THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF CHEMICAL ENGINEERING

# DEVELOPMENT AND APPLICATION OF THE ANNOTATION INTEGRATION METHOD FOR DEVELOPING GENOME-SCALE MODELS

THOMAS J. MUELLER
FALL 2012

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Chemical Engineering
with honors in Chemical Engineering

Reviewed and approved* by the following:

Costas Maranas
Donald B. Broughton Professor of Chemical Engineering
Thesis Supervisor

Ali Borhan
Professor of Chemical Engineering
Honors Adviser

* Signatures are on file in the Schreyer Honors College.

# ABSTRACT

There are an increasing number of organisms for which a developed genome scale metabolic reconstruction would be beneficial, but which do not have a high level of manual genome annotation. This paper presents the Annotation Integration Method, which uses a combination of phylogenetically similar models and additional sources for reliable gene annotations to develop accurate metabolic reconstructions. This method does not introduce errors that may arise from the use of homology searches separate from manual oversight. The efficacy of this method was displayed through the development of reconstructions for 5 species of *Cyanothece* that have important inherent metabolic qualities. Statistics from the development of the model for *Cyanothece* PCC 7424 were used to illustrate this process.

# TABLE OF CONTENTS

# LIST OF FIGURES

# ACKNOWLEDGEMENTS

The author would like to thank all parties involved in this thesis and undergraduate education. Dr. Costas Maranas provided inspiration and support for the research published here. The author would lastly like to recognize the sources of this educational opportunity: the Department of Energy for providing funding; and The Schreyer Honors College and Pennsylvania State University's Department of Chemical Engineering

# Chapter 1

## Introduction

The number of, and applications for, genome-scale metabolic reconstructions have been growing significantly. Reconstructions allow for increased investigation into current organism functions. A well-curated model will provide insights into the essentiality of genes through simulated gene deletions, as well as efficiently determining synthetic lethals, which are combinations of non-essential gene knockouts that when implemented at the same time, prohibit growth [1]. Metabolic models can be used to identify changes to the growth medium that would allow for higher product yields [2]. Reconstructions can also be used to study the implementation or improvement of functions foreign to the organism. When coupled with optimization programming, strategies for metabolite overproduction [3, 4] and novel pathway inclusion [5] can be developed. Likewise, previously developed metabolite production strategies can be further refined with the use of a metabolic model [6].

A genome scale metabolic reconstruction is the assembly of known reactions present within the metabolism of an organism, including information such as localization of metabolites along with enzyme and gene function [7]. An annotated genome, combined with biochemical evidence from literature forms the basis for original metabolic reconstructions [8]. The overall metabolic network is defined through the restrictions placed on reaction fluxes through constraint-based modeling [9]. Not only must a reaction be determined to be present within the organism's metabolism, but the reaction must be both charge and elementally balanced to assure faithful replication of the overall stoichiometry [10]. Metabolic reconstructions contain a detailed

biomass equation, typically developed for the specific organism, or a closely related species. A biomass equation accounts for all internal metabolites that constitute biomass [11]. A final crucial component of a metabolic reconstruction is the proper development of the gene-protein-reaction (GPR) relationships. These GPR relationships elaborate on which genes must be transcribed and translated to produce each protein, and which reactions those proteins then catalyze. The proteins in these relationships are most commonly recorded using their Enzyme Commission or EC number. Genome scale models are never truly finished; new biological discoveries and improved modeling algorithms should frequently be applied to expand and refine reconstructions [12-14].

Once a reconstruction's reactions and constraints are defined, the flow of metabolites through the metabolism is evaluated with the use of flux balance analysis (FBA) [15]. FBA is used to mathematically represent the constraints placed on an actual metabolism. It specifies the maximization of biomass and imposes a steady state mass balance constraint, in that any internal metabolite that is produced by a reaction is also used by another [15, 16]. The formulation of this steady state is enumerated in more detail in the methods section.

As previously stated, metabolic models are useful in a predictive capacity, most prominently in the overproduction of target metabolites. Once an accurate and complete metabolic reconstruction is developed, numerous methodologies can be applied, such as OptForce, [4]which can be used to develop a set of organism modifications required to reach a designated production level for a target metabolite [5].

While there are a select number of organisms for which there exists an abundance of published literature, most organisms have comparatively little published work. Many of the metabolic reconstructions that have been built up to this point have been constructed

for model organisms, or for those organisms that already have a large amount of published literature. Certain organisms, such as *Escherichia coli* and *Saccharomyces cerevisiae* already have entire databases devoted to annotations of their genes [17-19]. There are a number of databases that contain annotations for a wide variety of organisms [20-23]. However, even in these databases the more heavily researched organisms have a larger number of reliable annotations. *E. coli K*-12, the strain that was modeled in the *i*AF1260 model [13], has approximately 16 times the number of reviewed annotations (4,326) in the Universal Protein Resource (Uniprot), as does the *Cyanothece* strain with the most reviewed annotations, *Cyanothece* PCC 7424 (271) [20].

Despite the comparative lack of research into other organisms, there are still a number of reconstructions created for organisms outside the restricted group of model organisms. Many of these reconstructions pull largely from a single database, such as KEGG, for their genome annotations [2, 24-27]. There have been a number of noted discrepancies within the KEGG database, such as multiple labels for a single metabolite and unbalanced or generic reactions [24]. This highlights an inherent risk with using only one database. There is the possibility for errors within the annotations, which if not detected, increase the risk of a large number of difficult-to-locate errors arising during the reconstruction process.

Sequencing technology is continually progressing in speed and accuracy, and the number of sequenced genomes is growing at an increasing rate. There is insufficient time to conduct research on each gene in newly sequenced genomes, but the function of those genes must still be determined to develop a genome-scale metabolic reconstruction. This lack of confirmed annotations was the impetus for the creation of the Annotation Integration Method outlined in this paper. Accurate genome scale metabolic reconstructions must be developed for newly prioritized

species that do not yet have an extensive collection of published literature. The Annotation Integration Method is designed to stringently select accurate annotations from a number of separate reconstructions and databases to acquire the GPR associations that will constitute the model.

