C. PHASE I FINAL REPORT

The budget period for our Phase I SBIR grant officially began on July 13, 2004 and ends on April 12, 2005. The actual work was conducted throughout this period and will continue afterward due to encouraging results. The key personnel involved in the project are listed below in Table 2 along with the approximate hours they have worked. Although we had originally intended to have a postdoctoral researcher perform the labeling experiments, we were unable to hire anyone with the appropriate background; therefore, the principal investigator performed the experiments himself. These experiments were done at the University of Washington under a subcontract with Dr. Mary Lidstrom, in which she provided laboratory resources but no personnel. Upon completion of Phase I, we have developed a comprehensive genome-scale metabolic model of *M. extorquens* AM1, utilized the *in silico* model to predict cell metabolism, and demonstrated the value of a combined experimental/modeling platform for microbial research. It is expected that some of the flux analysis results presented here will be published in a joint article with Dr. Lidstrom's group.

Table 1: Key individuals involved in the project, location of work, project role, and approximate hours.

Name	***Location***	***Project Role***	***Approximate Hours Spent***
Stephen Van Dien	Genomatica	Principal Investigator *M. extorquens* model development; isotope labeling experiments; data analysis	495
Christophe Schilling	Genomatica	Technical guidance and consulting	83
Anthony Burgard *	Genomatica	Development of optimization routines using GAMS	80
Krishna Mahadevan *	Genomatica	Assisting with atomic mapping matrix calcualtion	40

* The work done by these individuals was beyond the allocated budget for Phase I, and although contributed significantly to the completion of this project was not supported under this grant.

C.1 Aim 1: Perform ^{13}C-labeling experiments using chemostat-grown *M. extorquens* cultures

Six labeling experiments were performed during the course of this project, using *M. extorquens* AM1 with a single-carbon (C_1) compound as the limiting nutrient. The experiments were designed to include growth of wild-type, deletion mutant, and overexpression strains at different growth rates, as well as on two different C_1 compounds. Flux distributions for wild-type cells growing at a relatively fast rate, 80% of maximum, have already been measured and reported (Van Dien et al. 2003). This provides a base case for comparison of our other results. The first two experiments were done using the same conditions, but with reduced growth rates ("slow" and "medium", corresponding to about 50% and 15% of the "fast" growth experiment). The

objective was to see if cell physiology and metabolism is a strong function of growth rate. A third experiment used wild-type cells with methylamine instead of methanol as the carbon source. Although the metabolism of these two substrates is schematically identical after the first step, it is possible that the generation of excess ammonia during the oxidation of methylamine has an effect on downstream reactions. We next addressed the issue of heterologous protein production, using a strain overexpressing the haloalkane dehalogenase gene from *Xanthobacter autotrophicus* on a medium copy plasmid behind the strong, methanol-inducible promoter P_{mxaF} (pDH80) (FitzGerald and Lidstrom 2003). As a control, a parallel experiment was performed using *M. extorquens* AM1 containing the parent vector pCM80 (Marx and Lidstrom 2001). Finally, to examine the effect of deleting pathways in central metabolism, we examined the strain Δfdh3, which is a mutant strain with all three formate dehydrogenase genes inactivated (Chistoserdova et al. 2004). Since *M. extorquens* AM1 normally oxidizes over 60% of the primary carbon source to CO_2 through a pathway that includes FDH, such a strain must exhibit significant redirection of carbon flux compared to the wild-type.

Chemostat cultivations were performed using carbon-limited mineral salts media (Attwood and Harder 1972) in a 1.3-L benchtop fermentor (New Brunswick Scientific, Edison, NJ) at a working volume of approximately 600 mL. The carbon source was used at a label fraction of 70% ^{13}C for each experiment, but the overall molar concentration in the growth media varied depending on the strain used. pH control was available but not needed, since the culture pH remained above 6.7. The temperature was set at 28°C, and the vessel was aerated with sterile air. The agitation rate was used to maintain dissolved oxygen at >95% of saturation. The vessel was inoculated with 50 mL of cultures already grown for several generations at the appropriate label concentration. Once stable optical density was reached, steady-state was insured by waiting an additional 3 vessel volume changes before sampling was begun. After taking each 50 mL sample, the vessel was immediately filled to the original volume using fresh media, and the system was allowed to return to steady-state (3 volume changes) before the next sample. Samples were harvested, total protein was extracted and hydrolyzed, and the resulting amino acids were derivatized and analyzed by gas chromatography-mass spectroscopy (GC-MS) as described previously (Van Dien et al. 2003). Each derivatized sample was injected 3 times, thus providing at least 9 total data points for the calculation of means and variances. Raw mass isotopomer data were corrected for naturally occurring ^{13}C in the derivatization reagents and non-carbon isotopes in the entire fragment using a well-established method (Fischer and Sauer 2003). It is not intended that this isotope correction procedure be incorporated into our SimPheny flux analysis module, so this will not be discussed further. For the work outlined in this proposal, it will be assumed that true experimental mass distribution vectors will be provided as inputs.

The specific experimental conditions and growth data are summarized in Table 2. Culture supernatants were assayed for extracellular formate, but only in the Δfdh3 strain was any formate detected (1.2 mM).

Table 2: Summary of Chemostat Experiments

Strain / Condition	Dilution Rate (hr-1)	Feed conc. (mM)	Average OD600
Wild-type AM1 / methanol	0.052	24.1	0.83

Wild-type AM1 / methanol	0.017	24.1	0.65
AM1 pCM80 / methanol	0.092	26.6	0.85
AM1 pDH80 / methanol	0.085	26.6	1.15
Wild-type AM1 / methylamine	0.087	25.0	0.64
Δfdh3 / methanol	0.091	48.3	1.20

The fact that the strain burdened by overexpression of a non-native protein grows better than the control strain containing an empty vector is cause for concern. It is likely that something went wrong during this "control" experiment, possibly due to a bad batch of media or contamination. Thus we will not consider the AM1 pCM80 experiment in the rest of the report. Due to time constraints of Phase I, we were unable to repeat this chemostat run. However, in Phase I we proposed to perform 5 experiments instead of 6, so we still fulfilled this aim even when pCM80 is removed.

C.2 Aim 2: Develop a Genome-Scale *in silico* Model of *M. extorquens* AM1

C.2.1 Model Construction

In this portion of work we developed a genome-scale metabolic model of *M. extorquens* AM1 in SimPheny, which will form the core structure for all *in silico* studies carried out with this organism. The starting point for model development was the annotated *M. extorquens* AM1 genome sequence (Integrated Genomics, http://ergo.integratedgenomics.com/ERGO/), which was obtained from the Lidstrom laboratory as a list of open reading frames (ORFs) and the corresponding putative gene assignments. Most of the assignments are based on automatic annotation by the ERGO software, although some key genes, particularly those involved in methylotrophy, have been hand-curated by lab members. After uploading the gene index into SimPheny, the genes were divided into included and excluded genes during the model-building procedure. Included genes are those for which we can associate a metabolic reaction. All others are excluded, for reasons such as "non-metabolic", "insufficient similarity", "unknown function", or "non-specific metabolic function". Obvious non-metabolic genes such as transcription factors, replication and repair proteins, structural proteins, and housekeeping genes, as well as genes clearly of unknown function, were placed in the Excluded category immediately. Other cases were not so clear, and required BLAST analysis to determine that the function is non-specific or that the similarity is insufficient. Throughout the procedure, such genes were placed in the Excluded category when encountered. A complete model must have all genes either associated with a reaction, or explicitly excluded.

Metabolic reconstruction, or the association of genes with metabolic reactions, was done by region. For a given metabolic region (for example, amino acids of the glutamate family), candidate genes encoding potential enzymes in the pathway were identified in the gene index based on annotation. Reaction information was compiled from metabolic databases such as KEGG and ECOCYC. The sequence annotation for each of the predicted open reading frames

was examined in great detail using BLAST searches against the GenBank database, and rational judgment on validity of the assignment was made. The decision as whether to make the gene-reaction assignment was based on the confidence of the BLAST hit, specificity of the annotation, perceived essentiality of the enzyme, and number of other genes found with the same assignment. Invalid genes were excluded because of unknown function, insufficient similarity, or non-specific function. If valid, the corresponding reaction was chosen from the Universal Reaction Database (URDB) in SimPheny, and added to the *M. extorquens* AM1 reaction list. All of the reactions in the URDB are balanced in terms of elements and charges, and each compound is evaluated for its ionic state at a pH of 7.4. Each of the metabolites involved in the reactions are defined by their location in terms of cytosolic or extracellular. Often the URDB contains several reactions that could be catalyzed by the same enzyme, differing by electron acceptors, energy requirements, or cosubstrates. In these cases, a choice was made based on either the BLAST results or current knowledge of *M. extorquens* physiology. In absence of any such information, redox carriers were assumed to be $NADP^+$/NADPH for biosynthetic reactions and NAD^+/NADH for catabolic reactions. Finally, a confidence score of 0-4 was assigned to the reaction based on how much evidence there is for its existence in *M. extorquens*. There were a few cases (<10) in which the reaction was not yet included in the URDB, so it was added following in-house quality control procedures.

