THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF CHEMICAL ENGINEERING

A COMPUTATIONAL COMPARISON OF GENOME-SCALE METABOLIC MODELS HIGHLIGHTING THE NEED FOR THE ADOPTION OF UNIVERSAL CONVENTIONS AND STANDARDS

STEPHEN THOMAS SPAGNOL
Spring 2010

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

Themis Matsoukas
Professor of Chemical Engineering
Honors Adviser

Andrew Zydney
Walter L. Robb Chair & Professor of Chemical
Engineering Chemical Engineering
Department Head

* Signatures are on file in the Schreyer Honors College.

Abstract

This study aims to compare some of the existing genome-scale metabolic models and display the inconsistencies between them in order to highlight the need for the adoption of universal conventions and standards of completeness and coverage to enhance the versatility and applicability of these models to metabolic engineering applications. These applications include uses for industry (i.e. biofuels, commodity chemicals, biochemicals, etc.), medicine (i.e. drug production, vaccines, antibiotics, drug target identification, etc.), bioremediation, and the growing number of problems to which metabolic engineering is being applied. However, the potential usage of genome-scale metabolic models for these applications is limited by the lack of congruency between models, which hinders attempts at strain optimization, gap filling, production of new metabolic reconstructions, and insertion of foreign pathways into a new host. These discrepancies primarily include incomplete reaction data, such as elementally and charge unbalanced reactions, and a lack of universal metabolite specificity and naming conventions. In this study, a Metabolite Rosetta Stone was created to allow for the translation of the different metabolite abbreviations from each model to a common form for comparison of their metabolic networks. In comparing 34 genome-scale metabolic models and the *Escherichia coli* core model, only three reactions were found to be common among all 35 reaction networks, which contradicts the fact that many of these organisms share several conserved metabolic pathways and, thus, reactions. However, in a comparison of seven models of more uniform conventions, a better agreement was observed with 40 reactions found common to all of them. This result conveys the need for the adoption of uniform conventions and standards and a reconciliation of these previous models in order to compare existing models, develop new ones from them, and incorporate existing pathways from one model into another. In addition, this study will detail what needs to be done to rectify the current problems and also provide some potential solutions to enhance the capabilities and effectiveness of genome-scale metabolic models.

Table of Contents

Acknowledgements

I would like to acknowledge Professor Costas D. Maranas, my Schreyer Honors Thesis and Research Adviser, for his guidance throughout the whole process of the researching and writing of my Schreyer Honors Thesis and, more importantly, his guidance during the course of my undergraduate career. My work with him has been the most inspiring and rewarding experience of my undergraduate career and has helped set me on my career path and provided me with a role model to be looked up to. In addition, Professor Andrew Zydney has played a critical role in helping me to formulate my career path and making sure I follow what I am passionate about. I would also like to recognize and thank Patrick Suthers for his contribution to my undergraduate research, as he assisted in formulating my project and providing me with all of the necessary help along the way. In addition, I would like to thank the Penn State Department of Chemical Engineering for their support by providing me with the Summer Research Fellowships in Biomolecular Engineering to do my research. Finally, I would like to acknowledge and thank Professor Themis Matsoukas, my Honors Adviser, for advising me throughout the course of my undergraduate studies in the Chemical Engineering Department at Penn State.

Introduction and Background

Metabolic engineering is the controlled manipulation of an organism's metabolic pathways, including its enzymes, regulatory networks, and transport processes, to achieve either the anabolism of a certain product, the catabolism of a metabolite present in the medium, or both. The field arose from the availability of complete genome sequencing and the development of new techniques for gene deletion, insertion, and controlled expression; namely, recombinant DNA technology [1]. However, metabolic engineering of organisms is inherently unique from other methods of cellular engineering in that it requires a systems-based approach due to the complexity of the metabolic networks and the interdependence of different pathways [1]. Whereas expression of a foreign protein requires insertion of the associated gene into the host genome under a promoter and simple expression using the host's transcriptional and translational machinery, metabolic engineering requires careful analysis of the entire system in order to determine the gene manipulations (deletions, insertions, and expression levels) needed to generate the optimal production of a certain metabolite. Consequently, using the same approach as in foreign protein expression with a focus on single genes (and their dependent enzymes) has been shown to be generally ineffective at increasing the production of a target metabolite [2]. This seemingly paradoxical response is caused by the evolution of organisms to direct their metabolism towards pathways that are critical for growth and development [1]. Thus, directing an organism's resources to pathways that are contrary to this evolution cannot be achieved by targeting single genes, as the organism will develop ways to work around such an obstacle and ensure that its resources are properly

devoted to maintaining its fitness. As a result, from the onset of research into metabolic engineering, it has been apparent that exhaustive experimentation must be done in order to obtain the right combination of gene manipulations for the optimal flux towards the production of a desired compound or growth on particular medium. To alleviate the number of experiments, computational algorithms and software were developed early on to predict the metabolic capabilities of an organism and to identify gene targets for manipulation [3, 4]. Examples of these methods include, but are not limited to, the COBRA toolbox [3], constraint-based flux analysis/flux balanced analysis [4, 5], Optknock [6], OptStrain [7], OptReg [8], Flux Coupling Finder [9], and the genome-scale metabolic models themselves [10, 11]. These methods, and others like them, have provided the systems-based approach needed for determining the genes to be manipulated for the optimal flux towards a target metabolic pathway or growth on a particular medium.

Metabolic engineering has found extensive use in many industries including the pharmaceutical industry, biotech industry, commodity chemical industry, biofuel industry, and others [12, 13]. The widening number of applications for metabolic engineering has necessitated further research into computational methods for predicting the metabolic behavior of an organism, identifying gene targets for manipulation, and determining the optimal growth media for a specific task [11]. To this end, genome-scale metabolic models have proven to be the most robust tools capable of these complex tasks [11].