This paper focuses on five species of *Cyanothece*, PCC 7424, 7425, 7822, 8801, and 8802 to use as examples. These five species all have promising industrial applications that could be enhanced through metabolic modeling, but they all lack direct literature evidence of most gene functions. The genus *Cyanothece* belongs to the phylum of Cyanobacteria. Cyanobacteria have a number of inherent metabolic properties that make them ideal candidates for use in an industrial setting. First, they are photosynthetic [28]; in fact cyanobacteria generally have a higher solar energy conversion efficiency, ranging from 3-9%, than that of C3 plants (2.4%) and C4 plants (3.7%) [29]. This has the potential to greatly restrict costs for industrial applications as low energy carbon in the form of carbon dioxide can be fed to reactors as opposed to higher energy substances such as glucose [30]. Secondly *Cyanothece* species contain the metabolic capability to produce a valuable fuel in hydrogen [28, 31-33]. *Cyanothece* possess a number of specific advantages over other hydrogen producing phototrophic microorganisms such as being able to grow in air and be easily fixed to solid matrices [34]. In particular *Cyanothece* strains are capable of fixing nitrogen from the atmosphere by temporally separating the hydrogen production and nitrogenase activities [35]. All five species are capable of fixing nitrogen and producing hydrogen, while *Cyanothece* PCC 7425 is the only one that is not capable of accomplishing this task in an aerobic environment [31]. 7425 also varies in a number of physical characteristics, enough so that it has been suggested that it should be reclassified to another genera pending further genetic review [36].

These five species all exhibit certain key characteristics, but there are a number of important variations between the species that makes the development of five separate reconstructions appropriate. A butanol pathway has been postulated to exist in varying levels of completion through an inspection of the *Cyanothece* genomes [37]. As was shown by Dr. Pakrasi's group at the University of Washington St. Louis, *Cyanothece* 7425 only has one enzyme in the pathway without an identified homolog, while 7822 is missing two enzymes. Additionally there is evidence for the presence of other specific enzymes or pathways, such as type III PHA synthase in *Cyanothece* 7424 and 8801 [38]. Assertions have been made about the varying presence of pathways such as the alkaline biosynthetic pathway and different pathways for breaking down arginine [31]. These differences are the main reason creating metabolic reconstructions for all five species is appropriate. The ability to discern which of the organisms would offer the greatest return on investment is critical.

Despite the immense benefits of genetically modifying an organism that already contains a large number of useful pathways, research being done into these species is still at an early stage. The most annotated *Cyanothece* species in Uniprot, *Cyanothece* 7424, still has only 6% of the annotations that a more highly curated organism in *E. coli K-12* contains. While the amount of research being done into *Cyanothece ATCC 51142* has increased in recent years, very few publications exist that are devoted to one or several of the five species focused on in this paper. Thus they are ideal targets for validating the Annotation Integration Method.

This method derives increased accuracy and reliability from using well curated and verified metabolic models from closely related species when available. In this paper, a previously developed metabolic reconstruction for *Cyanothece* ATCC 51142, *i*Cyt773, was used [39]. *i*Cyt773, with 946 reactions and 811 metabolites, contains 4 compartments, the periplasm,

cytosol, lumen, and thylakoid. All reactions are charge and elementally balanced. A comprehensive GPR was developed for the model, which does not contain any thermodynamically infeasible cycles. The *i*Cyt773 model is an improvement upon the previously published *i*Cce804 model [40]. *i*Cyt773 contains 43 genes and 266 reactions which are unique from the *i*Cce804 model. As illustrated in the phylogenetic tree in Figure 4, *Cyanothece* ATCC 51142, is very closely related to the 5 species, and therefore will share a large portion of its metabolism with them. Given this relationship and the detailed development of the reconstruction, it was deemed a suitable model to be used by the Annotation Integration Method to aide the accuracy and speed at which the new models are developed.

The Annotation Integration Method was applied using the *i*Cyt773 model [39] along with a number of databases to construct models for 5 different *Cyanothece* species, *Cyanothece* PCC 7424, PCC 7425, PCC 7822, PCC 8801, and PCC 8802. This development demonstrates the method's effectiveness as well as provides models for the analysis of promising metabolic features present in some or all of the species. Curation of the five models has not yet been completed but initial results align with expectations derived from phylogenetic comparisons.

# Chapter 2 Methods

## Formation of Intermediate Models

The Annotation Integration Method begins with the retrieval of select reactions from a previously curated model. This set of retrieved reactions will be referred to as the intermediate model. The statistics of the *Cyanothece* 7424 model will be used to illustrate the implementation of the method. Finding reactions from curated models that should be present in

an organism's metabolism is accomplished through a stringent homology search. In other implementations of this method, the number of models used will vary based on what well-curated reconstructions are available. In this case the *i*Cyt773 model of *Cyanothece* ATCC 51142 was used as the starting point.

The stringent homology search is achieved through the implementation of bidirectional BLASTp searches on all genes in the new genome. A BLASTp search is performed for each open reading frame (ORF) present in the new genome against the 51142 genome. An E value cut off of 10^-30 is used in order to help guarantee similarity of function of any found matches. The top result from this forward BLAST is then taken and used as the query sequence for another BLASTp search against the initial genome. If the initial ORF is a result that falls below the E value restriction, it is classified as a bidirectional hit. If the initial query is the top hit for the reverse BLAST search, it is categorized as a bidirectional best hit. The number of bidirectional best hits varied from 1887 shared between 51142 and 7425, as would be expected given the phylogenetic differences between 7425 and the other organisms, and 2648 best hits between 51142 and 8802.

The curation level of the model and specifically its GPR relationships become more important at this stage. Logic gates are used to evaluate each gene association relationship in the model using only bidirectional best hits. If enough genes in the association are bidirectional best hits to satisfy the logic "and" and "or" operators, the linked reaction is deemed to also occur in the metabolism of the new species. The reactions that have a correlation are transferred to form the intermediate model. The gene associations of these transferred reactions are modified to only reflect those genes that are bidirectional best hits. Figure 1 shows examples of this logic gate analysis. As can be seen with rxn00536 and rxn05231 any components of the gene association

that are not satisfied are not transferred. Correctly modifying these associations is vital to the final steps of model curation and application, where gene knockouts are being used either as model verification or proposed modifications for metabolite overproduction. A final logic gate analysis of the remaining gene associations is performed using all bidirectional hits. Any reactions found through this method can either be added to the intermediate model or be reserved for use with a high confidence level in the later GapFill steps. In the case of the 7424 example these reactions were held for later use.

| *i*Cyt 773 Reaction | *i*Cyt773 gene association | Association modified and translated to 7424 genes |
|---|---|---|
| rxn11595 | *cce_1674 and cce_0284 and cce_1710* | PCC7424_3567 and PCC7424_3997 and PCC7424_4217 |
| rxn00536 | *cce_0002 or cce_4888 or cce_1588* | PCC7424_5710 or PCC7424_3972 |
| rxn05231 | *(cce_1354 and cce_4877) or (cce_1354 and cce_1640)* | PCC7424_5337 and PCC7424_5334 |
| MOBDabc | *cce_0578 and cce_0847 and cce_0844* | Reaction does not translate |

**Bidirectional Best Hit**
**Gene Without Best Hit**

**Figure 1: Examples of logic gate analysis for generation of *Cyanothece* PCC 7424 intermediate model.** Bidirectional best hits are shown in green, other genes are shown in red. If the bidirectional best hits satisfy the gene association the satisfied portions are transferred along with the reaction to the intermediate model and translated into 7424 genes using bidirectional best-hit relationships.