The systematic gene evaluation procedure was completed by associating genes to the functional proteins that they encode, and then linking the protein(s) to the reaction(s) that they enable or catalyze. These functional links are referred to as gene-protein-reaction (GPR) associations. Many are simple one-to-one linear relationships, but more complex situations arise in cases of multiple gene candidates, subunits of a holoenzyme, and multiple reactions catalyzed by the same enzyme. The reactions in a pathway were then augmented with reactions that were introduced based on biochemical or physiological information but for which no associated gene was located. These reactions are termed “non-gene associated” reactions. A non-growth associated maintentance term of 4.8 mmol/hr-g DCW was calculated by extrapolating chemostat yield data to zero growth rate. Finally, a metabolic map was created for each region using the Atlas feature of SimPheny. The map of central metabolism is shown in Figure 1. The entire set is comprised of 12 maps.

After completing the above procedure for the common metabolic regions, several hundred unevaluated genes remained in the gene index. Some of these genes may encode metabolic functions not part of the major pathways, or could be poorly annotated genes encoding some of the non-gene associated reactions created above. The genes were examined one at a time by BLAST analysis to determine if they should be included in the model or not. The majority were excluded because they could only be characterized to a broad functional category.

The final step of the model building procedure was to define an expression for biomass synthesis. Biomass composition of wild-type *M. extorquens* AM1 has been measured previously (Van Dien and Lidstrom 2002), and except for PHB content it was assumed to be the same in all experiments. PHB content is known to be condition-dependent, and was measured independent of this project but under similar conditions. A lumped biomass reaction for *M. extorquens* AM1 (called BIO_Mex) was created to include all macromolecues in a ratio dictated by the biomass composition. The coefficients for each component were calculated by dividing the weight

percent composition of the macromolecule in the cell by the molecular weight of the macromolecule; for example,

protein = $\frac{0.481 \quad g \quad protein / gDW}{108527 \quad moles \quad protein / gDW} = 0.00443 \; mmol\; protein / g\; DCW$.

The growth-associated maintenance energy requirement was assumed to be 51 mmol/g DCW, which is the value used in the *E. coli* model. A Microsoft Excel spreadsheet was created to take the biomass composition as an input (weight % each component) and provide coefficients for BIO_Mex reaction (mmol component / g DCW). The biomass equation also includes very small amounts (1-100 nmol/g DCW) of the following vitamins and cofactors: biotin, thiamin, CoA, FAD, folate, tetrahydromethanopterin, cob(I)alamin, NADP+, NAD+, protoheme, ubiquinone-8, PQQ, pyridoxal-5-phosphate, bacteriochlorophyll, and phytoene (as a representative carotenoid). Because the concentrations for most of these have never been measured the values are not quantitative, but the fluxes are so small that they have negligible effect on the model results. They were included to ensure a non-zero flux to their biosynthetic pathways.

Key aspects of the *M. extorquens* AM1 model are summarized in the table below.

Table 3: Statistics of the *M. extorquens* AM1 genome scale metabolic model

Genes		
Total Number of Genes	7462	
Included Genes	1143	15.32%
Proteins		
Total Number of Proteins	631	
Intra-System Reactions		
Total Number of Model Reactions	721	
Gene Associated Model Reactions	648	89.88%
Non-Gene Associated Model Reactions	73	10.12%
Metabolites		
Total Number of Metabolites	738	
Number of Extracellular Metabolites	55	7.45%
Number of Intracellular Metabolites	683	92.55%
Exchange Reactions (Input/Output)		
Total Number of Exchange Reactions	20	

We also performed a thorough "gap analysis" of the network to determine "dead-ends" in the network- metabolites that are only consumed or produced. This helps to identify potential sites where there is missing information in our metabolic knowledge of the organism, and represent focus areas for future model refinement. Currently, the model contains 60 metabolites that are produced only, and 62 that are consumed only. It is expected that the genome sequence will be completed within the next year, which may uncover new functionalities involving the dead-end metabolites. In addition, reverse BLAST against a gene encoding the missing functionality in another organism can be performed to identify potential missing genes. This work is outside the scope of this SBIR, and will be performed by the Lidstrom laboratory.

C.2.2 Simulations

The second objective of aim 2 was to use the *in silico* model to make predictions for the internal fluxes using the constraint-based approach. Methanol (or methylamine) uptake rate and cell growth rate (i.e., the chemostat dilution rate) were calculated from the data in Table 1 using the "Fermentation Module", a SimPheny extension recently created under a separate SBIR for the analysis of bioprocess data. The growth rates were set to the measured dilution rates, and substrate uptake was minimized as the objective function. The predicted uptake rate was compared to that calculated from the chemostat data (Table 4). The first simulation was for wild-type cells at the "medium" growth rate. The agreement is well within the 20% threshold imposed to define "accurate predictions" for Phase I, suggesting that the model is well defined and that the cells are near optimality when growing under these conditions. The predicted intracellular flux distribution for central metabolism is shown in Figure 1. For the slow growth rate experiment, the predicted flux distribution is identical to that in Figure 1, except with the absolute values scaled proportionally to the decrease in growth rate. The agreement between measured and predicted uptake rates in these cases is also good. In contrast, the accuracy is poor for the methylamine growth experiment. This suggests a systematic shift in flux modes away from the optimum, and will be explored using the ^{13}C results described in the next section.

Table 4: Comparison of model predictions with measurements for substrate uptake rate

Strain / Condition	**Substrate Uptake (mmol/hr-gDCW)**		
	Predicted	**Measured**	**% Error[a]**
Wild-type AM1 / methanol	4.72	5.42	12.9
Wild-type AM1 / methanol	2.19	2.35	6.8
AM1 pDH80 / methanol	7.06	6.96	-1.4
Wild-type AM1 / methylamine	7.25	11.70	38.0
Δfdh3 / methanol	10.13	13.0	22.1