The use of genome-scale metabolic models, by limiting the number of costly and time-consuming experiments, will help the field of metabolic engineering advance into new

fields and further optimize the use of microorganisms in the fields in which they are already utilized. New applications for metabolic engineering are already being researched through genome-scale metabolic models. Current industrial applications include the production of biofuels such as ethanol and butanol [1, 4, 11, 13], biodiesel [11], commodity chemicals (i.e. propane diol, acetone, and acetaldehyde) [1, 11, 13], biochemicals (i.e. lactic acid) [1, 11, 13], and biopolymers [13]. In addition to these, an increasing number of resources are being devoted to research into medical applications such as the production of therapeutic proteins and hormones [1], amino acid biosynthesis [11], synthetic drugs and drug intermediates [1], nutrients and dietary supplements (i.e. lycopene, L-lysine, etc.) [13], and pharmaceuticals (including antibiotics and vaccines) [13]. Recent developments in the field include use in bioremediation to remove toxic waste while producing useful chemicals, as genome-scale metabolic models can be utilized for the determination of metabolic capabilities for growth on a specific medium [13]. In addition, genome-scale metabolic models have been utilized to identify potential drug targets for pathogens such as the multi-drug resistant Acinetobacter baumannii AYE [14] as well as respiratory, gastrointestinal, and oral pathogens [13] since, by identifying the essential genes, reactions, and metabolites, drugs can be developed to combat these pathogens [13]. Based on these same ideas, genome-scale metabolic models are being utilized to model the characteristics of many diseased human cells and specific tissue cells to enhance our understanding and research potential solutions [13]. Finally, another interesting fuel and energy application that has emerged is the use of photosynthetic microorganisms to harness the energy of the sun and produce new energy resources such as biofuels and biohydrogen [13], which is being explored with genome-scale metabolic

models. All of these applications expose the need to reduce the costly experiments and expedite the process through the use of genome-scale metabolic models and other computational approaches.

Currently, there are two primary obstacles facing the use of genome-scale metabolic models. The first type includes deviations of the models from experimental results and "gaps" within the models themselves (i.e. isolated metabolites without connecting pathways or missing reactions within pathways [15]). The second includes the lack of universal conventions for producing genome-scale metabolic models and the production of wholly inadequate models, which lack the quality and coverage necessary to be utilized in conjunction with other models. While much research has been devoted to analyzing the complications of the former [16, 17, 18, 19], little has been done in the way of the latter. Yet, the lack of universal conventions and the production of incomplete models hinder the utility of the models and their applications to metabolic engineering. Problems commonly associated with incomplete models include stoichiometrically inconsistent reactions, reactions that are elementally or charge unbalanced, and differing metabolite naming conventions which can lead to duplication of a metabolite in different models. Whereas uniform and complete models would allow for easy comparison of similar pathways between organisms in order to determine which has the optimal flux towards a particular metabolite or to incorporate pathways from other organisms into new hosts using the gene-protein-reaction (GPR) associations and gene manipulation, the current state of genome-scale metabolic models causes these novel applications to fall short of their full potential. Other potential advantages would include the production of phylogenic trees based on metabolic capabilities, the production of new genome-scale

metabolic models from the pathways of old ones, strain optimization, and even gap-filling.

The lack of uniform conventions for metabolite naming and specificity is particularly enigmatic because the probability of the duplication of a given metabolite in a genome-scale metabolic model or database is highly likely. The presence of these duplicates results in the increased likelihood of "gaps" in the metabolic networks and further limits the ability of these models or databases to aid in the determination of target genes for manipulation and evaluation of the metabolic capabilities of an organism. Thus, there is an inherent need to reconcile the different metabolite abbreviations in order to avoid these shortcomings in the genome-scale metabolic models.

Recently, an attempt has been made to set forth a "comprehensive protocol" for the production of future genome-scale metabolic models to ensure that subsequent models follow uniform conventions with standards for completeness and quality [20]. While those working in the field of genome-scale metabolic models begin to adopt these common conventions for future models, steps must be taken to reconcile the existing models in order to adapt them for future uses. The objective of this work is to probe a number of the existing models and highlight the incompatibilities between models to emphasize the need for universal conventions and standards. In addition, this work aims to detail the necessary measures that must be taken to rectify these incompatibilities and recommend potential developments that may improve the efficacy of genome-scale metabolic models.

Materials and Methods

Analysis of the Available Genome-scale Metabolic Models

Prior to beginning the comparisons between the genome-scale metabolic models, each of the models available were probed to determine the different variations in the representations of reactions between models. Since several metabolic pathways have been conserved throughout all life forms (such as glycolysis, pentose-phosphate pathway, amino acid biosynthesis, etc.), reactions from these pathways were used to probe the models and observe these differences. From this process, it was determined that there were at least nine variations of the same reaction among the different models even when differing metabolite abbreviations were neglected (as differences in reaction reversibility, charge/elemental balancing, and cofactor usage were the primary means of comparison for this analysis). The nine major differences are shown for the glucose-6-phosphate dehydrogenase reaction in Table 1.

Table 1. The glucose-6-phosphate deyhdrogenase reaction of the glcolysis pathway in selected genome-scale metabolic models to compare the variation in its representation. In comparing these models, the inconsistency between models became apparent as even highly conserved reactions varied in completeness and coverage. The varying cofactor usage, charge and elemental balancing (or lack thereof), and directionality were all found to be inconsistent throughout.

Reaction Listing	Model
[c] : g6p + nadp <==> 6pgl + h + nadph	*Escherichia coli* iAF1260 [21], *Lactobacillus plantarum* [22], *Pseudomonas aeruginosa* [23], *Staphylococcus aureus* [24]
[c] : g6p + nadp --> 6pgl + h + nadph	*Bacillus subtilis* [25], *Mycobacterium tuberculosis* iNJ661 [26], *Pseudomonas putida* [27], *Rhizobium etli* [28], *Saccharomyces cerevisiae* iND750 [29], *Saccharomyces cerevisiae* [30]
[c]g6p + nadp <==> 6pgl + h + nadph	*Escherichia coli* iJR904 [31]
[c] : f420-2 + g6p --> 6pgl + f420-2h2	*Methanosarcina barkeri* [32]
G6P + NADP <-> D6PGL + NADPH	*Escherichia coli* [33]; *Mus musculus* [34], *Saccharomyces cerevisiae* iLL672 [35], *Saccharomyces cerevisiae* iFF708 [36]
G6P + NADP -> D6PGL + NADPH	*Aspergillus nidulans* [37], *Mannheimia succiniciproducens* [38], *Streptomyces coelicolor* [39]
G6P + NAD -> D6PGL + NADH	*Helicobacter pylori* [40]
C01172 + C00006 = C01236 + C00005 + C00080	*Mus musculus* (GSM mouse) [41]
C00092 + C00006 <=> C01236 + C00005 + C00080	*Halobacterium salinarum* [42]

Upon comparing the various models, a list of models with complete metabolite listings (for translating the metabolite abbreviations) and reaction data (for pathway comparison) was developed for further analysis. The Metabolite Rosetta Stone includes the metabolite abbreviations of 38 models, as well as the *E. coli* core model [43]. Since many of these models had their reactions in file types that could not easily be extracted for parsing, only 34 models and the *E. coli* core model were included for metabolic network comparisons. The two *Helicobacter pylori* models had the identical metabolite abbreviation conventions. Thus, even though the older model [40] did not include a listing of its metabolite abbreviations, its reactions were included for the comparison using the metabolite abbreviations of the newer model [44]. These 35 metabolic networks spanned all three kingdoms and are displayed in Table 2.

