.

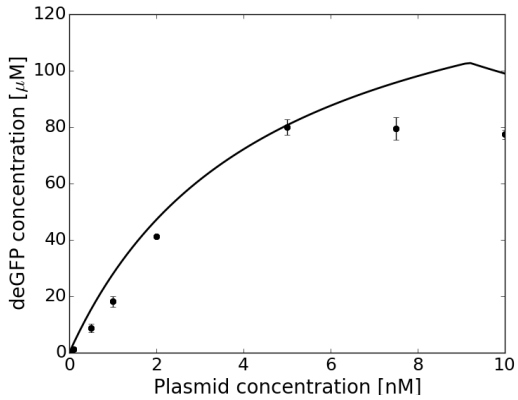

Fig. 1: Model simulation (solid line) versus deGFP protein concentration (dots) at different plasmid concentrations produced in TX-TL 2.0.

Results

The protein concentration levels of deGFP produced with TX-TL 2.0 at different plasmid concentrations was modeled using a sequence specific constraints based modeling approach in a *E. coli* cell free network (Fig. 1). The metabolic network was formulated by removing all cell wall associated reactions and incorporating amino acid synthesis and transcription/translation associated reactions. The network consisted of 281 reactions and 132 species. The transcription and translation rates were constrained to experimental measurements [**?**] for the RNA polymerase and ribosome elongation rates. The production time for TX-TL 2.0 is approximately 8 hours [**?**]. The concentration of deGFP was determined at each plasmid concentration by multiplying the corresponding production flux by the production time. The sequence specific constraints based model was able to capture the deGFP concentration up to a plasmid concentration of 5 nM and overestimates deGFP measurements above that. TX-TL 2.0 has a saturation on protein production mostly due to limited resources which our model begins to see saturating effects at 9 nM plasmid concentration. This may be due to oxidative phosphorylation activity which has been observed to take place in other cell free systems [**?**].

Oxidative phosphorylation activity influences the protein production flux as well as

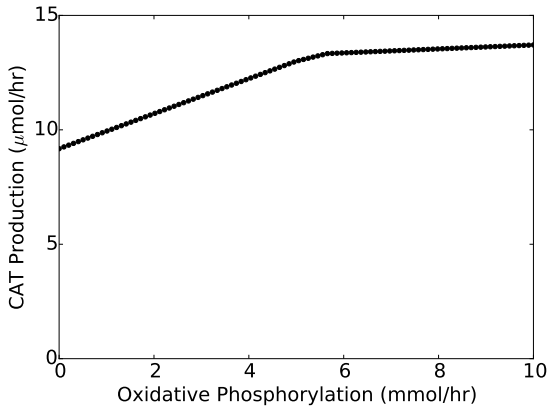

Fig. 2

acetate and lactate flux. Our model suggests higher oxidative phosphorylation activity increases the protein production rate for CAT (Fig. **??**) and other proteins. Consistent with literature, limiting oxidative phosphorylation results in lower protein yields in cell free systems as seen in our model. Oxidative phosphorylation activity was limited to have a maximum flux of 3.0 mmol/hr in it's reactions, determined from a tradeoff between acetate and lactate flux (Fig. **??**). Experimental data has shown lactate and acetate to have similar fluxes during glucose consumption during the production of CAT (data not shown). Low oxidative phosphorylation leads to NADH overflow which is re-oxidized with lactate dehydrogenase reducing pyruvate to lactate leading to high lactate formation. High oxidative phosphorylation relative to our cell free system leads to lower lactate and higher acetate formation.

Theoritical carbon yields were calculated for nine different proteins using a sequence specific constraints based modeling approach in a *E. coli* cell free network. Carbon yields were calculated for three different cases: unconstrained, limited oxidative phosphorylation and TX-TL constraints (Fig. 3). In the unconstrained case, amino acids uptake and oxidative phosphorylation activity were unbounded. Most of the proteins studied in this paper resulted in a carbon yield near forty percent for a cell free system. In the limited

oxidative phosphorylation case, carbon yields were reduced for each of the proteins.

In the TX-TL 2.0 case, oxidative phosphorylation activity was limited in addition to limited transcription and translation fluxes. Following these constraints, an optimal plasmid concentration was determined for each protein which was used to determine the carbon yield for TX-TL 2.0 at that plasmid concentration. Carbon yields are significantly reduced compared to the unconstrained and limited oxidative phosphorylation cases. The carbon yields are being lost mostly due to lactate and acetate formation. The constrained TX-TL case requires additional ATP to optimize for the protein of interest production leading to acetate formation.