[a] % error defined as 100*(measured-predicted)/measured

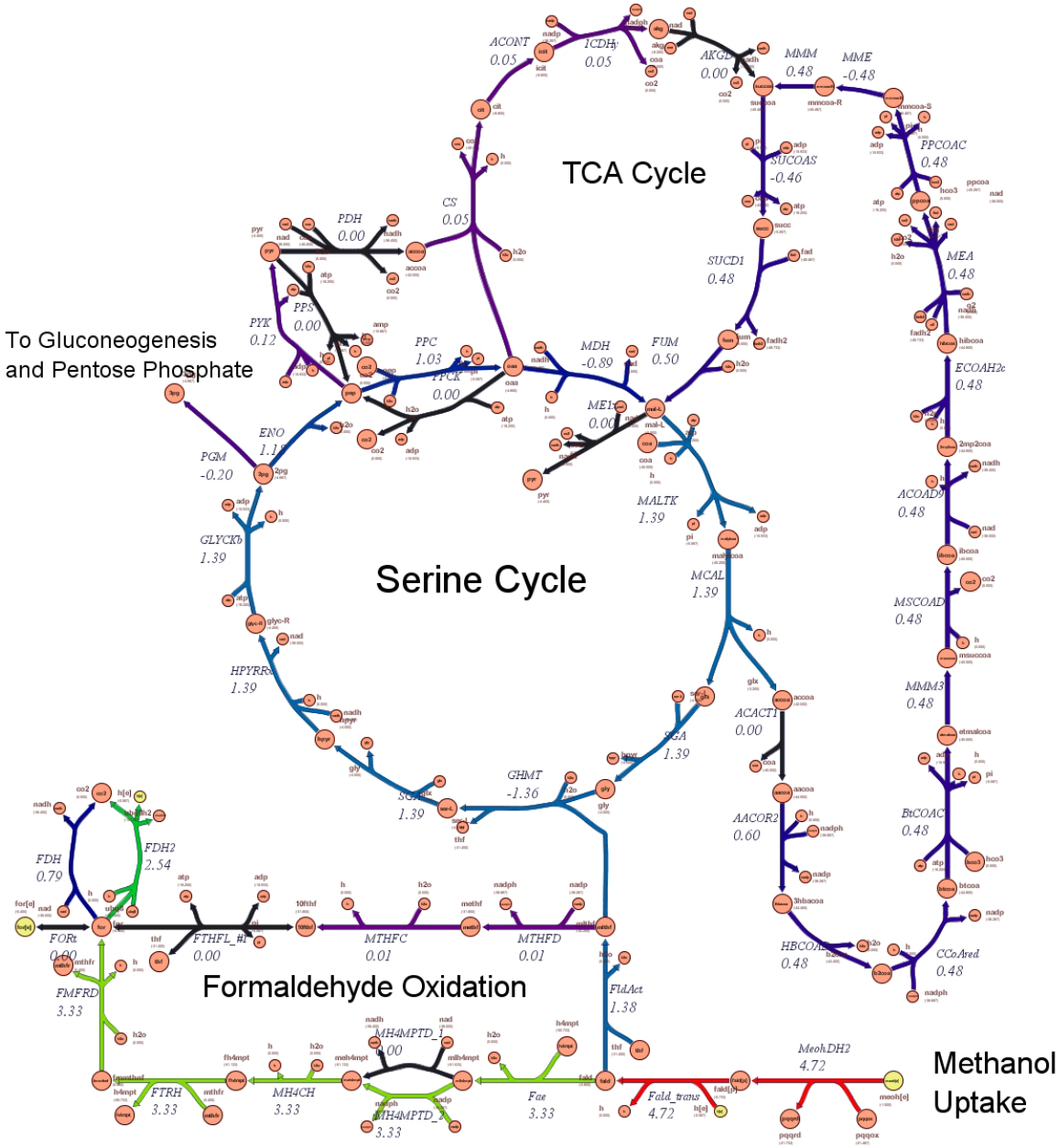


Figure 1: Predicted flux distribution for wild-type *M. extorquens* AM1 central metabolism grown on methanol. Growth rate was set to the experimental dilution rate, and methanol uptake minimized. Briefly, methanol is oxidized to formaldehyde, which is taken up into the cytoplasm and coupled to one of carriers, tetrahydrofolate (H_4F) and tetrahydromethanopterin (H_4MPT), to generate methylene-folate derivatives (Chistoserdova et al. 1998). The model predicts that methylene-H_4MPT is exclusively oxidized to formate and subsequently to CO_2, generating ATP and NAD(P)H, while only methylene-H_4F enters central metabolism via the serine cycle. For each pass through the serine cycle, one carbon atom enters as CO2 in addition to that from methylene-H_4F. Other pathways such as the TCA cycle and pentose phosphate pathway are predicted to be used for biosynthesis only, and do not form complete cycles for energy production. These predictions have been confirmed experimentally (Van Dien et al. 2003).

To simulate Δfdh3, all three FDH reactions were constrained to zero flux. As a result, the primary means of formaldehyde oxidation are effectively removed, and the cell must find other means to generate energy and NAD(P)H from formaldehyde. Simulations for the mutant strain were performed by two methods. The first is identical to that used for the wild-type, except that the missing enzymes were removed from the model. The second approach is termed minimization of metabolic adjustment (MOMA), and is based on the fact that mutants are more likely to resist deviations in metabolic fluxes, rather than to maximize growth rate (Segre et al. 2002). The flux distributions predicted by both methods are similar, in that the cell generates NADH by channeling a high flux through the TCA cycle (Figure 2). The predicted minimum methanol uptake rate for the applied dilution rate is reasonable, but not as accurate as for the wild-type.

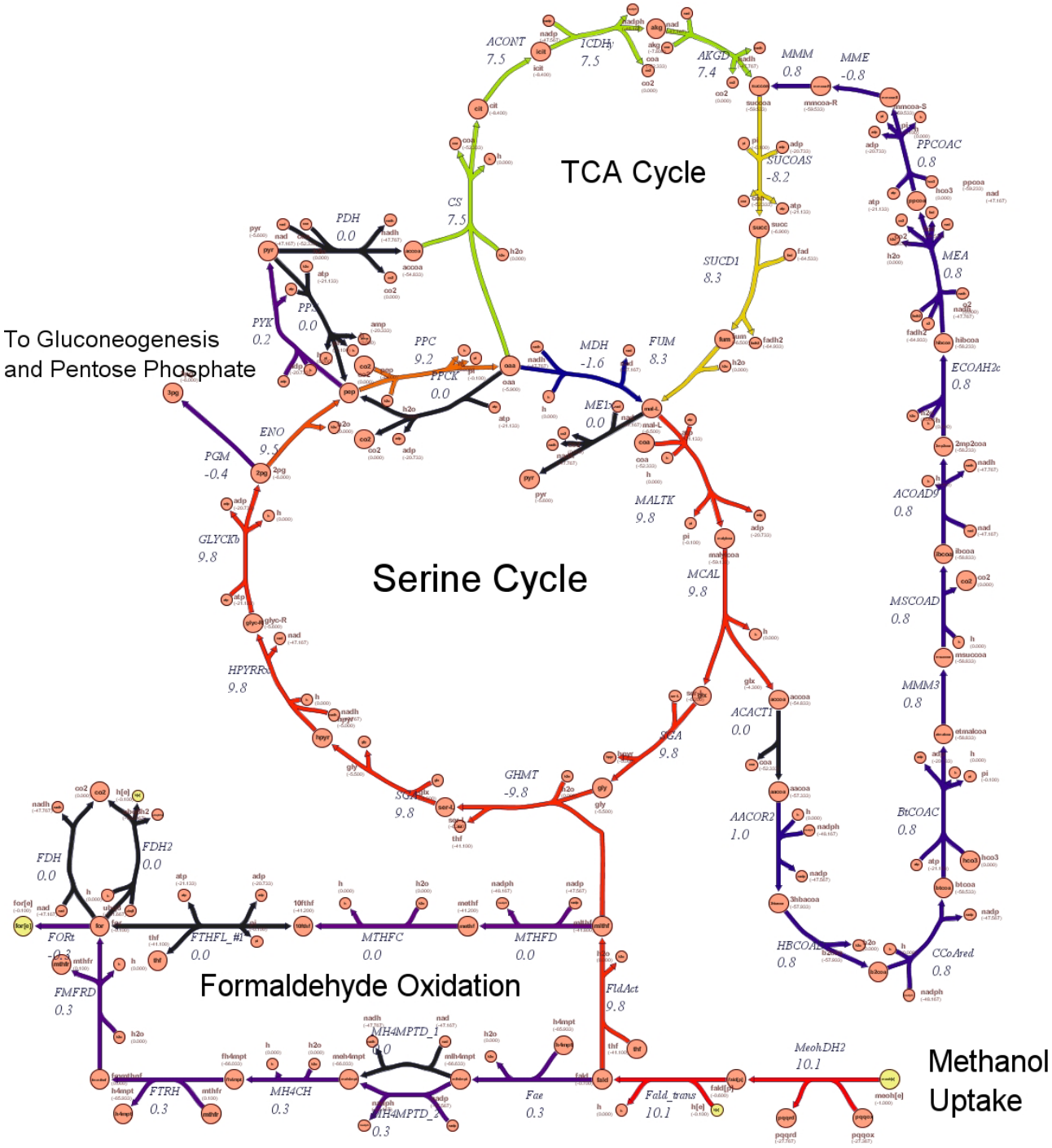


Figure 2: Predicted flux distribution for the Δfdh3 strain grown on methanol. Growth rate was set to the experimental dilution rate, and methanol uptake minimized.

C.3 Aim 3: Calculation of Intracellular Fluxes from Isotopomer Distribution Data

C.3.1 Construction of a Reduced Stoichiometric Model

The calculation of an isotopomer balance is much more complex than the overall mass balances done in traditional constraint-based analysis, both due to the non-linearity of the equations and the number of variables involved. The number of total isotopomers in the system is equal to

$$\sum_{i}^{M} 2^{n_i}$$

Where M is the number of metabolites in the system and n_i is the number of carbon atoms in metabolite i. The entire *M. extorquens* AM1 genome-scale model is therefore unwieldy for the computationally-intense process of isotopomer balancing, and there is strong incentive to reduce the size of the metabolic network for this procedure. For the initial Phase I study, we decided to simplify the network as much as possible, to include just those portions of the model involved in central metabolism and amino acid biosynthesis. For commercial applications, we would likely retain more pathways. We developed a semi-automated procedure for generating a reduced model containing any desired subset of the original model. This is summarized as follows

1. Download the stoichiometric matrix from SimPheny, as well as lists of reactions and metabolites that correspond to each row or column.
2. Remove all reactions contained only in excluded portions of metabolism, and delete corresponding columns from the stoichiometric matrix
3. Remove all metabolites not involved in the transfer of carbon atoms, and delete corresponding rows from the stoichiometric matrix. These include inorganic ions, cofactors, and energy carriers. The formaldehyde oxidation pathways were also removed, since these involve exclusively one-carbon compounds and cannot be quantified using carbon labeling (Van Dien et al. 2003).
4. Remove "duplicate" reactions that only differ in cofactor or direction. Merge opposite reactions (ex., PEP carboxylase and PEP carboxykinase) into a single reversible reaction.
5. After above manipulations, remove any rows or columns containing only zeros from stoichiometric matrix, and remove corresponding reactions and metabolites from lists.
6. Remove "dead-end" reactions, which produce a compound that is not used in any other reaction.
7. Lump all reactions between branch points into a single pathway, thus reducing the size of the matrix without any loss of information.
8. Construct a new biomass reaction based on the simplified network, and add as a final column to the stoichiometric matrix. The biomass requirements were obtained by observing the fluxes of the metabolites toward biosynthesis using the full SimPheny model.

The reduced metabolic network contains 35 reactions and 32 metabolites. Since CO_2 can freely exit the cell and is thus not constrained by a balance, there are 4 degrees of freedom; i.e., the network can be completely defined by specification of 4 independent fluxes. A further complication is introduced by reversible reaction steps, since both the forward and reverse reaction rates affect the observed isotopomer distribution. These are generally expressed in terms of the net flux and the exchange coefficient between 0 and 1 (Dauner et al. 2001; Wiechert and

de Graaf 1997). The reduced network contains 10 reversible reactions, so these 10 exchange coefficients are additional adjustable parameters which are determined using the isotopomer model.