Table 2. The complete list of genome-scale metabolic models and the *E. coli* core model utilized to compare the different metabolic pathways and variation between models. The list includes organisms from all three kingdoms of life. The final list was developed based on the availability of both the metabolite and reaction lists in a suitable format.

Bacteria	*Acinetobacter baylyi* [45], *Bacillus subtilis* [25], *Corynebacterium glutamicum* [46], *Escherichia coli* iAF1260 [21], *Escherichia coli* iJR904 [31], *Escherichia coli* core model [43], *Escherichia coli* iJE660 (revised) [33], *Escherichia coli* iJE660 [33], *Geobacter metallireducens* [47], *Geobacter sulfurreducens* [48], *Helicobacter pylori* [40], *Helicobacter pylori* iIT341 [44], *Lactobacillus plantarum* [22], *Mannheimia succiniciproducens* [38], *Mycobacterium tuberculosis* GSMN-TB [49], *Mycobacterium tuberculosis* iNJ661 [26], *Mycoplasma genitalium* iPS189 [50], *Porphyromonas gingivalis* [51], *Pseudomonas aeruginosa* [23], *Pseudomonas putida* [27], *Rhizobium etli* [28], *Salmonella typhimurium* [52], *Staphylococcus aureus* [24], *Streptomyces coelicolor* [39].
Eukarya	*Aspergillus nidulans* [37], *Leishmania major* [53], *Mus musculus* [34], *Mus musculus* GSM mouse [41], *Mus musculus* Cardiomyocyte [54], *Saccharomyces cerevisiae* iFF708 [36], *Saccharomyces cerevisiae* iND750 [29], *Saccharomyces cerevisiae* iLL672 [35], *Saccharomyces cerevisiae* [30].
Archaea	*Halobacterium salinarum* [42], *Methanosarcina barkeri* [32].

Experimental Procedure

To compare the genome-scale metabolic models, three steps were needed to allow them to be assembled into a common form for analysis and comparison. The first of these steps was to unify their metabolite names since many groups that have developed a model have utilized their own unique naming convention. Second, the reactions needed to be assembled into a common form, including reaction directionality and compartment designation conventions, in order to be parsed and compared. Finally, where possible, the GPR associations needed to be coupled to their respective reactions for future analysis.

With many different metabolite naming conventions utilized among the genome-scale metabolic models, it is impossible to compare the reactions without first reconciling the different nomenclature. In addition, the lack of consistency in model metabolite abbreviations results in the potential presence of duplicated metabolites in models or databases because the differing nomenclature causes them not to be recognized as the same. To accomplish this first objective, a Metabolite Rosetta Stone table of full metabolite names and each model's abbreviation was manually created using the *Escherichia coli* iAF1260 model [21] as the standard template to be augmented for the purposes of translating the different abbreviations. Each of the models with a metabolite list had their respective metabolite abbreviations inserted into this Metabolite Rosetta Stone in order to match each model's abbreviation with the full metabolite name and other models' abbreviations for translation. The rows of each column provided the full name for a given metabolite and the next six columns were devoted to additional information (if available) about that compound, including KEGG ID, molecular formula,

and synonyms. Each of the remaining columns listed the metabolite abbreviation from a given model for the specific metabolite of that row. Thus, much like the true Rosetta Stone, this table could be utilized to translate the abbreviations of one model into those of another or into the full name of the compound. New metabolites were appended to the list as needed for cases in which previous models did not contain that specific metabolite. For this comparison, the different metabolite abbreviations were translated into a uniform style of the form CM00001 (with subsequent numbering up to CM04636) to be utilized for analyzing the reactions. For purposes of illustration, a sample of the Metabolite Rosetta Stone is included in Table 3 (some columns were eliminated for simplicity).

Table 3. A sample from the manually created Metabolite Rosetta Stone. Columns for the CAS number and the alternate molecular formula were not included since neither of these metabolites had them.

Full Name (Taken from E. coli iAF1260 model)	Maranas Group Database Abbreviation	Metabolite Molecular Formula (Taken from E. coli iAF1260 model)	Synonyms	KEGG ID	Escherichia coli Abbreviation (iAF1260) [21]	Acinetobacter baylyi Abbreviation [45]
10-Formyltetrahydrofolate	CM00001	C20H21N7O7	10-Formyl-THF	C00234	10fthf	10-FORMYL-THF
1,2-Diacyl-sn-glycerol (didodecanoyl, n-C12:0)	CM00002	C27H52O5	1,2-Diacylglycerol; D-1,2-Diacylglycerol	C00641	12dgr120	DIACYLGLYCEROL

After manually creating the Metabolite Rosetta Stone to translate the different abbreviations into a common nomenclature, the reactions of each model needed to be transformed into a similar form to be parsed and compared. This included the conversion of the different symbols for reaction directionality and the various ways of denoting the compartment in which the reaction takes place to a unified format. The format adopted was the same as the *E. coli* iAF1260 model. For genome-scale metabolic models of eukaryotes, which have a vast number of compartments, the compartment designation conventions of the *Saccharomyces cerevisiae* model iND750 were utilized [29] and appended. Table 4 lists how the compartments were denoted for comparison between models.

Table 4. A complete list of the compartment designation notations. These notations were adopted from the *E. coli* iAF1260 model and appended with the *S. cerevisiae* iND750 model for eukaryote department designation.

Extracellular	[e]
Cytoplasm	[c]
Periplasm	[p]
Chloroplast	[h]
Mitochondria	[m]
Endoplasmic Reticulum	[r]
Golgi Apparatus	[g]
Lysosome	[l]
Peroxisome/Glyoxysome	[x]
Glycosome	[y]
Vacuole	[v]
Acidocalcisome	[a]
Nucleus	[n]

For purposes of identifying common reactions between models, the reaction stoichiometry and the compartment in which it took place were eliminated for comparison. In addition, reaction directionality was ignored such that reactions were recognized as the same whether or not they where reversible or irreversible (thus, a reaction of the form A <==> B was considered to be identical to B <==> A, A --> B, etc.). A Rosetta Stone of unique reactions and how they appear in each model was also assembled in a similar format as that for the metabolites, including reaction frequency, a listing of the reaction as written in the database (with and without the stoichiometry included), and the listing of the reaction for each particular model. This is displayed in Table 5 (for simplicity, only the database reaction with stoichiometry is displayed).