Both the E-value cut off and the uses of a bidirectional BLAST are implemented to be conservative in the initial construction of the model. The E value cut off is set at a value low

enough that any genes with different function should not be deemed similar by the homology search. The merits of bidirectional BLAST can be proven from the presence of bidirectional hits that have different top hits in the reverse and forward directions. Such variation undermines some of the confidence in a direct correlation of function. This restrictive nature is vital to keeping unwarranted reactions from inclusion. It is much easier to restore connectivity [41] than it is to search through an entire network to locate a reaction or pathway providing functionality not present within the actual organism. The conservative nature of reaction additions is also necessary in order to help mitigate issues with the creation of thermodynamically infeasible cycles. These cycles are in direct violation of the fact that thermodynamic driving forces around a metabolic loop must add up to zero [42, 43]. Such cycles typically arise from reactions that are not elementally or charge balanced, or which should have irreversible directionality. The methodology for the formation of intermediate models can be seen in Figure 2.

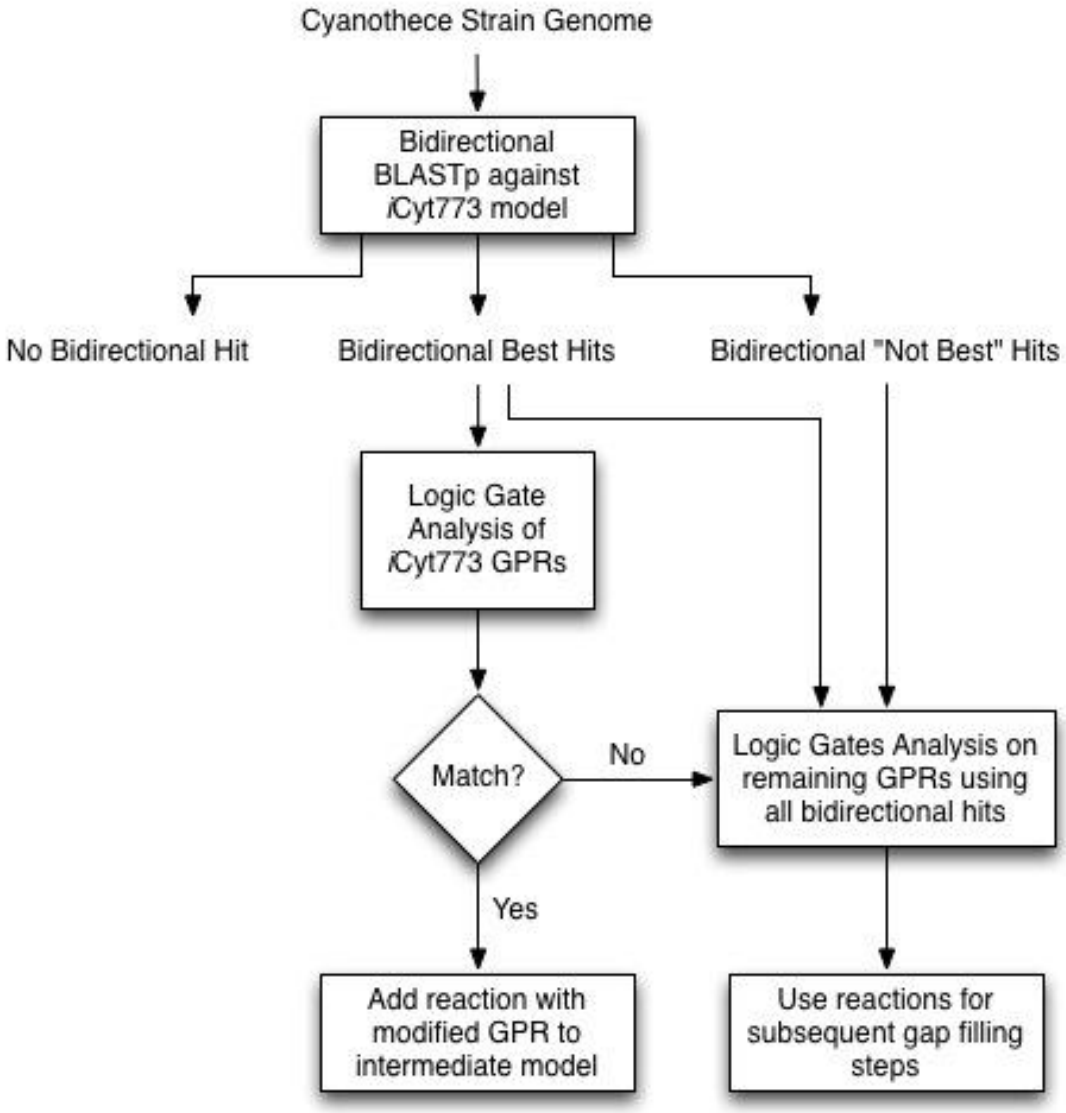



**Figure 2: Flow chart representation of intermediate model development using Annotation Integration Method.** Flow chart designed for specific process of 7424 model creation. For this intermediate model development the *i*Cyt773 model was used. Reactions found using all bidirectional hits in logic gate analysis were restricted to use in GapFill procedure.

## Retrieval of Annotations

While a large portion of the metabolism of each species is already encapsulated in the intermediate models through the bidirectional BLAST reaction transfer process, a number of reactions have not yet been included, especially reactions unique to each species. In order to supplement the reactions already added, genome annotations are retrieved from four different sources, Uniprot, NCBI Protein Clusters, the Pathosystems Resource Integration Center (PATRIC), and the Rapid Annotations using Subsystems Technology (RAST) method [20, 21, 44, 45]. These annotations are used to develop new GPR relationships to help complete the models. Pulling from a variety of different sources is vital to the correct development of the reconstructions. It increases the pool of annotations and therefore helps reduce the probability of an annotation error propagating into the model.