C.3.2 An Automated Procedure for Isotopomer Mapping

Isotopomer mapping matrices (IMMs) describe the transfer of carbon atoms from the reactants to products, and are a property of a given reaction independent of the particular model (Schmidt et al. 1997). IMMs are based on mechanistic information, but can also be predicted by finding the best structural match between all atoms in the reactants and products. We employed Pipeline Pilot™ (SciTegic Inc., San Diego), a high-throughput data analysis and mining system for chemoinformatic applications, to predict IMMs using a structure matching algorithm. Consultants from SciTegic helped to create a workflow process in Pipeline Pilot specific for our application. The input is a list of reactions with associated reactants and products and their KEGG ID numbers (http://www.genome.jp/kegg/ligand.html). Pipeline Pilot then extracts the molecule files (.mol format) from our in-house database, and calculates the predicted IMM as well as a score indicating the quality of the match in the optimal and suboptimal cases.

The above procedure was done for the 35 reactions in the reduced *M. extorquens* AM1 model, and checked manually based on known biochemistry. In only two cases the predictions were incorrect: transaldolase and transketolase. Inspection of the results show that in these two cases, the score for the actual solution is only slightly lower than that of the predicted (incorrect) solution.

We therefore propose a semi-automated strategy for calculating IMMs of large sets of reactions. The Pipeline Pilot workflow will be run for the entire set, and for each reaction the scores of the best and next-best solution compared. If the optimal solution is much better, than the prediction will be accepted. If two or more solutions have very similar scores, then the prediction may not be correct, and the IMM is calculated manually. Although the procedure is not completely automated, it will likely reduce the manual work by over 90%.

C.3.3 Calculate Bounds using Fermentation Data with Flux Variability Analysis

Using the measured substrate uptake and biomass yield values from each experiment, and assuming 10% experimental error, we used a technique called Flux Variability Analysis (FVA) (Mahadevan and Schilling 2003) to calculate the range of variability that can exist in each flux in the network and still satisfy the imposed constraints. These ranges were used as the upper and lower bounds for the flux calculation.

C.3.4 Perform Isotopomer Balance

The isotopomer balance algorithm calculates the predicted set of isotopomer distribution vectors (IDVs) for all metabolites in the network given the flux distribution. The input is a random flux distribution and set of exchange coefficients that are within specified bounds and satisfy the overall metabolite balance $S*v=0$. The program was written based on previous work done by the

principal investigator, but was generalized in anticipation of the needs to apply to systems other than *M. extorquens* AM1. Previous programs used in this step (Van Dien et al. 2003) contained each reaction explicitly, and were thus model-specific. In this work, we used multi-dimensional matrices and nested loops through both reactions and compounds to achieve generality. Through each iteration, the IDV of compound i was calculated as

$$\mathrm{IDV_i} = \frac{1}{\sum_{k=1}^{M} v_{i,out}^{k}} \times \sum_{k=1}^{M} \left[v_{i,in}^{k} \prod_{j=1}^{n_k} (\mathrm{IMM}_{\mathrm{j}\to\mathrm{i}}^{\mathrm{k}} \times \mathrm{IDV_j}) \right]$$

where M is the total number of reactions in the network, n_k is the number of substrates in reaction k, $v^k_{i,out}$ is the flux of metabolite i in reaction k if it is consumed, and $v^k_{i,in}$ is the flux if it is produced. The resulting program is model-independent. It generates the isotopomer balance once the stoichiometric matrix and list of IMMs are supplied, and calculates all IDVs given the input flux distribution and isotopomer distribution of the feed molecules (in this case, methanol and CO_2).

Mass spectrometry data do not provide the entire isotopomer distribution of a compound, but rather the mass distribution of each amino acid fragment analyzed. This information is written in the form of a mass distribution vector (MDV), a column vector containing mole fractions of a group of isotopomers all with the same mass (Wittmann and Heinzle 1999). The final step of the isotopomer balance routine was to calculate the predicted MDVs for all observed products. The user supplies the set of measured MDVs, along with an identifier to indicate the compound and fragment type, and the program compares the MDVs calculated by isotopomer balance to the measurements. The objective value, used for the optimization described below, is the sum-of-squares difference between these values, weighted by the standard deviations in order to favor the most accurate measurements. This function contained 167 terms for our experiments, each corresponding to a different mass isotopomer of an amino acid fragment detected by GC-MS. The isotopomer balance was coded in both Matlab (The Mathworks, Natick, MA) and GAMS (Brooke 1998a) for use with the different optimization routines described below.

C.3.5 Optimization Algorithm

The ultimate goal of the flux analysis routine is to find the values of fluxes and exchange coefficients that minimize the value of the objective function. The isotopomer balance equations contain bi-linear terms, so this minimization is a non-linear programming problem. Due to the non-convex nature of the solution space, the existence of multiple local optima is likely and conventional gradient-based algorithms will converge upon on of these local solutions. A number of algorithms exist for these types of problems, though none guarantees that the global optimum will be found (Ghosh et al. 2005). In this work we compared several optimization routines to determine which are best suited for our particular application, using as criteria the speed of calculation and ability to minimize the objective function. We chose the first set of data obtained, wild-type culture at medium growth rate, to use in the evaluation procedure.

The most commonly used optimization routine for flux analysis is the genetic strategy, which creates diversity in a “population” of flux distributions through small changes in the parameter

values (mutation) or combination of parameters from two different "parent" flux distributions (Christensen and Nielsen 2000; Gombert et al. 2001). Beginning with a randomly generated population, the cells with lowest objective value are selected in each generation while those with high objective value are removed to maintain a constant population size. In our case the initial population was not completely random, since each flux distribution was required to satisfy the overall metabolite balance S*v=0. Thus for each randomly generated set of fluxes, we chose the nearest feasible flux distribution as a member of the population. For the *M. extorquens* reduced network, we previously estimated a required population size of at least 10,000 to ensure representative sampling of the solution space (Van Dien et al. 2003). The Matlab isotopomer balance requires about 10 seconds per function call, so use of a population this large required over 24 hours computer time per generation. To reduce the time required, in the first generation the best 100 out of 10,000 points were chosen, and the subsequent generations were carried out using the population of 100. The routine consistently converged to a solution within 100 generations, requiring 1.5-2 days.

We next tried the FMINCON function in the Matlab Optimization Toolbox, which is a gradient-based method that converges to a local optimum in about 4 hours. Alone this is not a reasonable option because the solution will vary depending upon the initial guess provided. To have a good chance of finding a global solution, we would need to run the minimization many times with different starting points. This would require many days, so the use of FMINCON alone was not explored further. However, we then considered a hybrid approach coupling FMINCON to the evolutionary strategy. First, this function was applied to the final result of the evolutionary algorithm. No improvement was made in the objective value, indicating that a local solution had already been found with the evolutionary algorithm. After realizing that the CONOPT2 (see below) was much faster than FMINCON, we abandoned this approach.

The large CPU times associated with the purely genetic strategy and hybrid approach are expected to become even more prohibitive in future applications where larger networks will be investigated. We thus explored the possibility of employing the high level modeling system, GAMS (Brooke 1998a; Brooke 1998b), which specializes in large-scale optimization problems. The entire problem was formulated in GAMS with the fluxes, IDVs, and MDVs serving as variables, and the isotopomer balances and mappings between the IDVs and MDVs as constraints. Three nonlinear solvers, BARON, CONOPT, and MINOS were tested to solve for the flux distribution that minimizes the value of the objective function (i.e., the sum-of-squares difference between the simulated and experimental mass distribution vectors). The global optimization solver, BARON, was evaluated first as this solver had shown success in tackling a related problem for a smaller network with one experimental observable (Ghosh et al. 2005). However, the network size and larger number of observables present in the *M. extorquens* model rendered BARON unable to find a solution for this application. We next explored the gradient-based solvers, MINOS and CONOPT, which require a multiple starting point procedure because these solvers converge to locally, not globally, optimal solutions (Figure 3). For each restart, the procedure employed here involves randomizing the fluxes between the upper and lower bounds calculated via flux variability analysis, solving for the closest feasible flux distribution, calculating the IDVs associated with this flux distribution, and providing the calculated fluxes and IDVs to MINOS or CONOPT as an initial guess. Unlike the FMINCON function in Matlab, both solvers accessed via GAMS are able to find locally optimal solutions on the order of

seconds and, in most cases, take less than a few minutes of restarts to find their best solution. Nevertheless, 1000 restarts are provided for each case to ensure a thorough sampling of potential solutions. All flux distributions from this SBIR were independently calculated using the GAMS and Matlab-based approaches, and in no case was Matlab able to outperform (i.e., in terms of CPU time or smaller objective value) the GAMS-based strategy. The approach was also successfully applied to a more complex network (~200 reactions) for *Escherichia coli* with similar results providing evidence that the procedure is scalable for larger applications.