Table 5. The Reaction Rosetta Stone. For simplicity, the Database Reaction without Stoichiometry is not included. Reaction directionality and compartment designation were eliminated for purposes of comparison and identification of unique reactions.

Frequency	Database Reactions with Stoichiometry	Acinetobacter baylyi [45]
1	(1) CM00562 + (1) CM00861 <==> (1) CM00261 + (1) CM00266	1 GLT + 1 PYRUVATE <-> 1 2-KETOGLUTARATE + 1 L-ALPHA-ALANINE

The GPR associations for models in which they were included were extracted as part of the reaction data, where possible. Organization of this data was only in the preliminary stages as attempts to compare the reaction networks met with major obstacles. However, the inclusion of this data with their respective reactions is necessary for future applications and comparison.

Experimental Outcomes

Results and Discussion

After compiling a list of metabolites in the form of the Rosetta Stone, a total of 4,636 unique metabolites were found. These "unique" metabolites included the organism-specific metabolites (i.e. "*H. pylori*-specific Lipid X" versus "Lipid X"), which were prevalent in several of the models. In addition, specific metabolites (i.e. β-D-glucose) were read as different from the more ambiguous, general metabolites (i.e. glucose). As described above, the unique metabolite abbreviations were then translated into the unified format (i.e. CM00001) for comparing the reactions in the metabolic networks of each model.

The compilation of this metabolite list allowed for the translation of each model's metabolite abbreviations into those of another or into a new unified format. However, the impact of this result is much larger, as translation of the metabolite abbreviations into a common format is the first step needed to allow one genome-scale metabolic model's reaction "language" to be translated into another in order to compare reactions (as was done here) and to analyze which foreign pathways to incorporate from another organism into a new host. This result also alleviates the potential duplication of metabolites and associated problems when developing new models from existing ones or incorporating non-native pathways into a new host.

Once the different reactions were transformed into a common format to be parsed (excluding reaction directionality, stoichiometry, and compartment designation), the comparisons between models and the number of unique reactions could be determined. From this analysis, it was found that there are 9,898 "unique" reactions. In mapping the

number of reactions that are common to all of the 35 reaction networks compared (34 genome-scale metabolic models and the *E. coli* core model), it was found that only three reactions were common to all of them. Given the number of conserved pathways among many of the organisms compared, this result stood in stark contrast to expectations. The reactions frequency across the 35 reaction networks is displayed in Figure 1. The general trend of very few common reactions across all reaction networks is surprising due to the inherent similarity of the organisms compared. While each organism undoubtedly has its own unique pathways, the complete lack of a common core group of reactions is inconsistent with reality.

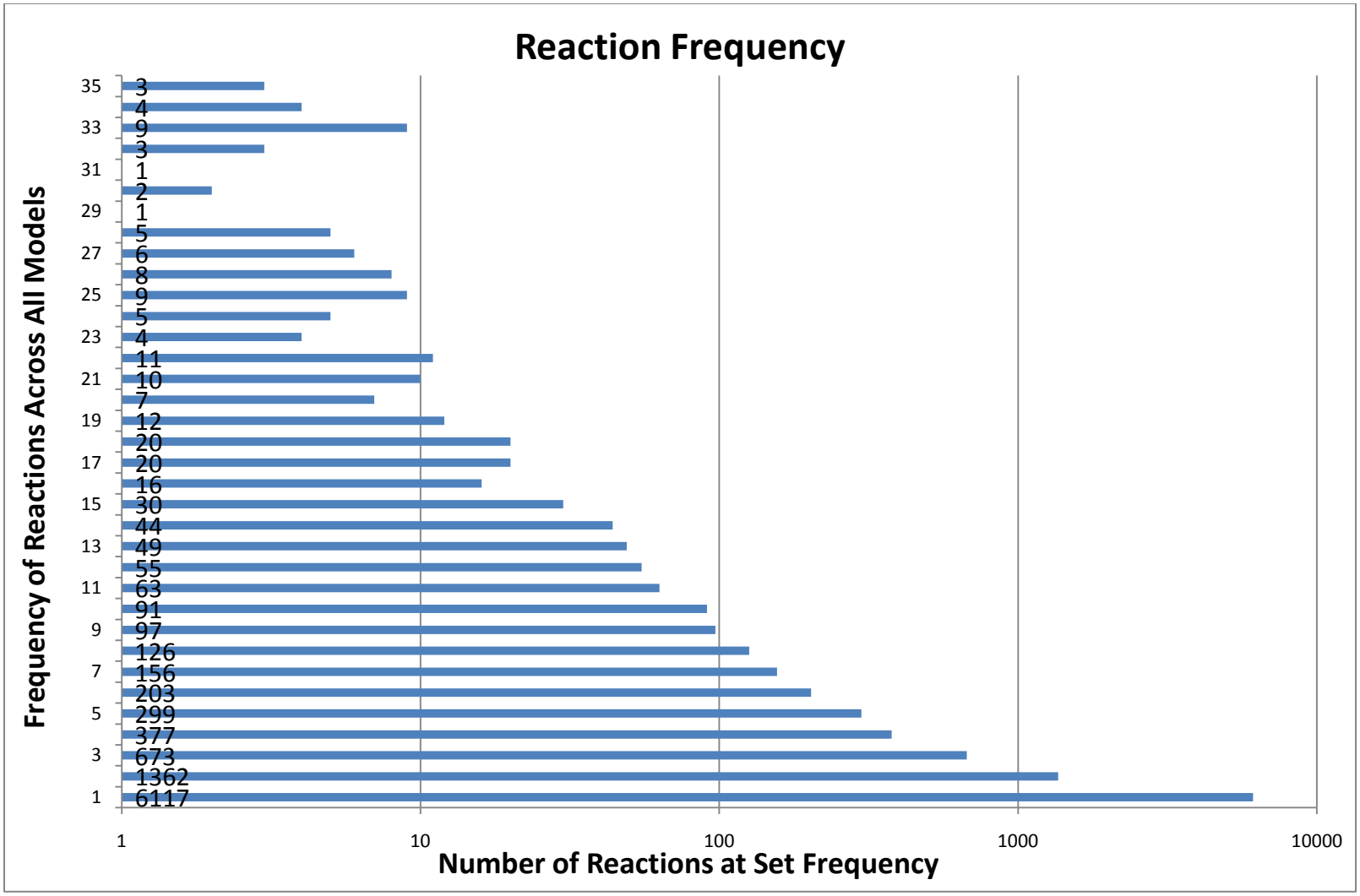

Figure 1. The frequency of the reactions across all 35 models available for comparison. For purposes of this comparison, reaction directionality and compartment designation were neglected. Metabolite abbreviations were translated to a common form for proper comparison.