Uniprot [20], was used as a primary resource for annotations in this work. Given the stringent requirements of the method, only reviewed entries in the Uniprot database were considered for inclusion into the reconstructions. In the initial iteration of the workflow, only the manually reviewed entries were retrieved. Of the species, 7822 and 8802 did not have any reviewed entries, while 7424, 7425, and 8801 had 271, 234, and 261 respectively. The NCBI Protein Clusters database contains a collection of protein clusters in varying stages of curation. Function annotations are shared throughout clusters, allowing for the propagation of annotations. Clusters are built through a modified BLAST all-against-all homology search, which results in a set of proteins where each protein has a higher modified BLAST score against any protein in the cluster than any protein not included in the cluster. A number of clusters are then manually curated to correct for things such as miscalled translation initiation sites and to insure no functional divergence within the cluster. Given that there is not a set cutoff for automated cluster

generation and the strict requirements of this method, only clusters that were manually curated, and therefore had a validated (moderate curation) or reviewed (extensive curation) status were considered [44]. Each gene was referenced against all reviewed and validated clusters. For the *Cyanothece* PCC 7424 genome a total of 132 genes were found in validated clusters, while 188 were found in reviewed clusters.

PATRIC is run by the Virginia Bioinformatics Institute. Initially designed to be a database to facilitate research into NIAIAD (National Institute of Allergy and Infectious Diseases) priority pathogens, PATRIC contains consistent annotations for a large number of bacterial species [21]. PATRIC also offers a wide variety of information and forms of presenting data for the organisms in its database. One of these files provides a relationship between genes and EC numbers, a relation vital for the cultivation of the GPR in a metabolic reconstruction. For the example organism, *Cyanothece* 7424, there were 455 genes with assigned E.C. numbers.

The RAST annotation server was utilized to annotate all five of the species' genomes. RAST provides a fully automated annotation of any submitted genome, through the identification of protein encoding genes and subsequent assignment of function utilizing a different set of manually curated protein clusters, referred to as FIGfams [46]. RAST selectively searches the closest phylogenetic neighbors to the submitted genome to both restrict computational cost and error in annotation [45]. Although this method is automated, and therefore allows for possible errors preventable through manual inspection, the manual curation steps involved in the creation of FIGfams increases the confidence in any annotation results. The RAST annotation method works by identifying protein-encoding genes, also referred to as PEGs, present within the genome. It assigns its annotations to these PEGs, as opposed to attributing them to ORFs. The relation between these two will not always be perfect, for example in the 7424 annotation there

are 6238 PEGs and 5710 ORFs, implying differences in splitting of the genome. These variations in splitting of the genome arise in part from the variation in determining protein encoded genes as opposed to open reading frames, as well as the fact that RAST separately identifies tRNA and rRNA genes [45]. A direct comparison of nucleotide sequences was determined to be highly effective at finding a large number of identical sequences between the two groups. For *Cyanothece* 7424, 4880 or roughly 85% of the ORFs had direct matches to a PEG. For those remaining unassociated PEGs, the amino acid sequences are blasted against the ORF labeled genome. Those results that have an E-value below 10^-30, zero mismatches, and a hit length that is at least 90% of both the PEG and ORF, are also determined to be matches.

Future expansions of this method will involve the inclusion of both the BIOCYC databases [47-49] as well as the recently published GLOBUS method [50]. BIOCYC contains what is referred to as tier 3 databases for all 5 species of *Cyanothece*. A tier 3 database is constructed using the PathoLogic program [51]. This program predicts the metabolic pathways present within a submitted genome. It predicts which transport reactions and enzymes are necessary to complete pathways, along with which genes may correspond to those enzymes [47-49]. Global Biochemical reconstruction Using Sampling (GLOBUS), is a novel method developed to attribute a function to a gene despite possible low sequence identity with the gene being used to ascribe its function. This is accomplished through a combination of sequence homology and context based correlations [50]. GLOBUS has the ability to provide unique functional annotations, given that it does not solely rely on homology searches, as many methods do. Yet given the comparative lack of testing done on this method, lower confidence will be placed on these results until more research has been done to prove the accuracy of this method.

## Evaluation and Processing of Annotations

At this stage in the Annotation Integration Method, annotations of varying specificity were retrieved for a number of ORFs. Some annotations were as complete as providing an EC number whereas some annotations provided only a protein or cluster name. For those annotations that did not specifically list an EC number, a number of relations are found using Uniprot. Those annotations that still could not be used to find an EC correlation are kept for use in the reaction confirmation portion of the GapFill process.

The GPR relationship is then completed with the addition of reactions from the file of balanced reactions from the SEED database using the determined EC numbers. These reactions are then filtered by a number of criteria to avoid inclusion of unrealistic reactions. First any reactions that contain metabolites confirmed not to be in the species are removed. Such is the case with metabolites such as ubiquinone and menaquinone, neither of which are present within cyanobacteria [52, 53]. Next, any reactions that contain generic metabolites are removed. These reconstructions deal with precise molecules, therefore any reactions containing metabolites such as a fatty acid with length R, are deemed unspecific and removed. If several reactions are still related to an EC number, they are reviewed to see if they vary only by a single metabolite or cofactor. If the varying metabolites are all determined to be within the organism, they are added. Further literature is consulted if there are several associated reactions with differing function in order to determine which actually corresponds to the EC number in question. All reactions are given one final manual review for plausibility before being accepted.

## Reconciliation of Annotations with Intermediate Model

These new GPR relations must be reconciled with those already present in the intermediate model. Every new relation that has an overlap with any part of a GPR from the intermediate must be manually inspected. If there are several reactions with different functions associated with the same genes or proteins, then precedence will be given to the GPRs from the intermediate model. Any conflicting GPRs will either be reevaluated if there is other annotation data, or else removed from the model entirely, as can be seen in the annotation retrieval workflow in Figure 3. If a new GPR is to be assigned to a gene it must pass the same restrictions as all other annotation assignments. These newly reconciled GPR are then added to the intermediate model, forming what is referred to as the draft model. The steps in the Annotation Integration Method that form a draft model from an intermediate model and annotations can be seen in Figure 3.