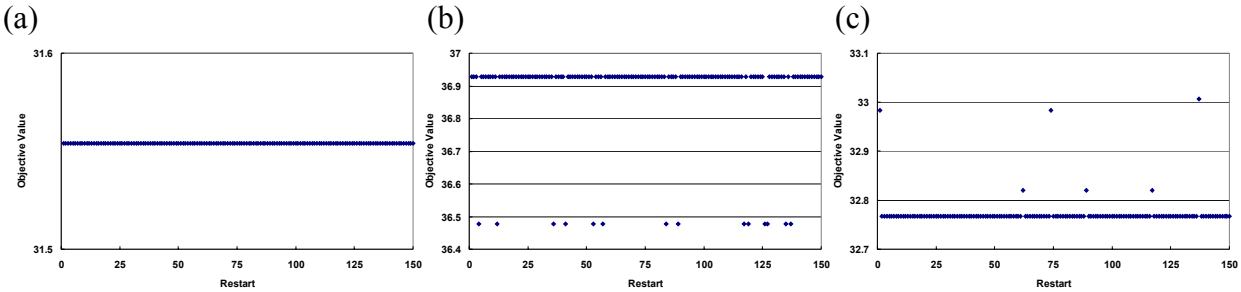


Figure 3: The presence of one, two, or multiple local optimal solutions are common in non-convex nonlinear optimization problems motivating the need for the multiple restart procedure. Figure (a) shows the solutions found for the wild-type, slow growth case, (b) wild-type, medium growth, and (c) the *fdh* mutant. For each case, the solution with the smallest objective function value was found after only a few iterations.

C.3.6 Summary of Computational Procedures

In summary, our work under this grant resulted in a number of significant improvements over the existing state-of-the-art computational methods in ^{13}C flux analysis.

1. Construction of a stoichiometric matrix specific for ^{13}C-labeling studies directly from a genome-scale model.
2. Rapid chemoinformatics-based method for calculation of isotopomer mapping matrices.
3. Generalized isotopomer balance routine that can be applied to any stoichiometric matrix and set of IMMs.
4. Found an efficient optimization strategy that can be applied to larger networks.

Combined with an efficient optimization routine, these steps provide the framework for eventual implementation within the SimPheny modeling platform.

C.4 Aim 4: Compare *in silico* Predictions with Measured Flux Distributions

Flux distributions for each of the experiments were calculated using the methods of section C.3. (Table 5). Since the simplified network has 4 degrees of freedom, the entire distribution of net fluxes can be defined by 4 carefully chosen "free fluxes" (Wiechert et al. 1999). Thus comparison of these key net fluxes is a compact means to show comparison of the entire flux distribution. Furthermore, since the isotopomer balance does not include the formaldehyde oxidation pathways, fluxes are calculated relative to the total amount of carbon entering the serine cycle through the glycine hydroxymethyltransferase reaction (GHMT), rather than the total amount of methanol entering the system. Predicted fluxes were likewise normalized, so that all fluxes in Table 5 are in units of mmol per 10 mmol carbon flux to serine cycle.

Table 5: Predicted vs. Measured Values of the "Free Fluxes": entries are Predicted / Measured (95% Confidence Interval)

Strain / Condition	α-KG DH	PDH	ME	G6P DH
Wild-type AM1 / methanol (μ=0.052)	0 / 0 (0.06)	0 / 0 (0.09)	0 / 0 (0.11)	0 / 0 (0.04)
Wild-type AM1 / methanol (μ=0.017)	0 / 0.83 (0.03)	0 / 0 (0.05)	0 / 0 (0.06)	0 / 0 (0.02)
AM1 pDH80 / methanol	0 / 0.21 (0.63)	0 / 0 (1.64)	0 / 0 (1.10)	0 / 5.55 (0.43)
Wild-type AM1 / methylamine	0 / 0 (0.48)	0 / 0.25 (0.13)	0 / 0 (0.45)	0 / 3.90 (0.15)
Δfdh3 / methanol	7.74 / 0 (0.44)	0 / 4.81 (0.17)	0 / 0 (0.09)	0 / 3.93 (0.37)

Abbreviations: α-KG DH, α-ketoglutarate (2-oxoglutarate) dehydrogenase; PDH, pyruvate dehydrogenase; ME, malic enzyme; G6P DH, glucose-6-phosphate dehydrogenase

The objective of this section was to determine the ability of the *in silico M. extorquens* AM1 model to predict the experimental results using flux balance analysis. Lack of agreement would indicate that such experiments are necessary to fully characterize metabolism. We defined agreement if the confidence intervals of the measured fluxes come within 10% of the predicted fluxes. Confidence intervals for key fluxes were calculated using established methods (Wiechert and de Graaf 1997). These intervals reflect the sensitivity of the objective function at the optimum to small changes in the free fluxes, as well as the standard deviations of the isotopomer measurements themselves, and can differ widely from one experiment to the next. It should be noted that we ran all simulations with both the growth rate and methanol uptake rate constrained to those values measured in the chemostat. The predictive power of the unconstrained models was already explored in section C.2.

We achieved excellent agreement for the wild-type culture at medium growth rates, and somewhat less agreement at low growth rates. It appears that there is a slight change in physiology at very low growth rates, and the cell utilizes the complete TCA cycle to a small extent. The model does not reflect this behavior, predicting that the flux distribution should be just a scaled-down version of that at medium or high growth rates.

Due to the poor agreement between the predicted and measured fluxes in cases other than the wild-type, it is clear that a purely theoretical approach is not always sufficient to understanding cell metabolism. This demonstrates the value of a combined theoretical/experimental approach to studying physiology and metabolism, and supports the need for developing such a capability within the SimPheny framework. The integration of experimental data to improve model predictions is addressed in the next section.

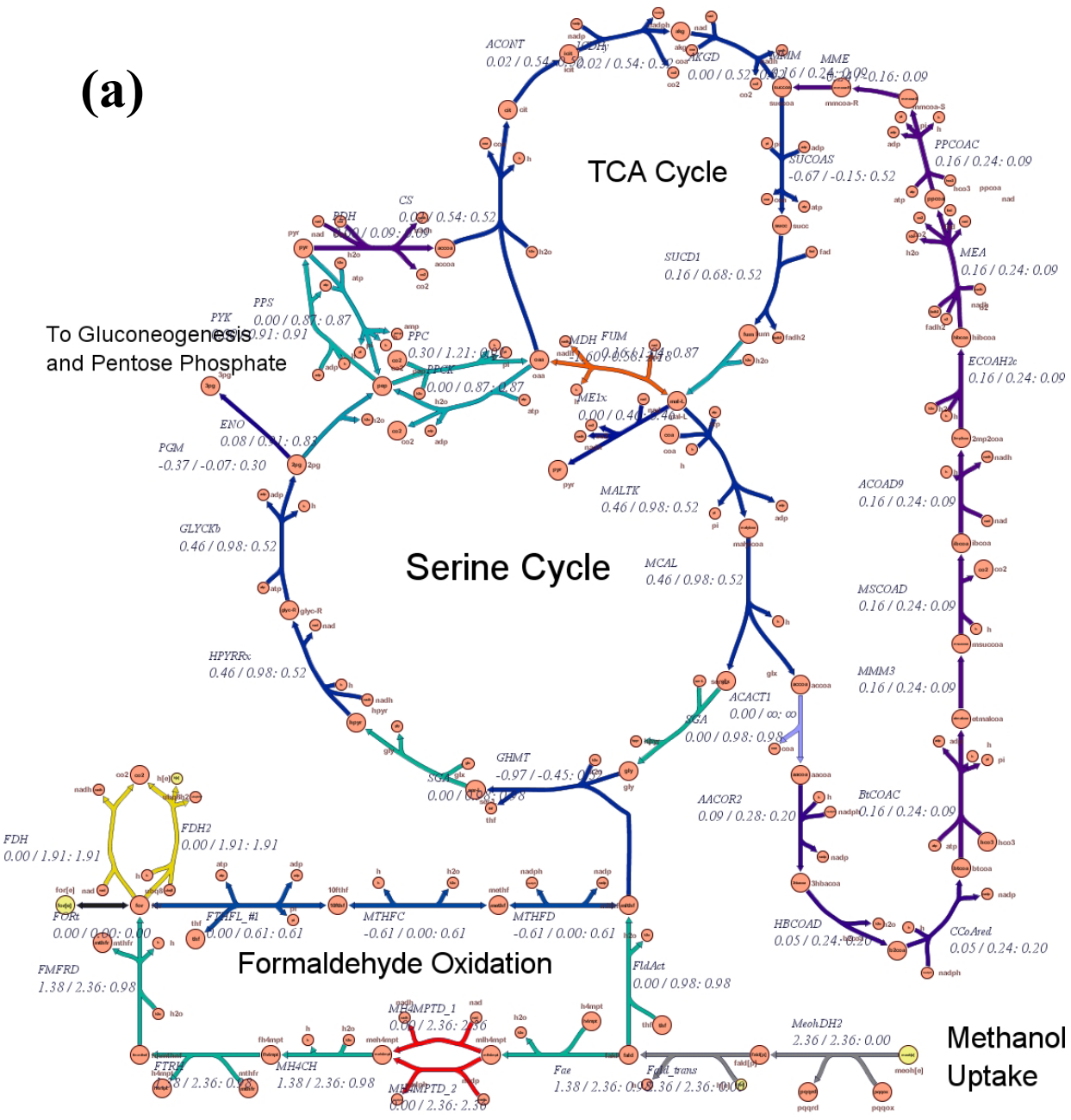
C.5 Aim 5: Use Experimental Data as Constraints on the Network

As the final objective of Phase I, here we demonstrate the ability to integrate experimental data with the constraints-based modeling techniques, for the creation of a powerful computational/ experimental approach to flux analysis.

