Final Progress Report
May 21, 2012

Grant Number: DE-FG02-05ER25684 Institution: The Pennsylvania State University
Street Address: Office of Sponsored Programs, 110 Technology Center
City/State/Zip: University Park, PA 16802
Principal Investigator: Costas D. Maranas
Address: 112 Fenske Laboratory, Chemical Engineering Department, University Park PA 16802
Telephone Number: 814-863-9958 Email: costas@psu.edu
DOE/Office of Science Program Office: The Office of Advanced Scientific Computing Research
DOE/Office of Science Program Technical Program Manager Contact: Christine Chalk & Susan Gregurick

A. Overall Technical Summary

Over the last few years, we have witnessed unprecedented progress in the use of microbial production systems for a variety of applications ranging from simple or complex industrial chemicals [1-7] to electrons in biological fuel cells and batteries [8-12]. The hope is that these success stories are the vanguards of novel and efficient bioconversions of biomass derived feeds into liquid fuels ranging from ethanol [13], butanol [14], branched-chain higher alcohols [15] and other high energy density molecules [16] in accordance with the Department of Energy's mission. In pursuit of these milestones, a number of modeling, algorithmic and computational bottlenecks were identified at a recent DOE Workshop on the Computational Research Needs on Alternative and Renewable Energy (CRNARE) (http://www.nrel.gov/crnare_workshop/). Key modeling and computational barriers to success were, among others, (i) the automated generation, curation, archiving and prototyping of metabolic models for microorganisms and plants; (ii) the development of isotope mapping models and computational approaches to support flux elucidation using labeled isotopes (MFA) for large-scale metabolic models; and (iii) the need for reliable algorithms to identify engineering strategies that lead to the targeted biomass conversion target. This

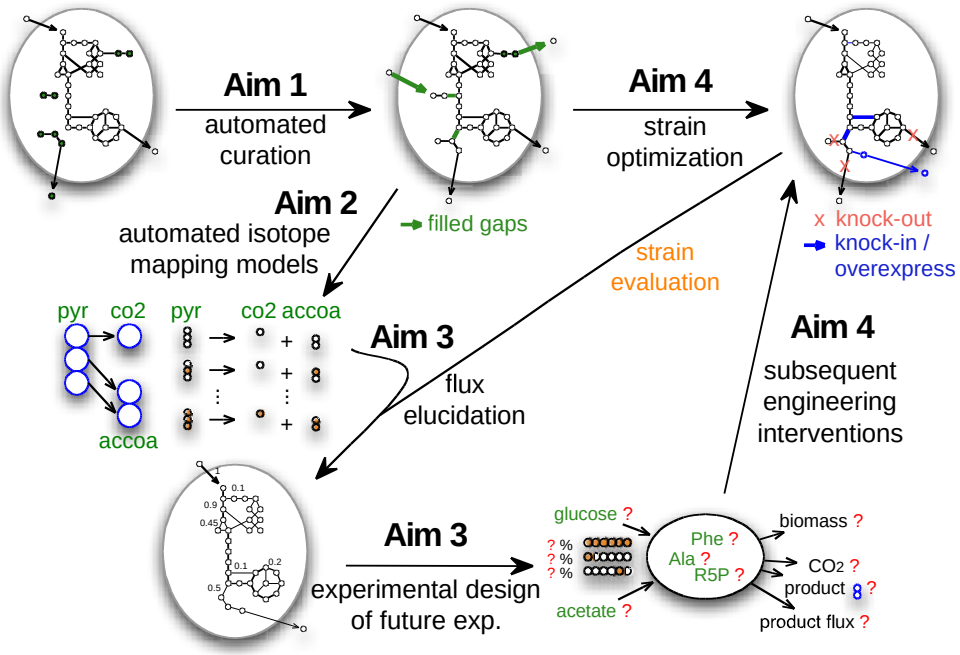

Figure 1: The general aims of this project are: (Aim 1) to generate integrated computational tools for the automated generation and curation of genome-scale models of metabolism for microbial and plant systems; (Aim 2) to automatically generate maps tracking the fate of labeled isotopes through genome-scale models; (Aim 3) to fully elucidate metabolic fluxes in genome-scale models using GC-MS or NMR data; (Aim 4) to leverage flux data for wild-type strain to identify all possible engineering strategies that lead to overproduction of a targeted product.

current research directly address these challenges by building on research milestones already reached with the support of DOE (DE-FG02-05ER25684). The research will lead to an improved capability of mapping, elucidating and re-directing metabolic flows in microbial strains and plants and thus directly contribute to DOE's bioconversion missions. To accomplish these objectives, we put forth four aims outlined in Figure 1.

We have promptly posted on the PIs webpage (http://maranas.che.psu.edu/) and broadly disseminated all data as well as the obtained models and computational tools in accordance with DOE's policy.