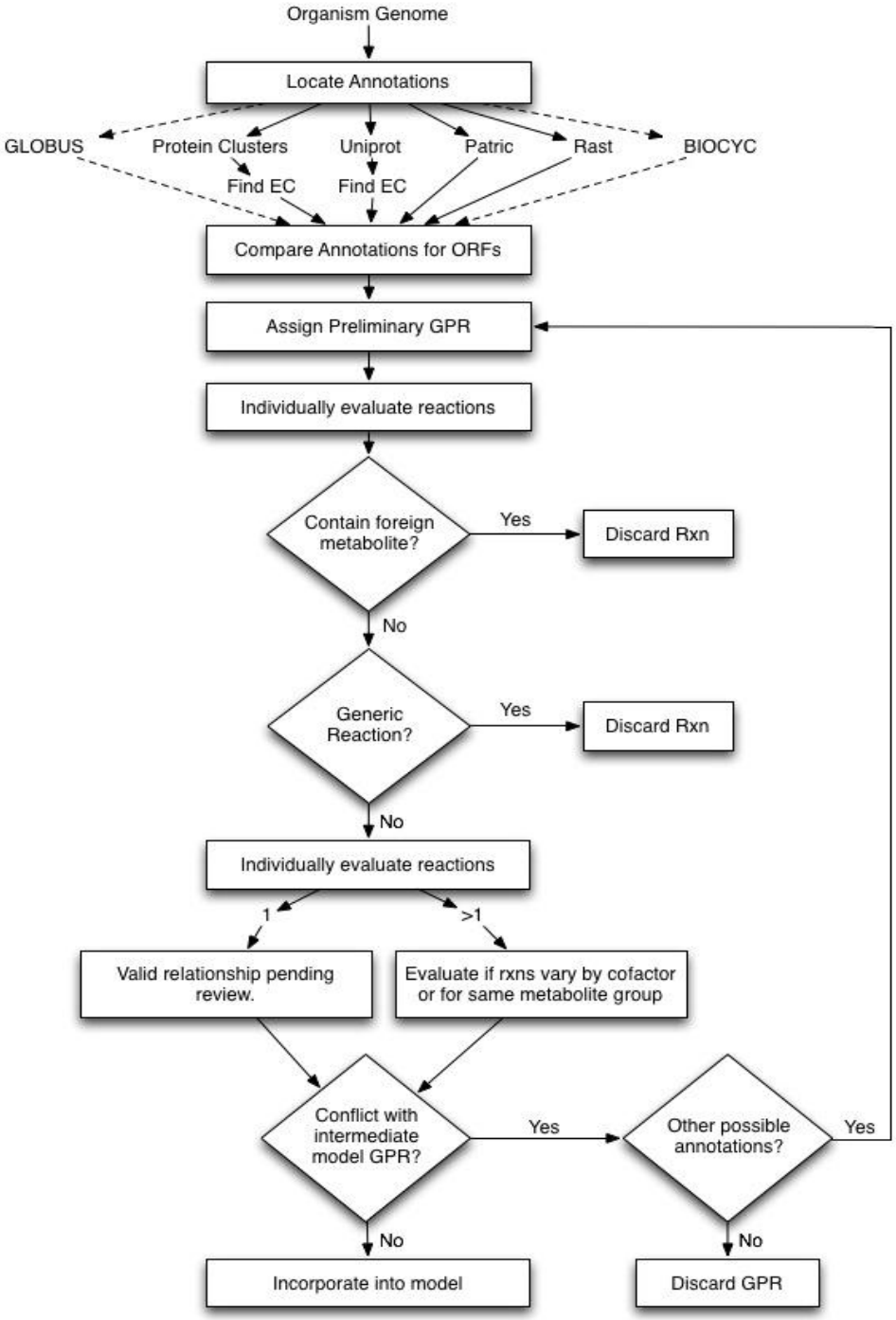



**Figure 3: Annotation retrieval workflow used in the Annotation Integration Method.** GLOBUS and BIOCYC annotation sources linked with dotted lines to delineate future inclusion in method.

**Annotation Retrieval and Model Simulation**

All query sequences for bidirectional homology searches were prepared from ORF delineated genomes from the PATRIC database. The homology comparison was automated through a PYTHON code developed by the author. The homology comparison was performed through use of the BLASTp algorithm. Complete annotation files were retrieved from the FTP servers from the respective websites. The files were processed and relevant annotations were retrieved through personally developed PYTHON codes. PYTHON will be used to code both the EC and reaction retrieval steps.

All flux balance analysis [15, 16], flux variability analysis, and GapFind and GapFill [41] optimization problems were solved using the CPLEX solver (version 12.1, IBM ILOG), which was used in the GAMS (version 23.3.3, GAMS Development Corporation) environment. Flux distributions were inferred using the FBA formulation as shown in equations 1-3 below.

$$Maximize\ V_{biomass} \quad 1)$$

$$\sum_{j=1}^{m} S_{ij} v_j = 0 \ \forall \ i \in 1, \ldots, n \quad 2)$$

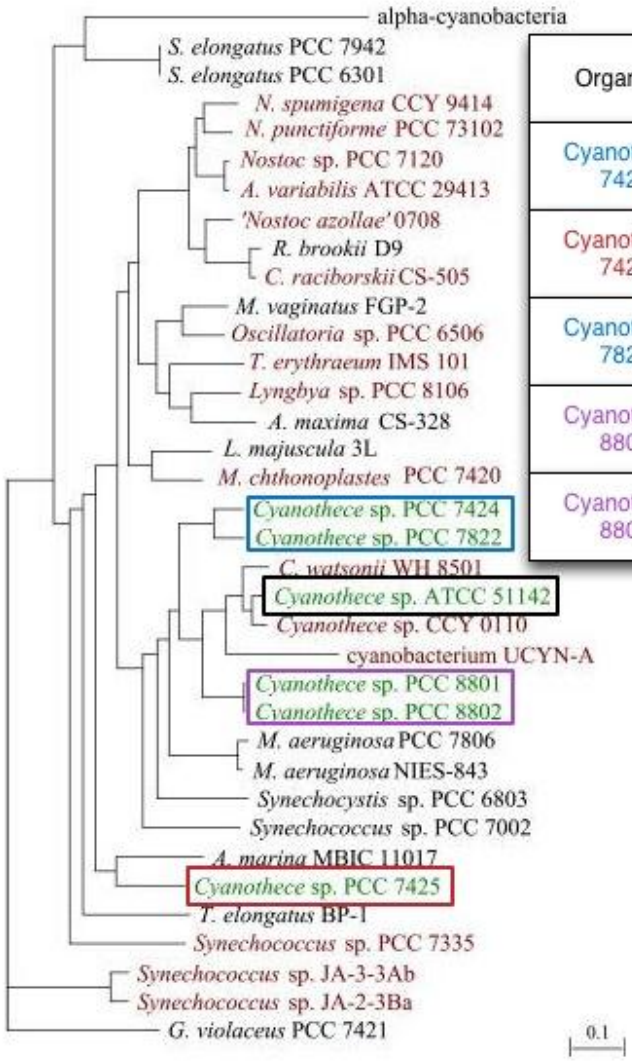
$$v_{j,min} \leq v_j \leq v_{j,max} \; \forall \, j \in 1,..,m \qquad 3)$$

In these equations j is the set of all reactions and i is the set of all metabolites. $V_j$ is the flux through reaction j, and $S_{ij}$ is the stoichiometric coefficient of metabolite i in reaction j. $V_{biomass}$ is the flux of the biomass equation, which when maximized simulates the actual organism growth.