C.5.1 Incorporate Measured Fluxes as Quantitative Constraints

Although the reduced system of equations developed for the simplified model is completely determined once the free fluxes are specified, such is not the case for the full model due to the addition of biosynthesis, formaldehyde oxidation pathways, and various energy and redox-associated reactions. The measured free fluxes were added as constraints (upper and lower bounds as 95% confidence intervals), serving to reduce the size of the solution space and thus narrow down the range of possible phenotypes predicted by the model. In addition, constraints on growth rate and methanol uptake rate were set from the chemostat data as discussed in section C.3. As mentioned above, all free fluxes measured and shown in Table 5 are relative to a fixed flux of 10 mmol/hr-gDCW through GHMT, which is the reaction through which single carbon compounds enter central metabolism. The flux of GHMT relative to the total methanol flux is unknown due to the formaldehyde oxidation pathways, and is in fact something we would like to learn using this approach. Therefore, we fixed the ratio of free fluxes to GHMT flux to agree with the measurements, rather than fixing the free fluxes themselves. This complication is unique to methylotrophs, and would not be an issue when dealing with organisms growing on sugars or organic acids.

Flux variablilty analysis was then applied to the constrained system, calculating a range of flux values for each reaction that can be adopted and still give a feasible solution. Thus bounds are set within which each flux must fall. It should be emphasized that these ranges are based only on the stoichiometry of the system and the applied constraints, and do not rely on optimization of any objective function. This analysis was performed for all cases, but for brevity is only reported here for the wild-type slow growth example. The results for the fully constrained system (including the flux analysis data) were compared to those in which the only constraints were growth rate and methanol uptake rate (Figure 4). Of particular importance are the formaldehyde oxidation pathways, near the bottom of the figure. Formaldehyde is produced from methanol by the methanol dehydrogenase complex in the periplasm, and is consumed in the cytoplasm. The formaldehyde in the cytoplasm reacts with two pools of folate compounds, tetrahydrofolate (H_4F) and tetrahydromethanopterin (H_4MPT), to generate methylene-folate derivatives (Chistoserdova et al. 1998). Each of these methylene adducts is then involved in further reactions, either for incorporation into cell material via the serine cycle or for energy metabolism, by oxidation to carbon dioxide (Chistoserdova, et al. 1998). Although these fluxes are not measured by ^{13}C-label tracing analysis, it is clear from the figures that by performing these experiments, they can be confined to a smaller range than by using the chemostat process data alone. This is because they are extremely important for energy and reducing potential generation, and the demands for ATP and NAD(P)H are set based on the fluxes in the remaining portions of metabolism.


(a)
TCA Cycle
Serine Cycle
Formaldehyde Oxidation
Methanol Uptake
To Gluconeogenesis and Pentose Phosphate
ACONT
0.02 / 0.54: 0.52
IKGD
0.00 / 0.52: 0.52
MME
0.16 / 0.24: 0.09
SUCOAS
-0.67 / -0.15: 0.52
PPCOAC
0.16 / 0.24: 0.09
CS
0.02 / 0.54: 0.52
PDH
0.00 / 0.09: 0.09
SUCD1
0.16 / 0.68: 0.52
MEA
0.16 / 0.24: 0.09
PPS
0.00 / 0.87: 0.87
PYK
0.91 / 0.91
PPC
0.30 / 1.21: 0.91
PPCK
0.00 / 0.87: 0.87
MDH
FUM
0.87
ME1x
0.00 / 0.40: 0.40
ECOAH2c
0.16 / 0.24: 0.09
ENO
0.08 / 0.91: 0.83
PGM
-0.37 / -0.07: 0.30
MALTK
0.46 / 0.98: 0.52
ACOAD9
0.16 / 0.24: 0.09
GLYCKb
0.46 / 0.98: 0.52
MCAL
0.46 / 0.98: 0.52
MSCOAD
0.16 / 0.24: 0.09
HPYRRx
0.46 / 0.98: 0.52
MMM3
0.16 / 0.24: 0.09
ACACT1
0.00 / ∞: ∞
SGA
0.00 / 0.98: 0.98
GHMT
-0.97 / -0.45: 0.52
SGT
0.00 / 0.00: 0.98
AACOR2
0.09 / 0.28: 0.20
BtCOAC
0.16 / 0.24: 0.09
FDH
0.00 / 1.91: 1.91
FDH2
0.00 / 1.91: 1.91
FORt
0.00 / 0.00: 0.00
FTHFL_#1
0.00 / 0.61: 0.61
MTHFC
-0.61 / 0.00: 0.61
MTHFD
-0.61 / 0.00: 0.61
HBCOAD
0.05 / 0.24: 0.20
CCoAred
0.05 / 0.24: 0.20
FMFRD
1.38 / 2.36: 0.98
FldAct
0.00 / 0.98: 0.98
MEHMPTD_1
0.00 / 2.36: 2.36
MeohDH2
2.36 / 2.36: 0.00
FTR18
1.38 / 2.36: 0.98
MH4CH
1.38 / 2.36: 0.98
MEHMPTD_2
0.00 / 2.36: 2.36
Fae
1.38 / 2.36: 0.98
Fald_trans
2.36 / 2.36: 0.00

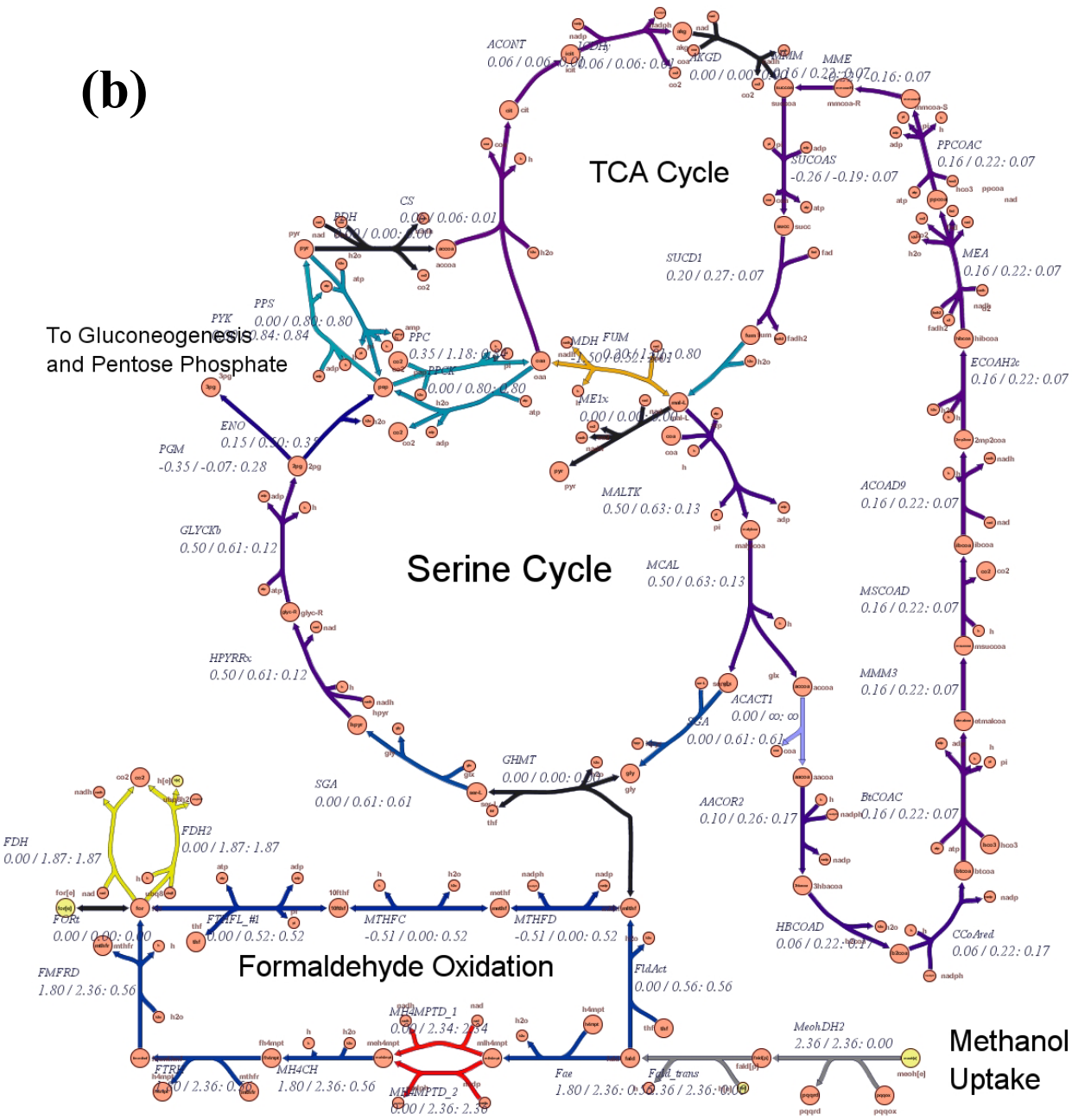


Figure 4: Range of possible flux distributions for wild-type *M. extorquens* AM1 under slow growth conditions (μ=0.017 hr^{-1}), as determined by flux variability analysis. (a) constrained only by chemostat process measurements from Table 4; (b) constrained by both chemostat data and free fluxes in Table 5. Results are expressed as: minimum flux / maximum flux : range.