TX-TL 2.0 carbon yields were then compared to the carbon number of each protein of interest (Fig. 4). There is a relative inverse relationship between the carbon yield and the carbon number of each protein with the exception of FGF21. In comparing the amino acids of each protein, the proteins with a carbon yield below ten percent have a high abundance of proline which is a non polar amino acid and may have a negative impact on increasing carbon yield.

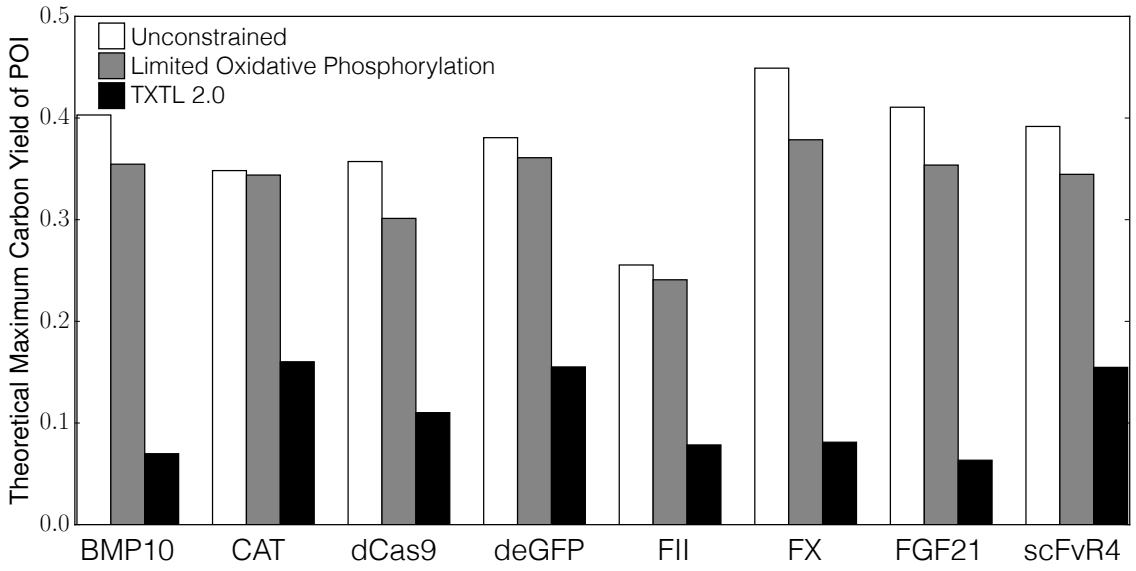

Fig. 3: Theoriticcal carbon yield of nine different proteins for three different cases in a cell free system. Unconstrained yields (white bar) produced the highest carbon yield out of the three cases. In the limited oxidative phosphorylation case (gray bar), carbon yields were slightly lower than for the unconstrained case. TXTL 2.0 (black bar) constraints had low carbon yields compared to unconstrained and limited oxidative phosphorylation cases.

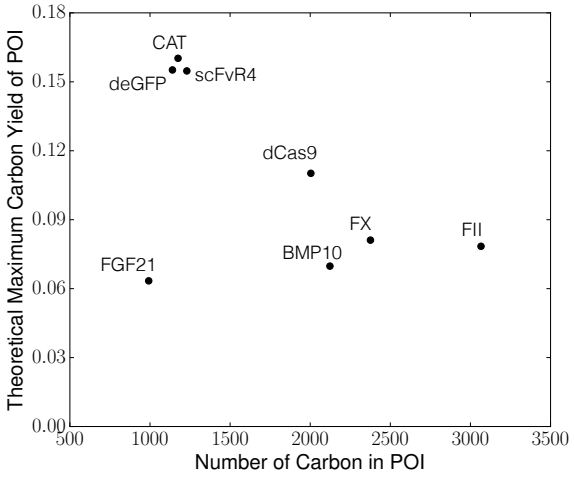

Fig. 4: Theoretical maximum carbon yield of nine proteins versus the corresponding carbon number of each protein. There is a relative inverse relationship of carbon yield to number of carbons in each protein except for the protein FGF21.