Progress has been made on all fronts since the time of our previous progress report and we are well on pace to meet and hopefully surpass all milestones that have been put forth in the proposal. The work has thus far yielded a number of successful developments both in the area of computational platforms to support all of our modeling aims for the coming year of this proposal, and in the area of scientific/technical advances. Below is further information on the progress related to the specific individual aims as outlined in our original proposal.

Aim 1: Automated Generation and Curation of Genome-Scale Metabolic Reconstructions

Aim 2: Automated Generation of Genome-Scale Isotope Mapping Models

Aim 3: Metabolic Flux Elucidation Algorithms Given GC-MS or NMR data

Aim 4: Computational Strain Design

The following sections detail our progress made during the second year of the award towards the project aims along a multitude of fronts in the development of computational tools to analyze, elucidate and redesign biological pathways. The ultimate outcome of the work will be a suite of computational aids for analyzing and optimizing the behavior of biological networks. We are confident that the success of the first two years of this three-year research program are a solid indication that we will be able to accomplish all of the stated objectives of this initiative.

B. Specific Aim 1: Automated Generation and Curation of Genome-Scale Metabolic Reconstructions

B.1. MetRxn: A Knowledgebase of Metabolites and Reactions Spanning Metabolic Models and Databases

The work in this section has been published [17].

B.1.1 Background

`The ever accelerating pace of DNA sequencing and annotation information generation [18] is spearheading the global inventorying of metabolic functions across all kingdoms of life. Increasingly, metabolite and reaction information is organized in the form of community [19], organism, or even tissue-specific genome-scale metabolic reconstructions. These reconstructions account for reaction stoichiometry and directionality, gene to protein to reaction associations, organelle reaction localization, transporter information, transcriptional regulation and biomass composition. Already over 75 genome-scale models are in place for eukaryotic, prokaryotic and archaeal species [20] and are becoming indispensable for computationally driving engineering interventions in microbial strains for targeted overproductions [21-24], elucidating the organizing principles of metabolism [25-28] and even pinpointing drug targets [29, 30]. A key bottleneck in the pace of reconstruction of new high quality metabolic models is our inability to directly make use of metabolite/reaction information from biological databases [31] (e.g., BRENDA [32], KEGG [33], MetaCyc, EcoCyc, BioCyc [34], BKM-react [35], UM-BBD [36], Reactome.org,

Rhea, PubChem, ChEBI etc.) or other models [37] due to incompatibilities of representation, duplications and errors, as illustrated in Figure B.1.1.

A major impediment is the presence of metabolites with multiple names across databases and models, and in some cases within the same resource, which significantly slows down the pooling of information from multiple sources. Therefore, the almost unavoidable inclusion of multiple replicates of the same metabolite can lead to missed opportunities to reveal (synthetic) lethal gene deletions, repair network gaps and quantify metabolic flows. Moreover, most data sources inadvertently include some reactions that may be stoichiometrically inconsistent [38] and/or elementally / charge unbalanced [39, 40], which can adversely affect the prediction quality of the resulting models if used directly. Finally, a large number of metabolites in reactions are partly specified with respect to structural information and may contain generic side groups (e.g., alkyl groups -R), varying degree of a repeat unit participation in oligomers, or even just compound class identification such as “an amino acid” or “electron acceptor”. Over 3% of all metabolites and 8% of all reactions in the aforementioned databases and models exhibit one or more of these problems.

There have already been a number of efforts aimed at addressing some of these limitations. The Rhea database, hosted by the European Bioinformatics Institute, aggregates reaction data primarily from IntEnz [41] and ENZYME [42], whereas Reactome.org is a collection of reactions primarily focused on human metabolism [43, 44]. Even though they crosslink their data to one or more popular databases such as KEGG, ChEBI, NCBI, Ensembl, Uniprot, etc., both retain their own representation formats. More recently, the BKM-react database is a non-redundant biochemical reaction database containing known enzyme-catalyzed reactions compiled from BRENDA, KEGG, and MetaCyc [35]. The BKM-react database currently contains 20,358 reactions. Additionally, the contents of five frequently used human metabolic pathway databases have been compared [45]. An important step forward for models was the BiGG database, which includes seven genome-scale models from the Palsson group in a consistent nomenclature and exportable in SBML format [46-48]. Research towards integrating genome-scale metabolic models with large databases has so far been even more limited. Notable exceptions include the partial reconciliation of the latest *E. coli* genome scale model *i*AF1260 with EcoCyc [49] and the aggregation of data from the *Arabidopsis thaliana* database and KEGG for generating genome-scale models [50] in a semi-automated fashion. Additionally, ReMatch integrates some metabolic models, although its primary focus is on carbon mappings for metabolic flux analysis [51]. Also, many metabolic models retain the KEGG identifiers of metabolites and reactions extracted during their construction [52, 53]. An important recent development is the web resource Model SEED that can generate draft genome-scale metabolic models drawing from an internal database that integrates KEGG with 13 genome scale models (including six of the models in the BiGG database) [54]. All of the reactions in Model SEED and BiGG are charge and elementally balanced.