C.5.2 Incorporate Gene Expression Data as Qualitative Constraints

Gene expression analysis was performed under a separate project in the Lidstrom laboratory. Chemostat cultures of wild-type *M. extorquens* AM1 grown both on methanol and succinate were sampled, and the relative gene expression levels between the two conditions were measured by microarray analysis. Genomatica received data in a spreadsheet listing each gene locus number followed by the $\log_{10}$ ratio of expression on succinate relative to methanol, and SimPheny was used to visualize this information on a metabolic map (Figure 5). As the Gene Expression Module of SimPheny is not intended to be a statistical analysis package, this data had already been pre-processed using available tools to remove unreliable data points. The utility of SimPheny is in the analysis of this processed data in the context of the genome-scale model, and in using this data to constrain the metabolic network. In this work, relative expression data was used to help characterize the flux distribution in cells growing on succinate, based on what is already known about growth on methanol. Starting with the predicted flux distribution for methanol-grown cells with μ=0.052 hr^{-1}, fluxes in the succinate simulation (using the same growth rate) were constrained in the changes that can occur between the two conditions. For example, if a gene's expression level decreases from one condition or strain to the next, then the reaction flux level calculated for the first experiment will be set as the maximum flux level for simulation of the second experiment. The simulation was performed with minimization of succinate uptake as the objective (Figure 6). The predicted succinate uptake rate was 1.295 mmol/hr-gDCW. This is in excellent agreement with a chemostat experiment performed in the Lidstrom lab, in which the uptake rate was measured to be 1.35 mmol/hr-gDCW. Finally, flux variability analysis was used to explore the constrained solution space. The results demonstrate that the entire flux distribution is very tightly constrained by imposing constraints from fermentation and gene expression data. The only fluxes of central metabolism showing any variability (range of 0.124 for all) are pyruvate kinase, PEP carboxykinase, malate dehydrogenase, and malic enzyme.

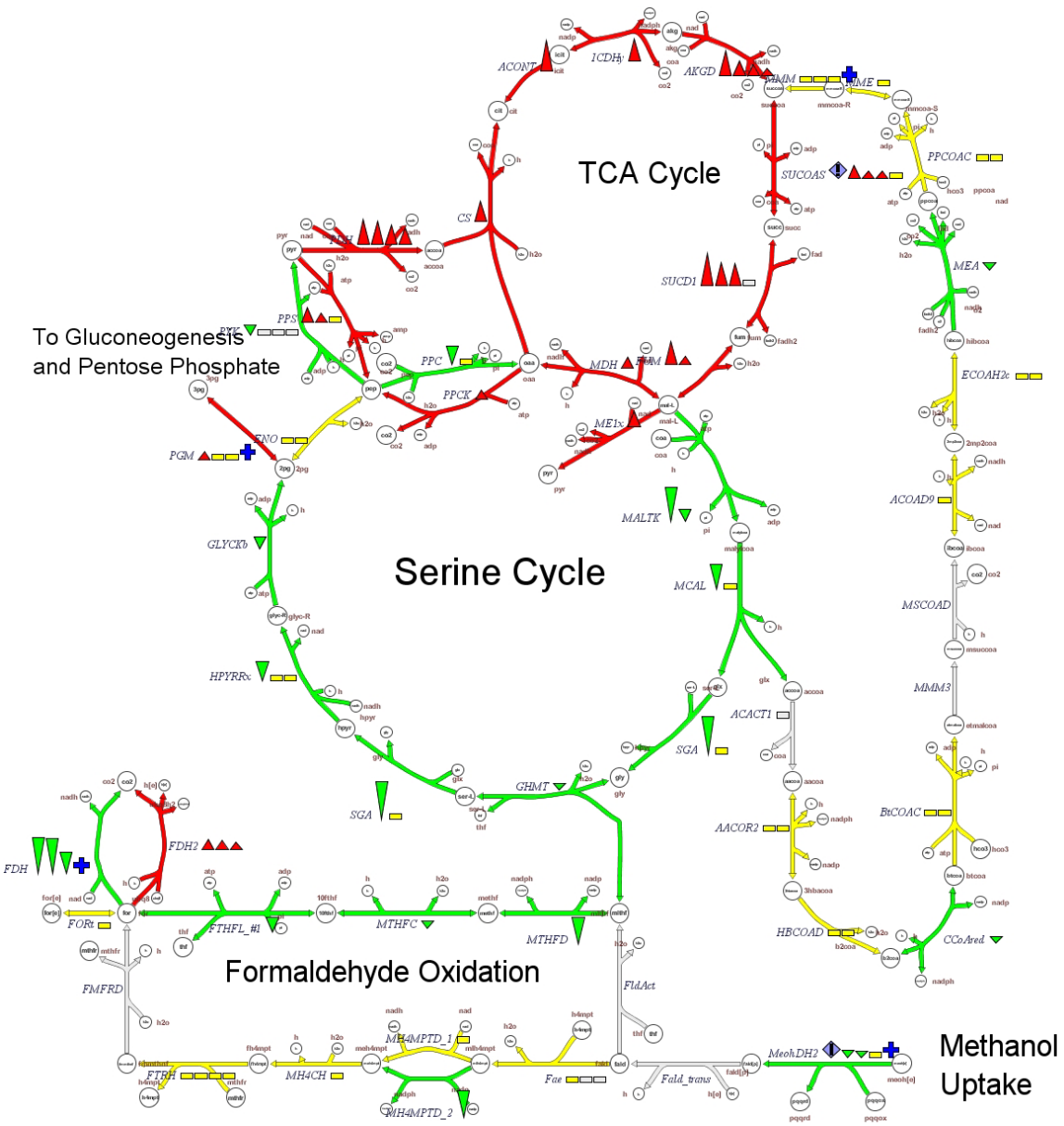


Figure 5: Predicted changes in flux distribution for *M. extorquens* AM1 grown on succinate relative to methanol, derived from $\log_{10}$ ratios of gene expression between the two conditions. Triangles show directional changes in each gene, whereas colors of arrows indicate predicted change in flux based on the gene expression changes. Green, decrease; red, increase; yellow, no change. Threshold was set to 0.15, meaning that log ratios between –0.15 and 0.15 were not considered to be significant.

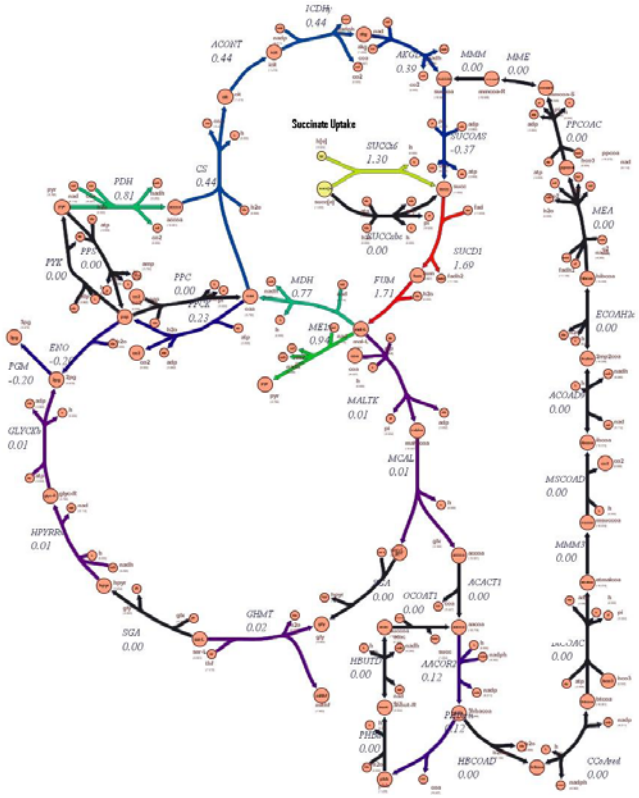


Figure 6: Predicted flux distribution for succinate grown cells with μ=0.052 hr^{-1}, after applying constraints derived from the relative gene expression data shown in Figure 6, with a threshold of 0.15. Objective function was the minimization of succinate uptake rate. Formaldehyde oxidation pathways are not shown because the fluxes are all equal to zero.