As is displayed above, the distribution was not as expected. In comparing the different reactions for each of the genome-scale metabolic models, it is readily apparent that there are many underlying reasons for the inconsistency observed. These inconsistencies include incomplete elemental and charge balancing (as displayed in Table 6), alternate cofactor usage among different organisms, and a lack of universal metabolite specificity and naming conventions. To elaborate briefly on this final point, the use of organism-specific metabolites in some models causes disagreement between the different models, and the validity of the difference between the organism-specific metabolite and the general one must be checked in order to determine whether or not these reactions are truly unique from each other. In addition, some models utilize the more ambiguous, general forms of a metabolite, while others are more specific in naming metabolites (i.e. beta-D-glucose versus D-glucose versus glucose). Again, such discrepancies between models must be checked in order to determine the level of dissemblance between the metabolites and the reactions in which they occur.

Table 6. The phosphoenolpyruvate synthesis reaction as it is found in two different models, one completely balanced and another lacking charge and elemental balancing as well as compartment designation. This highlights the inconsistencies that are a common problem for comparing reactions.

Unbalanced Reaction	PYR + ATP -> PEP + AMP + PI
Complete Reaction	[c]: atp + h2o + pyr --> amp + (2) h + pep + pi

By limiting the scope of the comparison to seven models that are more homogeneous in their metabolite and reaction conventions, a correspondence much closer to that expected was observed, as shown in Figures 2 and 3. For these models, 40 reactions were found common to all of them and the general distribution of the reaction frequency data was more in line with reality with a core list of reactions common to all models and all of the models having metabolic network similarity between 20.7-38.5%, despite a wide range of

organisms. The overlap between these models showed better agreement because of the use of more homogeneous conventions. The results observed after comparing these seven models exposes the need for universal conventions for genome-scale metabolic models, as attempts to compare models, incorporate non-native pathways, fill "gaps", achieve strain optimization, or even produce new models from existing ones cannot be successful if the full reaction languages cannot be unified.

Model Name	Bacillus subtilis [25]	Escherichia coli iAF1260 [21]	Methanosarcina barkeri iAF692 [32]	Mycobacterium tuberculosis iNJ661 [26]	Mycoplasma genitalium iPS189 [50]	Saccharomyces cerevisiae iMM904 [30]	Salmonella typhimurium iRR1083 [52]
Bacillus subtilis [25]	100.0	32.7	29.2	34.2	33.6	28.9	32.5
Escherichia coli iAF1260 [21]	32.7	100.0	26.3	34.0	22.5	27.7	37.6
Methanosarcina barkeri iAF692 [32]	29.2	26.3	100.0	38.5	26.7	29.9	24.7
Mycobacterium tuberculosis iNJ661 [26]	34.2	34.0	38.5	100.0	20.7	33.6	32.3
Mycoplasma genitalium iPS189 [50]	33.6	22.5	26.7	20.7	100.0	22.0	25.4
Saccharomyces cerevisiae iMM904 [30]	28.9	27.7	29.9	33.6	22.0	100.0	22.3
Salmonella typhimurium iRR1083 [52]	32.5	37.6	24.7	32.3	25.4	22.3	100.0

Figure 2. Percent Similarity of the balanced models. The percent similarity is defined by 100*√((number of similar reactions^2)/(the product of the total number of reactions in each model)

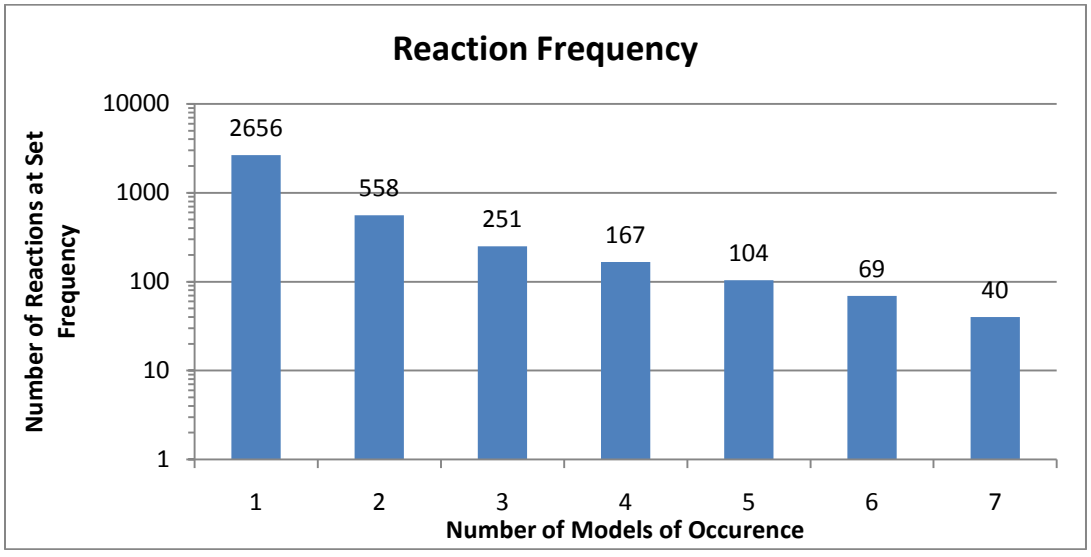

Figure 3. The Reaction Frequency of the seven models of homogeneous conventions. For purposes of this comparison, reaction directionality and compartment designation were neglected. Metabolite abbreviations were translated to a common form for proper comparison.

Future Research

As is evident based on the lack of consistency between the models, steps need to be taken to set standards for completeness, provide universal conventions of genome-scale metabolic models, and rectify problematic models in order to expand their utility for future uses. Until these steps are taken, the full advantages of genome-scale metabolic models associated with biofuel production, drug development, and other advances in biotechnology [13] will not be obtained. Thus, this problem needs to be of primary importance to all of the members of the genome-scale metabolic model community.