## Chapter 3 Results and Discussion

The results from the annotation retrieval and evaluation of the intermediate models are indications of the efficacy of the Annotation Integration Method. Annotation reconciliation and final model curation are continuing and have been validated through phylogenetic based comparisons.

Comparisons can be made between the intermediate models and the *i*Cyt773 model, as well as between each other. Given the close phylogenetic relationship between *Cyanothece* species, a large percentage of the reactions in *i*Cyt773 were transferred during the intermediate model generation. As can be seen by the phylogenetic tree in Figure 4, the two species most closely related to 51142, 8801 and 8802, also share the highest number of reactions. Similarly the most removed species, 7425, shares the fewest reactions.

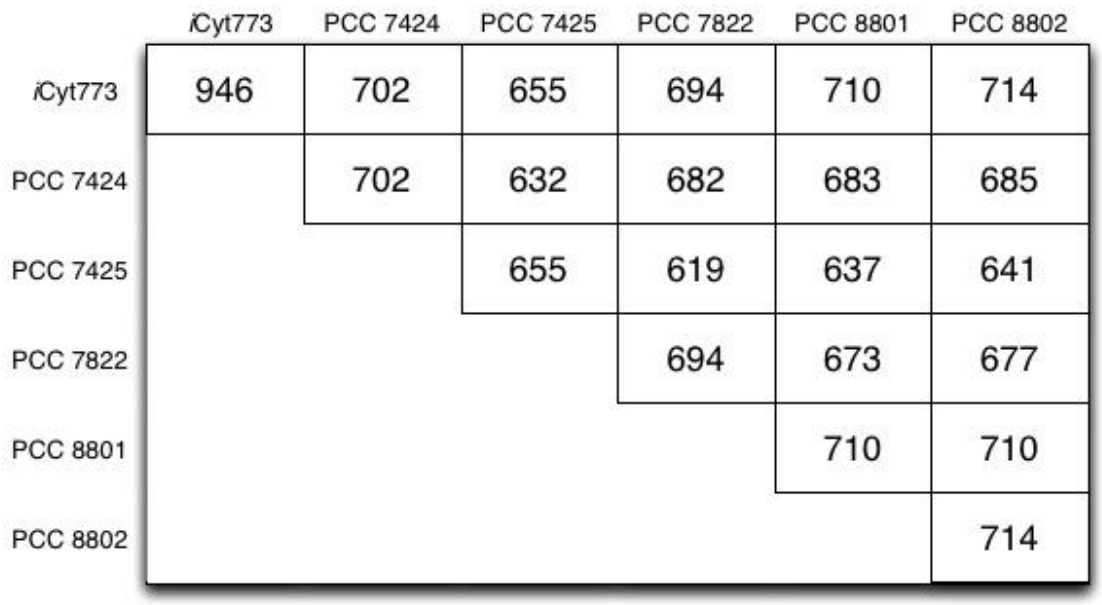



| Organism | Reactions in Intermediate Model | % of *i*Cyt773 model |
|---|---|---|
| Cyanothece 7424 | 702 | 74.21 |
| Cyanothece 7425 | 655 | 69.24 |
| Cyanothece 7822 | 694 | 73.36 |
| Cyanothece 8801 | 710 | 75.05 |
| Cyanothece 8802 | 714 | 75.48 |

**Figure 4: Comparison of number of reactions in intermediate models with positioning in phylogenetic tree.** The species are highlighted in the phylogenetic tree in the same color as they are written in the table. *Cyanothece ATCC* 51142 is highlighted in the black box. Phylogenetic tree published in Bandyopadhyay et al [31].

Pairwise comparisons performed between the intermediate models, shown in Figure 5, display a continued adherence to trends exhibited by the phylogeny. The number of shared reactions directly correlates to the separation between organisms on the phylogenetic tree.

7425, being the farthest removed on the phylogenetic tree, shares the fewest reactions with the other species. As anticipated, the two pairs most closely linked on the tree, 7424/7822 and 8801/8802, share a larger percentage of their reactions with each other than the other intermediate models.



| | *i*Cyt773 | PCC 7424 | PCC 7425 | PCC 7822 | PCC 8801 | PCC 8802 |
|---|---|---|---|---|---|---|
| *i*Cyt773 | 946 | 702 | 655 | 694 | 710 | 714 |
| PCC 7424 | | 702 | 632 | 682 | 683 | 685 |
| PCC 7425 | | | 655 | 619 | 637 | 641 |
| PCC 7822 | | | | 694 | 673 | 677 |
| PCC 8801 | | | | | 710 | 710 |
| PCC 8802 | | | | | | 714 |

**Figure 5: Pairwise comparison of intermediate models and *i*Cyt773.** Number of shared reactions between the five developed intermediate models and the *i*Cyt773 model.

The Annotation Integration Method focuses on both standardizing and streamlining the creation of metabolic models for organisms with few reviewed annotations. In addition, through the use of a number of reliable databases and models, the Annotation Integration Method serves to mitigate many errors that can arise from the practices of using a single source or un-reviewed annotations that are more commonly requisite with less researched organisms.

# Chapter 4 Future Steps

For initial testing the biomass equation from iCyt773 was used. For final development individual biomass equations are being developed using information from Dr. Pakrasi's group at the University of Washington St. Louis. Biomass compositions will be accurately measured in a method similar to that discussed by Saha et al [39]. Stoichiometric values will be calculated for all major biomass components, including DNA, RNA, carbohydrates, amino acids, lipids, and pigments.

Upon inclusion of the novel biomass equations, the GapFind and GapFill algorithms will be implemented [41]. These algorithms identify and propose network modifications to reconnect root no production and root no consumption metabolites. Reactions from previous stages in the Annotation Integration Method that were restricted from the model, namely reactions from all bidirectional hit logic gate analysis, and conflicting annotations, will be given higher priority for gap filling. Additional consideration is taken to avoid the creation of thermodynamically infeasible cycles through reaction addition or modification of reaction reversibility. The formation of thermodynamically infeasible cycles was in large part mitigated through taking a large number of reactions from the cycle-free iCyt773 model. Flux variability analysis will be performed to verify the lack of cycles. Any reactions whose fluxes hit the bounds are determined to be involved in a cycle. Cycle resolution commonly involves restriction of the reversibility of a reaction or in rare instances the complete removal of certain reactions. After cycle resolution, certain reactions, such as the nitrogenase in *Cyanothece* 7425, will also be restricted to the specific conditions under which they are active.