C.6 Summary of Phase I

We set forth an ambitious list of tasks for Phase I, focusing on developing a computational strategy for the analysis of ^{13}C-label tracing data, and demonstrating the commercial potential of this method in conjunction with our SimPheny modeling platform. In summery, we achieved everything that we had set out to do, including the following:

- Performed a series of six ^{13}C-labeling experiments with chemostat cultures of *M. extorquens* AM1.
- Constructed a genome-scale model for *M. extorquens* AM1, used it to make predictions of growth phenotype with a variety of strain/condition combinations, and compared predictions to process data from the chemostat.
- Used process data as constraints to reduce the size of the feasible solution space, thus improving the predictive power of the model.
- Developed and optimized a computational protocol for the calculation of intracellular fluxes from isotopomer labeling data (see section C.3.6).

- Calculated intracellular fluxes and compared to predictions. In not all cases was there agreement, which shows that flux analysis provides information critical to understanding physiology and metabolism.
- Used experimental data from both isotopomer labeling experiments and microarrays to constrain simulations, thus improving predictive capability of the model.

C.7 References

Attwood MM, Harder W. 1972. A rapid and specific enrichment procedure for *Hyphomicrobium* spp. Antonie Van Leeuwenhoek 38:369-377.

Bonarius HPJ, Schmid G, Tramper J. 1997. Flux analysis of underdetermined metabolic networks: The quest for the missing constraints. Trends Biotechnol. 15(308-314).

Brooke A. 1998a. GAMS: A User's Guide. Washington, D.C.: GAMS Development Corporation.

Brooke A. 1998b. GAMS: The Solver Manuals. Washington, D.C.: GAMS Development Corporation.

Chistoserdova L, Laukel M, Portais JC, Vorholt JA, Lidstrom ME. 2004. Multiple formate dehydrogenase enzymes in the facultative methylotroph Methylobacterium extorquens AM1 are dispensable for growth on methanol. J Bacteriol 186(1):22-8.

Chistoserdova L, Vorholt JA, Thauer RK, Lidstrom ME. 1998. C1 transfer enzymes and coenzymes linking methylotrophic bacteria and methanogenic archaea. Science 281:99-102.

Christensen B, Nielsen J. 2000. Metabolic network analysis of *Penicillium chrysogenum* using ^{13}C-labeled glucose. Biotechnol. Bioeng. 68(6):652-659.

Dauner M, Bailey JE, Sauer U. 2001. Metabolic flux analysis with a comprehensive isotopomer model in *Bacillus subtilis*. Biotechnol. Bioeng. 76(2):144-156.

Edwards JS, Palsson BO. 1999. Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. J. Biol. Chem. 274(25):17410-17416.

Edwards JS, Palsson BO. 2000. The *Escherichia coli* MG1655 *in silico* metabolic genotype: Its definition, characteristics, and capabilities. Proc. Nat. Acad. Sci. USA 97(10):5528-5533.

Fischer E, Sauer U. 2003. Metabolic flux profiling of Escherichia coli mutants in central carbon metabolism using GC-MS. Eur. J. Biochem. 270:880-891.

FitzGerald KA, Lidstrom ME. 2003. Overexpression of a heterologous protein, haloalkane dehalogenase, in a poly-β-hydroxybutyrate-deficient strain of the facultative methylotroph *Methylobacterium extorquens* AM1. Biotechnol Bioeng 81(3):263-268.

Ghosh S, Zhu T, Grossmann IE, Ataai MM, Domach MM. 2005. Closing the loop between feasible flux scenario identification for construct evaluation and resolution of realized fluxes via NMR. Computers & Chem. Eng. 29:459-466.

Gombert AK, Moreira dos Santos M, Christensen B, Nielsen J. 2001. Network identification and flux quantification in the central metabolism of *Saccharomyces cerevisiae* under different conditions of glucose repression. J. Bacteriol. 183(4):1441-1451.

Karp PD. 1998. Metabolic databases. Trends Biochem. Sci. 23(3):114-116.

Klamt S, Schuster S, Gilles ED. 2002. Calculability analysis in underdetermined metabolic networks illustrated by a model of the central metabolism in purple nonsulfur bacteria. Biotechnol. Bioeng. 77:734-751.

Mahadevan R, Schilling CH. 2003. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab. Eng.

Marx A, de Graaf AA, Wiechert W, Eggeling L, Sahm H. 1996. Determination of the fluxes in central metabolism of *Corynebacterium glutamicum* by NMR spectroscopy combined with metabolite balancing. Biotechnol. Bioeng. 49:111-129.

Marx CJ, Lidstrom ME. 2001. Development of improved versatile broad-host-range vectors for use in methylotrophs and other Gram-negative bacteria. Microbiology.

Palsson BO. 2000. The challenges of in silico biology. Nat. Biotechnol. 18:1147-1150.

Palsson BO. 2002. In silico biology through "omics". Nat. Biotechnol. 20:649-650.

Phalakornkule C, Fry B, Zhu T, Kopesel R, Ataai MM, Domach MM. 2000. 13C NMR evidence for pyruvate kinase flux attenuation underlying suppressed acid formation in Bacillus subtilis. Biotechnol Prog 16(2):169-75.

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(4):398-421.

Reed JL, Vo TD, Schilling CH, Palsson BO. 2003. An expanded genome-scale model of *Escherichia coli* K-12 (*i*JR904 GSM/GPR). Genome Biology 4:R54.

Riascos CAM, Gombert AK, Pinto JM. 2005. A global optimization approach for metabolic flux analysis based on labeling balances. Computers & Chem. Eng. 29:447-458.

Sauer U, Lasko DR, Fiaux J, Hochuli M, Glaser R, Szyperski T, Wuthrich K, Bailey JE. 1999. Metabolic flux ratio analysis of genetic and environmental modulations of *Escherichia coli* central carbon metabolism. J. Bacteriol. 181(21):6679-6688.

Savinell JM, Palsson BO. 1992. Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism. J. Theor. Biol. 154:421-454.

Schilling CH, Edwards JS, Palsson BO. 1999. Toward metabolic phenomics: analysis of genomic data using flux balances. Biotechnol. Prog. 15:288-295.

Schilling CH, Palsson BO. 2000. Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis. J. Theor. Biol. 203(3):249-283.

Schmidt K, Carlsen M, Nielsen J, Villadsen J. 1997. Modeling isotopomer distributions in biochemical networks using isotopomer mapping matrices. Biotechnol. Bioeng. 55(6):831-840.

Segre D, Vitkup D, Church GM. 2002. Analysis of optimality in natural and perturbed metabolic networks. Proc. Nat. Acad. Sci. USA 99(23):15112-15117.

Selkov E, Maltsev N, Olsen GJ, Overbeek R, Whitman WB. 1997. A reconstruction of the metabolism of *Methanococcus jannaschii* from sequence data. Gene 197:GC11-26.

Vallino JJ, Stephanopoulos G. 1993. Metabolic fluc distributions in *Corynebacterium glutamicum* during growth and lysine overproduction. Biotechnol. Bioeng. 41(6):633-646.

Van Dien SJ, Lidstrom ME. 2002. Stoichiometric model for evaluating the metabolic capabilities of the facultative methylotroph *Methylobacterium extorquens* AM1, with application to reconstruction of C3 and C4 metabolism. Biotechnol. Bioeng. 78(3):296-312.

Van Dien SJ, Strovas T, Lidstrom ME. 2003. Quantification of central metabolic fluxes in the facultative methylotroph *Methylobacterium extorquens* AM1 using ^{13}C-label tracing and mass spectrometry. Biotechnol. Bioeng. 84(1):45-55.

Varma A, Palsson BO. 1994. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild type *Escherichia coli*. Appl. Environ. Microbiol. 60(10):3724-3731.

Wiechert W. 2001. ^{13}C metabolic flux analysis. Metab. Eng.(3):195-206.

Wiechert W, de Graaf AA. 1997. Bidirectional reaction steps in metabolic networks: I. Modeling and simulation of carbon isotope labeling experiments. Biotechnol. Bioeng. 55(1):101-117.

Wiechert W, Mollney M, Isermann N, Wurzel M, de Graaf AA. 1999. Bidirectional reaction steps in metabolic networks: III. Explicit solution and analysis of isotopomer labeling systems. Biotechnol. Bioeng. 66:69-85.

Wiechert W, Siefke C, de Graaf AA, Marx A. 1997. Bidirectional reaction steps in metabolic networks: II. Flux estimation and statistical analysis. Biotechnol. Bioeng. 55(1):118-135.

Wittmann C, Heinzle E. 1999. Mass spectrometry for metabolic flux analysis. Biotechnol. Bioeng. 62(6):739-750.

Wittmann C, Heinzle E. 2001. Modeling and experimental design for metabolic flux analysis of lysine-producing *Corynebacteria* by mass spectrometry. Metab. Eng. 3:173-191.