One possible solution for future work to this end is the creation of a database which includes a complete list of known metabolites and metabolic reactions with their directionality, compartment designations, and occurrence in different organisms. Each genome-scale metabolic model would also be included in this database as part of its functionality. The creation of this database would impose uniformity by making the complete reaction data readily available for inclusion in new models and forcing the reconciliation of the problems in current models in order to be included. As was the case in comparing the models, the first step needed is to standardize the metabolite names. This includes automation of the procedure utilized for the creation of the Metabolite Rosetta Stone, such that the list can be appended as new genome-scale metabolic models are developed (as discussed below). In addition, a standard set of conventions for metabolite specificity and naming need to be imposed. The use of PubChem identification numbers appears to be a logical solution since the list of chemicals in that database is comprehensive, specific, and unique; all the characteristics that are currently lacking in many of the genome-scale metabolic models. This first step also includes remedying the organism-specific designation for some metabolites (in the cases where

these metabolites are not unique from the general ones) and ensuring specificity (especially for stereoisomers) of all metabolites for models in which they are ambiguous in order to establish the usage of specific metabolites without ambiguity. Next, the reactions will need to be reconciled with regards to elemental and charge balancing, as unbalanced reactions are essentially useless for comparison with other models and other analysis, such as pulling pathways into new or existing models and filling the "gaps" in other models. It is also necessary that standards for completeness of the reaction data and compartment and directionality conventions be imposed for easier comparisons.

Much of the groundwork for this undertaking has already begun from the results of this work. A Metabolite Rosetta Stone that incorporates 38 different models and the *E. coli* core model and their abbreviations utilized for the metabolites along with their associated KEGG identifications (where possible) for translation has already been prepared. However, the Rosetta Stone will need to be made more adaptive and automated for syncing it with other databases and allowing for the inclusion of new models (and their new metabolites) as they become available. In addition, programs have already been developed that are needed to create a database of unique reactions from the genome-scale metabolic models, which will allow for a complete compilation of all of the metabolic networks from the different models once the above problems are rectified.

Finally, there is potential to make such a database more extensive by allowing the inclusion of the kinetic data for each of the enzymes involved in a particular model. The BRENDA database has an extensive listing of kinetic parameters that would be invaluable to experimentalists if it could be included with a database of complete metabolic networks. With the developing tools such as text mining, which has shown to be useful in a variety of similar cases [55], such a comprehensive database is within reach.

Summary and Conclusion

The lack of universal conventions for genome-scale metabolic models including uniform reaction directionality notation, compartment designation, metabolite naming conventions, and standards of completeness and coverage severely limit the effectiveness of these models in solving many of the problems facing the field of metabolic engineering and their utility in applications for a variety of industrial and medical problems. In the current state of the genome-scale metabolic model community, with many groups developing models individually or in close partnership with only a limited number of others, there exists a complete lack of congruency between models and more collaboration is needed to ensure the adoption a certain set of standards. The current lack of collaboration in accepting these standards stands in the way of progress in gap-filling, strain optimization, incorporation of non-native functionalities into new models and organisms, and the development of new metabolic reconstructions. This work highlights the shortcomings of the current genome-scale metabolic models by conveying the inconsistencies between them.

The development of a Metabolite Rosetta Stone allowed for the translation of each genome-scale metabolic model's metabolites into a common form for comparison and analysis of the metabolic pathways. From the 34 models and the *E. coli* core model utilized for determination of the similarity between models, only 3 reactions were found to be common to all 35 reaction networks. This result fell far short of reality, as the disagreement between the models in terms of elemental and charge balancing, cofactor usage, and specific metabolite usage interfered with the recognition of many of the same reactions as being the same. This fact stands as a testament to the need for the adoption of universal conventions and reconciliation of previous models for future use.

In comparing seven models that followed more uniform and consistent conventions, a better result was achieved. For these models, 40 reactions were found to be common throughout all seven models and the percent similarity in the metabolic pathways of these models was in closer agreement and was more consistent with reality. This result portrayed the utility associated with using universal conventions, as the pathways in these models could be readily compared and adapted for use in other models of similar quality and conventions.

To further the progress and utility associated with genome-scale metabolic models, the development of a comprehensive database is a potential solution for these problems. Functionalities including a complete list of unique metabolites and reactions with compartment designation, directionality, and the organism in which it occurs would be of great value for the database and enhance the knowledgebase in the field by compiling all this data in a readily accessible format. Metabolites would be converted to a unified format for specific and unique metabolites (such as PubChem) and existing genome-scale metabolic models would be included once any of the above problems were rectified. In addition, such a database would allow for quicker metabolic reconstructions of new organisms, easier incorporation of non-native functionalities into new hosts, a more aggressive approach to gap-filling efforts, and simplified strain optimization. Finally, it would alleviate the problems currently facing genome-scale metabolic models by aiding in the enforcement of the adoption of universal standards for the models.

In order to allow genome-scale metabolic models to achieve their full potential, the creation of this database is a necessity. While the recent methods to impose more uniform standards will help to enhance the quality of new models [20], steps still must be taken to reconcile existing models to aid in the production of future models and allow for the comparisons to be drawn between new and existing models. With greater

collaboration among all of the contributors to the genome-scale metabolic model community, the full potential of the models can be realized and the resources currently focused on the time-consuming reconciliation and reconstruction processes for models can be shifted towards to ultimate objectives of strain optimization and the insertion of foreign pathways into other organisms for use in both industry and medicine.

References

1. Raab, R. M., K. Tyo, and G. Stephanopoulos. "Metabolic engineering." *Adv Biochem Eng Biotechnol* 100 (2005): 1-17.

2. Niederberger P., R. Prasad, G. Miozzari, H. Kacser (1992) "A strategy for increasing an in vivo flux by genetic manipulations. The tryptophan system of yeast. " *Biochem J* 287:473–479.

3. Becker, S. A., A. M. Feist, M. L. Mo, G. Hannum, B. Ø. Palsson, and M. J. Herrgard. "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox." *Nat Protoc* 2.3 (2007): 723-38.

4. Kim, H. U., T. Y. Kim, and S. Y. Lee. "Metabolic flux analysis and metabolic engineering of microorganisms." *Mol Biosyst* 4.2 (2008): 113-20.

5. Kauffman, K. J., P. Prakash, and J. S. Edwards. "Advances in flux balance analysis." *Curr Opin Biotechnol* 14.5 (2003): 491-96.

6. Burgard, A. P., P. Pharkya, and C. D. Maranas. "Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization." *Biotechnol Bioeng* 84.6 (2003): 647-57.

7. Pharkya, P., A. P. Pharkya, and C. D. Maranas. "OptStrain: a computational framework for redesign of microbial production systems." *Genome Res* 14.11 (2004): 2367-76.