Discussion

Materials and Methods

Formulation and solution of the model equations. The flux balance analysis problem was formulated as:

$$\max_{\boldsymbol{w}} \left(w_{obj} = \boldsymbol{\theta}^T \boldsymbol{w}\right)$$

$$\text{Subject to : } \mathbf{Sw} = \mathbf{0}$$

$$\alpha_i \leq w_i \leq \beta_i \qquad i = 1, 2, \ldots, \mathcal{R}$$

where S denotes the stoichiometric matrix, $\mathbf{w}$ denotes the unknown flux vector, $\boldsymbol{\theta}$ denotes the objective selection vector and α_i and β_i denote the lower and upper bounds on flux w_i, respectively. The flux balance analysis problem was solved using the GNU Linear Programming Kit (v4.52) [**?**]. In this study, the objective w_{obj} was to maximize the production of circuit output. The specific glucose uptake rate was constrained to allow a maximum flux of 10 mmol/hr [**?**]; the amino acids were also bound to allow a maximum flux of 10 mmol/hr, but did not reach this maximum flux.

Transcription and translation template reactions. The transcription and translation template reactions are based off sequence specific FBA [**?**] involving transcription initiaion, transcription, mRNA degradation, translation initiation, translation, and tRNA charging. The mRNA and protein sequence of each protein was determined from literature. The transcription rate was constrained by the following formulation:

$$w_{tx} = [RNAP]\frac{v_{RNAP}}{l_{mRNA}}\left(\frac{[Gene]}{km + [Gene]}\right)P$$

where $[RNAP]$ is the concentration of RNA polymerase which was determined from literature values based on the number of copies per cell, v_{RNAP} is the elongation rate (nucleotides/hr) of the RNA polymerase, l_{mRNA} is the number of nucleotides in the mRNA for the protein of interest, $[Gene]$ is the gene concentration of the protein of interest, km is the plasmid saturation coefficient, and P is the promoter activity. The promoter activity

was formulated following Moon et al. for synthetic circuits by the following:

$$P = \frac{K_1 + K_2 f_{p70}}{1 + K_1 + K_2 f_{p70}}$$

where K_1 represents the state of RNA polymerase binding, K_2 is the state of sigma-70 binding along with RNA polymerase, and f_{p70} is the fraction of the transcription factor, sigma-70, bound to the promoter following Hills kinetics.

The translation rate was constrained by the following formulation:

$$w_{tl} = [Ribo] K_P \frac{v_{Ribo}}{l_{protein}} [mRNA_{ss}]$$

where $[Ribo]$ is the ribosome concentration determined from literature values based on the number of copies per cell, K_P is the polysome amplification constant, v_{Ribo} is the elongation rate (amino acids/hr) of the ribosome, $l_{protein}$ is the number of amino acids in the protein of interest, and $[mRNA_{ss}]$ is the mRNA concentration at steady state determined by the transcription rate divided by the degradation rate of mRNA.

Theoritical carbon yield. The theortical carbon yield of each protein was formulated as:

$$Yield = \frac{C_{POI} v_{POI}}{\sum_{i=1}^{\mathcal{R}} C_i v_i}$$

where C_{POI} and C_i denote the carbon number of the protein of interest (POI) and substrate i, respectively, v_{POI} and v_i denote the flux of the POI and substrate i, respectively, and $\mathcal{R}$ denotes the number of substrates consumed.

Global sensitivity analysis. We conducted a global sensitivity analysis, using the variance-based method of Sobol, to estimate which parameters controlled the performance of synthetic circuits [**?**]. We computed the total sensitivity index of each parameter relative

to two performance objectives, the peak thrombin time and the area under the thrombin curve (thrombin exposure). We established the sampling bounds for each parameter from the minimum and maximum value of that parameter in the parameter set ensemble. We used the sampling method of Saltelli *et al.* [**?**] to compute a family of $N\,(2d+2)$ parameter sets which obeyed our parameter ranges, where N was the number of trials, and d was the number of parameters in the model. In our case, N = 10,000 and d = 22, so the total sensitivity indices were computed from 460,000 model evaluations. The variance-based sensitivity analysis was conducted using the SALib module encoded in the Python programming language [**?**].

Acknowledgements

This study was supported by an award from the Army Research Office (ARO #59155-LS).

References