There are a number of methods for additional verification of the metabolic reconstructions, three of which are most promising. The first of which involves comparing metabolic functions between the reconstructions and publications detailing the activities of specific pathways or enzymes. Second, if any known gene knockout or mutant data is present, then model testing similar to the testing performed by the author of this paper for the publication by Saha et al [39], could be implemented. One mutant has already been reported for *Cyanothece* 7822 [54]. Finally the use of flux data through metabolic flux analysis [39, 55] can be used to verify the flux distributions predicted through FBA.

In future refinements, the potential exists to pair the Annotation Integration Method with METRXN0 [56] to provide an accessible and streamlined platform for the development of metabolic reconstructions.

## Chapter 5 Conclusion

This paper outlines the Annotation Integration Method, a new approach to developing metabolic reconstructions for organisms with relatively few reviewed gene annotations. Models are currently being developed for 5 different strains of *Cyanothece* both as verification of the efficacy of the method, as well as to help determine the differences and promising features of their metabolisms. The developed intermediate models are following expected phylogenetic trends, and the number of retrieved annotations for *Cyanothece* PCC 77424 is promising.

## References


1. Suthers, P.F., A. Zomorrodi, and C.D. Maranas, *Genome-scale gene/reaction essentiality and synthetic lethality analysis.* Mol Syst Biol, 2009. **5**: p. 301.
2. Cuauhtemoc Licona-Cassani, et al., *Reconstruction of the Saccharopolyspora erythraea genome-scale model and its use for enhancing erythromycin production.* Antonie van Leeuwenhoek, 2012. **102**(3): p. 493-502.
3. Zomorrodi, A.R., et al., *Mathematical optimization applications in metabolic networks.* Metab Eng, 2012. **14**(6): p. 672-86.
4. Sridhar Ranganathan, Patrick F. Suthers, and C.D. Maranas, *OptForce: An Optimization Procedure for Identifying All Genetic Manipulations Leading to Targeted Overproductions.* PLOS Computational Biology, 2010.
5. Ranganathan, S. and C.D. Maranas, *Microbial 1-butanol production: Identification of non-native production routes and in silico engineering interventions.* Biotechnol J, 2010. **5**(7): p. 716-25.
6. Sónia Carneiro, Isabel Rocha, and E. Ferreira, *Application of a genome-scale metabolic model to the inference of nutritional requirements and metabolic bottlenecks during recombinant protein production in Escherichia coli.* Microbial Cell Factories, 2006. **5**.
7. Adam M. Feist, et al., *Reconstruction of Biochemical Networks in Microbial Organisms.* Nat. Rev Microbiol, 2009. **7**(2): p. 129-143.
8. Edwards, J.S. and B.O. Palsson, *The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities.* Proc Natl Acad Sci U S A, 2000. **97**(10): p. 5528-33.
9. Puchalka, J., et al., *Genome-scale reconstruction and analysis of the Pseudomonas putida KT2440 metabolic network facilitates applications in biotechnology.* PLoS Comput Biol, 2008. **4**(10): p. e1000210.
10. Suthers, P.F., et al., *A genome-scale metabolic reconstruction of Mycoplasma genitalium, iPS189.* PLoS Comput Biol, 2009. **5**(2): p. e1000285.
11. Thiele, I. and B.O. Palsson, *A protocol for generating a high-quality genome-scale metabolic reconstruction.* Nat Protoc, 2010. **5**(1): p. 93-121.
12. Reed, J.L., et al., *An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR).* Genome Biol, 2003. **4**(9): p. R54.
13. Adam M Feist, et al., *A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information.* Molecular Systems Biology, 2007. **3**(121).
14. Zomorrodi, A. and C. Maranas, *Improving the iMM904 S. cerevisiae metabolic model using essentiality and synthetic lethality data.* BMC Systems Biology, 2010. **4**(178).

15. Jeffrey D. Orth, Ines Thiele, and B.Ø. Palsson, *What is flux balance analysis.* Nat. Biotechnology, 2011. **28**(3): p. 245-248.
16. Varma, A. and B. Palsson, *Metabolic Flux Balancing - Basic Concepts, Scientific and Practical Use.* Bio-Technology, 1994: p. 994-998.
17. Karp, P.D., et al., *Multidimensional annotation of the Escherichia coli K-12 genome.* Nucleic Acids Res, 2007. **35**(22): p. 7577-90.
18. Christie, K.R., et al., *Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms.* Nucleic Acids Res, 2004. **32**(Database issue): p. D311-4.
19. Guldener, U., et al., *CYGD: the Comprehensive Yeast Genome Database.* Nucleic Acids Res, 2005. **33**(Database issue): p. D364-8.
20. Consortium, T.U., *Reorganizing the protein space at the Universal Protein Resource (UniProt).* Nucleic Acids Research, 2012.
21. Gillespie, J.J., et al., *PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species.* Infect Immun, 2011. **79**(11): p. 4286-98.
22. Kanehisa, M., et al., *KEGG for integration and interpretation of large-scale molecular datasets.* Nucleic Acids Research, 2012.
23. Kanehisa, M. and S. Goto, *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Research, 2000. **28**: p. 27-30.
24. Dal'Molin, C.G., et al., *AlgaGEM--a genome-scale metabolic reconstruction of algae based on the Chlamydomonas reinhardtii genome.* BMC Genomics, 2011. **12 Suppl 4**: p. S5.
25. de Oliveira Dal'Molin, C.G., et al., *AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis.* Plant Physiol, 2010. **152**(2): p. 579-89.
26. Balagurunathan, B., et al., *Reconstruction and analysis of a genome-scale metabolic model for Scheffersomyces stipitis.* Microb Cell Fact, 2012. **11**: p. 27.
27. Ulas, T., et al., *Genome-scale reconstruction and analysis of the metabolic network in the hyperthermophilic archaeon Sulfolobus solfataricus.* PLoS One, 2012. **7**(8): p. e43401.
28. Tamagnini, P., et al., *Hydrogenases and hydrogen metabolism of cyanobacteria.* Microbiol Mol Biol Rev, 2002. **66**(1): p. 1-20, table of contents.
29. Dismukes, G.C., et al., *Aquatic phototrophs: efficient alternatives to land-based crops for biofuels.* Curr Opin Biotechnol, 2008. **19**(3): p. 235-40.
30. Ducat, D.C., J.C. Way, and P.A. Silver, *Engineering cyanobacteria to generate high-value products.* Trends Biotechnol, 2011. **29**(2): p. 95-103.
31. Anindita Bandyopadhyay, et al., *Novel Metabolic Attributes of the Genus Cyanothece, Comprising a Group of Unicellular Nitrogen-Fixing Cyanobacteria.* mBio, 2011. **2**(5).
32. Min, H. and L.A. Sherman, *Hydrogen production by the unicellular, diazotrophic cyanobacterium Cyanothece sp. strain ATCC 51142 under*