8. Pharkya, P. and C.D. Maranas. "An Optimization Framework for Identifying Reaction Activation/Inhibition or Elimination Candidates for Overproduction in Microbial Systems." *Metabolic Engineering* 8.1 (2006): 1-13.

9. Burgard, A.P., E.V. Nikolaev, C.H. Schilling, and C.D. Maranas. "Flux Coupling Analysis of Genome-scale Metabolic Reconstructions." *Genome Research* 14.2 (2004): 301-312.

10. Park, J. M., T. Y. Kim, and S. Y. Lee. "Constraints-based genome-scale metabolic simulation for systems metabolic engineering." *Biotechnol Adv* 27.6 (2009): 979-88.

11. Blazeck, J., and H. Alper. "Systems metabolic engineering: Genome-scale models and beyond." *Biotechnol J* (2010). *PubMed (Epub ahead of print)*. Web. 3 Feb. 2010.

12. Patil, K. R., M. Akesson, and J. Nielsen. "Use of genome-scale microbial models for metabolic engineering." *Curr Opin Biotechnol* 15.1 (2004): 64-69.

13. Milne, C. B., P. J. Kim, J. A. Eddy, and N. D. Price. "Accomplishments in genome-scale in silico modeling for industrial and medical biotechnology." *Biotechnol J* 4.12 (2009): 1653-70.

14. Kim, H. U., T. Y. Kim, and S. Y. Lee. "Genome-scale metabolic network analysis and drug targeting of multi-drug resistant pathogen Acinetobacter baumannii AYE." *Mol Biosyst* 6.2 (2010): 339-48.

15. Palsson, B. "Metabolic systems biology." *FEBS Lett* 583.24 (2009): 3900-04.

16. Kumar, V. S., and C. D. Maranas. "GrowMatch: an automated method for reconciling in silico/in vivo growth predictions." *PLoS Comput Biol* 5.3 (2009). *PubMed (Epub e1000308)*. Web. 5 Feb. 2010.

17. Edwards, J. S., R. U. Ibarra, and B Ø. Palsson. "In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data." *Nat Biotechnol* 19.2 (2001): 125-30.

18. Reed, J. L., T. R. Patil, *et al.*, "Systems approach to refining genome annotation." *Proc Natl Acad Sci U S A* 103.46 (2006): 17480-84.

19. Fuhrer, T., L. Chen, U. Sauer and D. Vitkup. "Computational prediction and experimental verification of the gene encoding the NAD+/NADP+-dependent succinate semialdehyde dehydrogenase in Escherichia coli" *J. Bacteriol* 189.22 (2007): 8073–8078.

20. Thiele, I., and B. Ø. Palsson. "A protocol for generating a high-quality genome-scale metabolic reconstruction." *Nat Protoc* 5.1 (2010): 93-121.

21. Feist, A. M., C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis, and B. Ø. Palsson. "A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information." *Mol Syst Biol* 3 (2007): 121.

22. Teusink, B., A. Wiersma, D. Molenaar, C. Francke, W. M. de Vos, R. J. Siezen, and E. J. Smid. "Analysis of growth of Lactobacillus plantarum WCFS1 on a complex medium using a genome-scale metabolic model." *J Biol Chem* 281.52 (2006): 40041-48.

23. Oberhardt, M. A., J. Puchalka, K. E. Fryer, V. A. Martins dos Santos, and J. A. Papin. "Genome-scale metabolic network analysis of the opportunistic pathogen Pseudomonas aeruginosa PAO1." *J Bacteriol* 190.8 (2008): 2790-803.

24. Becker, S. A., and B Ø. Palsson. "Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation." *BMC Microbiol* 5 (2005): 8.

25. Oh, Y. K., B. Ø. Palsson, S. M. Park, C. H. Schilling, and R. Mahadevan. "Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data." *J Biol Chem* 282.39 (2007): 28791-99.

26. Jamshidi, N., and B. Ø. Palsson. "Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets." *BMC Syst Biol* q (2007): 26.

27. Nogales, J, B Ø. Palsson, and I Thiele. "A genome-scale metabolic reconstruction of Pseudomonas putida KT2440: iJN746 as a cell factory." *BMC Syst Biol* 2 (2008): 79.

28. Resendis-Antonio, O., J. L. Reed, S. Encarnación, J. Collado-Vides, and B Ø. Palsson. "Metabolic reconstruction and modeling of nitrogen fixation in Rhizobium etli." *PLoS Comput Biol* 3.10 (2007): 1887-95.

29. Duarte, N. C., M. J. Herrgard, and B. Ø. Palsson. "Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model." *Genome Res* 14.7 (2005): 1298-309.

30. Mo, M. L., B. Ø. Palsson, and M. J. Herrgård. "Connecting extracellular metabolomic measurements to intracellular flux states in yeast." *BMC Syst Biol* 3 (2009): 37.

31. Reed, J. L., T. D. Vo, C. H. Schilling, and B Ø. Palsson. "An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR)." *Genome Biol* 4.9 (2003): R54.

32. Feist, A. M., J. C. Scholten, B. Ø. Palsson, F. J. Brockman, and T. Ideker. "Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri." *Mol Syst Biol* 2 (2006).

33. Edwards, J. S., and B Ø. Palsson. "The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities." *Proc Natl Acad Sci U S A* 97.10 (2000): 5528-33.

34. Sheikh, K., J. Förster, and L. K. Nielsen. "Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of Mus musculus." *Biotechnol Prog* 21.1 (2005): 112-21.

35. Kuepfer, L., U. Sauer, and L. M. Blank. "Metabolic functions of duplicate genes in Saccharomyces cerevisiae." *Genome Res* 15.10 (2005): 1421-30.

36. Förster, J., I. Famili, P. Fu, B. Ø. Palsson, and J. Nielsen. "Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network." *Genome Res* 14.2 (2003): 244-53.

37. David, H., I. S. Ozçelik, G. Hofmann, and J. Neilsen. "Analysis of Aspergillus nidulans metabolism at the genome-scale." *BMC Genomics* 9 (2008): 163.

38. Kim, T. Y., H. U. Kim, J. M. Park, H. Song, J. S. Kim, and S. Y. Lee. "Genome-scale analysis of Mannheimia succiniciproducens metabolism." *Biotechnol Bioeng* 97.4 (2007): 657-71.

39. Borodina, I., P. Krabben, and J. Neilsen. "Genome-scale analysis of Streptomyces coelicolor A3(2) metabolism." *Genome Res* 15.6 (2005): 820-29.

40. Schilling, C. H., M. W. Covert, I. Famili, G. M. Church, J. S. Edwards, and B. Ø. Palsson. "Genome-scale metabolic model of Helicobacter pylori 26695." *J Bacteriol* 184.16 (2002): 4582-93.

41. Quek, L. E., and L. K. Nielsen. "On the reconstruction of the Mus musculus genome-scale metabolic network model." *Genome Inform* 21 (2008): 89-100.

42. Gonzalez, O., S. Gronau, M. Falb, F. Pfeiffer, E. Mendoza, R. Zimmer, and D. Oesterhelt. "Reconstruction, modeling & analysis of Halobacterium salinarum R-1 metabolism." *Mol Biosyst* 4.2 (2008): 148-59.

43. Palsson, Bernhard Ø. *Systems Biology: Properties of Reconstructed Networks*. New York: Cambridge University Press, 2006.

44. Thiele, I., T. D. Vo, N. D. Price, and B. Ø. Palsson. "Expanded metabolic reconstruction of Helicobacter pylori (iIT341 GSM/GPR): an in silico genome-scale characterization of single- and double-deletion mutants." *J Bacteriol* 187.16 (2005): 5818-30.

45. Durot, M., F. Le Fèvre, V. de Berardinis, A. Kreimeyer, D. Vallenet, C. Combe, S. Smidtas, M. Salanoubat, J. Weissenbach, and V. Schachter. "Iterative reconstruction of a global metabolic model of Acinetobacter baylyi ADP1 using high-throughput growth phenotype and gene essentiality data." *BMC Syst Biol* 2 (2008): 85.

46. Kjeldsen, K. R., and J. Nielsen. "In silico genome-scale reconstruction and validation of the Corynebacterium glutamicum metabolic network." *Biotechnol Bioeng* 102.2 (2009): 583-97.

47. Sun, J., B. Sayyar, J. E. Butler, P. Pharkya, T. R. Fahland, I. Famili, C. H. Schilling, D. R. Lovely, and R. Mahadevan. "Genome-scale constraint-based modeling of Geobacter metallireducens." *BMC Syst Biol* 3 (2009): 15.

48. Mahadevan, R., D. R. Bond, J. E. Butler, A. Esteve-Nuñez, M. V. Coppi, B. Ø. Palsson, C. H. Schilling, and D. R. Lovely. "Characterization of metabolism in the Fe(III)-reducing organism Geobacter sulfurreducens by constraint-based modeling." *Appl Environ Microbiol* 72.2 (2006): 1558-68.

49. Beste, D. J., T. Hooper, G. Stewart, B. Bonde, C. Avignone-Rossa, M. E. Bushell, P. Wheeler, S. Klamt, A. M. Kierzek, and J. McFadden. "GSMN-TB: a web-based genome-scale network model of Mycobacterium tuberculosis metabolism." *Genome Biol* 8.5 (2007): R89.

50. Suthers, P. F., M. S. Dasika, V. S. Kumar, G. Denisov, J. I. Glass, and C. D. Maranas. "A genome-scale metabolic reconstruction of Mycoplasma genitalium, iPS189." *PLoS Comput Biol* 5.2 (2009). *PubMed (Epub e1000285)*. Web. 13 Feb. 2009.

51. Mazumdar, V., E. S. Snitkin, S. Amar, and D. Segrè. "Metabolic network model of a human oral pathoge." *J Bacteriol* 191.1 (2009): 74-90.

52. Raghunathan, A., J. Reed, S Shin, B. Palsson, and S. Daefler. "Constraint-based analysis of metabolic capacity of Salmonella typhimurium during host-pathogen interaction." *BMC Syst Biol* 3 (2009): 38.

53. Chavali, A. K., J. D. Whittemore, J. A. Eddy, K. T. Williams, and J. A. Papin. "Systems analysis of metabolism in the pathogenic trypanosomatid Leishmania major." *Mol Syst Biol* 4 (2008): 177.

54. Vo, T. D., and B. Ø. Palsson. "Isotopomer analysis of myocardial substrate metabolism: a systems biology approach." *Biotechnol Bioeng* 95.5 (2006): 972-83.

55. Wiegers, T. C., *et al.* "Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics database (CTD)." *BMC Bioinformatics* 10 (2009): 326.

Curriculum Vitae for Stephen Thomas Spagnol

(stephen.spagnol@gmail.com)
Permanent Address
602 Overlook Court
Wexford, PA 15090
Home: (724)933-0818
Cellular: (412)952-5961

Summary of Qualifications

Hardworking, motivated, and persevering Chemical Engineering Student with passion for problem solving and research. Dedicated team worker with strong written and oral communication skills. Skilled and effective in preparing and giving presentations.

Education

The Pennsylvania State University, Schreyer Honors College
University Park, PA
Bachelor of Science in Chemical Engineering, Bioprocess and Biomolecular Engineering Option
Expected Graduation (with Highest Distinction and Honors): May 2010
Cumulative GPA: 3.96/4.00

North Allegheny Senior High School
High School Diploma as Top Scholar with Highest Honors
Cumulative GPA: 4.33/4.00

Research

Chemical & Biological Systems Optimization Lab – Dr. Costas D. Maranas
Pennsylvania State University – Chemical Engineering Department
2008-present

Heart, Lung & Esophageal Surgical Institute – Dr. Rodney Landreneau
University of Pittsburgh Medical Center
2008

Honors and Awards

- Sparks Award for maintaining 4.0 GPA through freshman and fall of sophomore year
- President's Freshman Award for maintaining 4.0 GPA freshman year
- Summer Research Fellowships in Biomolecular Engineering
- National Starch and Chemical Foundation, Inc. Scholarship
- Endowment for the Chemical Engineering Dept. Undergraduate Scholarship (Penn State)
- PHEAA New Economy Technology Scholarship
- Dean's List eight consecutive semesters
- Marsh W. White Scholarship

Extracurricular Activities & Affiliations

- Delta Chi Fraternity
- Omega Chi Epsilon Chemical Engineering Honors Society
- National Society of Collegiate Scholars
- Penn State Dance MaraTHON for Pediatric Cancer

Computational and Laboratory Skills

- Python Programming, MatLab, Mathematica, and HYSYS
- Recombinant DNA Techniques
- DNA Isolation and Purification Techniques
- Protein Isolation and Purification Techniques
- Enzyme Catalysis and Kinetics Techniques
- Differential Scanning Calorimetry
- X-Ray, IR, and NMR Spectroscopy