*conditions of continuous light.* Appl Environ Microbiol, 2010. **76**(13): p. 4293-301.
33. Melnicki, M.R., et al., *Sustained H(2) production driven by photosynthetic water splitting in a unicellular cyanobacterium.* mBio, 2012. **3**(4): p. e00197-12.
34. David O. Hall, et al., *The potential applications of cyanobacterial photosynthesis for clean technologies.* Photosynthesis Research, 1995. **46**: p. 159-167.
35. Welsh, E.A., et al., *The genome of Cyanothece 51142, a unicellular diazotrophic cyanobacterium important in the marine nitrogen cycle.* Proc Natl Acad Sci U S A, 2008. **105**(39): p. 15094-9.
36. Porta, D., R. Rippka, and M. Hernandez-Marine, *Unusual ultrastructural features in three strains of Cyanothece (cyanobacteria).* Arch Microbiol, 2000. **173**(2): p. 154-63.
37. Wu, B., et al., *Alternative isoleucine synthesis pathway in cyanobacterial species.* Microbiology, 2010. **156**(Pt 2): p. 596-602.
38. Hai T, Hein S, and S. A, *Multiple evidence for widespread and general occurrence of type-III PHA synthases in cyanobacteria and molecular characterization of the PHA synthases from two thermophilic cyanobacteria:Chlorogloeopsis fritschii PCC 6912 andSynechococcus sp. strain MA19.* Microbiology 2001. **147**(11): p. 3047-60.
39. Saha, R., et al., *Reconstruction and Comparison of the Metabolic Potential of Cyanobacteria Cyanothece sp. ATCC 51142 and Synechocystis sp. PCC 6803.* PLoS One, 2012. **7**(10): p. e48285.
40. Trang T. Vu, et al., *Genomescale modeling of light-driven reductant partitioning and carbon fluxes in diazotrophic unicellular cyanobacterium Cyanothece sp. ATCC 51142.* PLOS Computational Biology, 2012. **8**(4).
41. Satish Kumar, V., M.S. Dasika, and C.D. Maranas, *Optimization based automated curation of metabolic reconstructions.* BMC Bioinformatics, 2007. **8**: p. 212.
42. Price, N.D., et al., *Extreme pathways and Kirchhoff's second law.* Biophys J, 2002. **83**(5): p. 2879-82.
43. Schellenberger, J., N.E. Lewis, and B.O. Palsson, *Elimination of thermodynamically infeasible loops in steady-state metabolic models.* Biophys J, 2011. **100**(3): p. 544-53.
44. William Klimke, et al., *The National Center for Biotechnology Information's Protein Clusters Database.* Nucleic Acids Research, 2009.
45. Ramy K Aziz, et al., *The RAST Server: Rapid Annotations using Subsystems Technology.* BMC Genomics, 2008.
46. Meyer F, Overbeek R, and R. A., *FIGfams: yet another set of protein families.* Nucleic Acids Research, 2009. **37**(20): p. 6643-54.
47. P.R. Romero and P.D. Karp, *Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases.* Bioinformatics, 2004. **20**(5): p. 709-717.

48. Green, M.L. and P.D. Karp, *A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases.* BMC Bioinformatics, 2004. **5**: p. 76.
49. Lee, T.J., I. Paulsen, and P. Karp, *Annotation-based inference of transporter function.* Bioinformatics, 2008. **24**(13): p. i259-67.
50. Plata, G., et al., *Global probabilistic annotation of metabolic networks enables enzyme discovery.* Nat Chem Biol, 2012.
51. Peter D Karp and S.M. Paley, *Evaluations of Computational Metabolic-Pathway Predictions for Helicobacter pylori.* Bioinformatics, 2003. **18**(5): p. 705-14.
52. Collins, M.D. and D. Jones, *Distribution of Isoprenoid Quinone Structural Types in Bacteria and Their Taxonomic Implications.* Microbiological Reviews, 1981. **45**(2): p. 316-354.
53. Sadre, R., C. Pfaff, and S. Buchkremer, *Plastoquinone-9 biosynthesis in cyanobacteria differs from that in plants and involves a novel 4-hydroxybenzoate solanesyltransferase.* Biochem, 2012. **442**: p. 621-629.
54. Min, H. and L.A. Sherman, *Genetic transformation and mutagenesis via single-stranded DNA in the unicellular, diazotrophic cyanobacteria of the genus Cyanothece.* Appl Environ Microbiol, 2010. **76**(22): p. 7641-5.
55. Ranganathan, S., et al., *An integrated computational and experimental study for overproducing fatty acids in Escherichia coli.* Metab Eng, 2012.
56. Kumar, A., P.F. Suthers, and C.D. Maranas, *MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases.* BMC Bioinformatics, 2012. **13**: p. 6.

# ACADEMIC VITA

Thomas Mueller

Tjm5293@psu.edu

---

## Education

B.S., Chemical Engineering, Fall 2012, Pennsylvania State University, University Park, Pa

## Honors and Awards

Awarded Hollenbach scholarship for four years (2009-2012) from College of Engineering
Received Penn State Alumni Association Scholarship (2009)
Schreyer Summer Research Grant (2010)
Summer Research grant from the Chemical Engineering Department (2010)
Schreyer Scholar
Received Lee and Mary Eagleton Award for Excellence in Design (2012)
Selected to participate in Chemical Energy storage and conversion REU (2011)

## Association Memberships/Activities

Member of AICHE

## Research Experience

Worked in the Chemical and Biological Systems Optimization lab from August 2009-present. Performed research in areas such as gene annotation, metabolic engineering, and optimization programming.

## Research Interests

I have broad interests in optimization programming and metabolic engineering, with a focus on genome scale metabolic reconstructions.

**Professional Presentations**

Presented poster on work at DOE Genomics Science Conference (2012)

**Publications and Papers**

Author on: Saha, R., et al., *Reconstruction and Comparison of the Metabolic Potential of Cyanobacteria Cyanothece sp. ATCC 51142 and Synechocystis sp. PCC 6803.* PLoS One, 2012. **7**(10): p. e48285.