In this work, we describe the development and highlight applications of the web-based resource MetRxn that integrates, using internally consistent descriptions, metabolite and reaction information from 8 databases and 44 metabolic models. The MetRxn knowledgebase (as of October 2011) contains over 76,000 metabolites and 72,000 reactions (including unresolved entries) that are charge and elementally balanced. By conforming to standardized metabolite and reaction descriptions, MetRxn enables users to efficiently perform queries and comparisons across models and/or databases. For example, common metabolites and/or reactions between models and databases can rapidly be generated along with connected paths that link source to target metabolites. MetRxn supports export of models in SBML format. New models are being added as they are published or made available to us. It is available as a web-based resource at http://metrxn.che.psu.edu.

B.1.2 Construction and Content

MetRxn construction

The construction of MetRxn largely followed the following steps, as illustrated in Figure B.1.2: 1) download of primary sources of data from databases and models, 2) integration of metabolite and reaction data, 3) calculation and reconciliation of structural information, 4) identification of overlaps between metabolite and reaction information, 5) elemental and charge balancing of reactions, 6) successive resolution of remaining ambiguities in description.

Step 1: *Source data acquisition*. Metabolite and reaction data was downloaded from BRENDA, KEGG, BioCyc, BKM-react and other databases using a variety of methods [6,52-57] based on protocols such as SOAP, FTP and HTTP. We preprocessed the data into flat files that were subsequently imported into the knowledgebase. All original information pertaining to metabolite name, abbreviations, metabolite geometry, related reactions, catalyzing enzyme and organism name, gene-protein-reaction associations, and compartmentalization was retained. For all 44 initial genome-scale models listed, the online information from the corresponding publications was also imported. The source codes for all parsers used in Step 1 are available on the MetRxn website.

Step 2: Source data parsing. The "raw data" from both databases and models was unified using standard SQL scripts on a MySQL server. The description schema for metabolites includes source, name, abbreviations used in the source, chemical formula, and geometry. The schema for reactions accounts for source, name, reaction string (reactants and products), organism designation, associated enzymes and genes, EC number, compartment, reversibility/direction, and pathway information. Once a source has been imported into the MySQL server, a data source-specific dictionary is created to map metabolite abbreviations onto names/synonyms and structures and metabolites to reactions.

Step 3: Metabolite charge and structural analysis. We used Marvin (Chemaxon) to analyze all 218,122 raw metabolite entries containing structural information (out of a total of 322,936, including BRENDA entries). Inconsistencies were found in 12,965 entries typically due to wrong atom connectivity, valence, bond length or stereo chemical information, which were corrected using APIs available in Marvin. A final corrected version of the metabolite geometries was calculated at a fixed pH of 7.2 and converted into standard Isomeric SMILES format. The structure/formula used corresponded to the major microspecies found during the charge calculation, which effectively rounds the charge to an integer value in accordance with previous model construction conventions. This format includes both chiral and stereo information, as it allows specification of molecular configuration [55-57]. Metabolites were also annotated with Canonical SMILES using the OpenBabel Interface from Chemspider. The canonical representation encodes only atom-atom connectivity while ignoring all conformers for a metabolite. Using bond connectivity information from the primary sources and resources such as PubChem and ChemSpider we used Canonical SMILES [58, 59] to resolve the identity of 34,984 metabolites and 32,311 reactions. Another 6,100 metabolites and 11,401 reactions involved, in various degrees, lack of full atomistic detail in their description (e.g., use an R or X as side-chains, are generic compounds like "amino acid" or "electron acceptor"). Over 25,000 duplicate metabolites and 27,000 reaction entries were identified and consolidated within the database. The metabolites and reactions present in the resolved repository were further classified with respect to the completeness of atomistic detail in their description.

Step 4: Metabolite synonyms and initial reaction reconciliation. Raw metabolite entries were assigned to Isomeric SMILES representations whenever possible. If insufficient structural information was available for a downloaded raw metabolite then it was assigned temporarily with the Canonical SMILES and revisited during the reaction reconciliation. Canonical SMILES retain

atom connectivity but not stereo-specificity and are used as the basic metabolite topology descriptors as many metabolic models lack stereo-specificity information. After generating the initial metabolite associations, we identified reaction overlaps using the reaction synonyms and reaction strings along with the metabolite SMILES representations. Directionality and cofactor usage were temporarily ignored. During this step, reactions were flagged as single-compartment or two-compartment (i.e., transport reactions). MetRxn internally retains the original compartment designations, but currently only displays these simplified compartment designations. In analogy to metabolites, reactions were grouped into families that shared participants but in the source data sets occurred in different compartments or differed only in protonation.

Step 5: *Reaction charge and elemental balancing.* Once metabolites were assigned correct elemental composition and protonation states, reactions were charge and elementally balanced. To this end, for charge balancing we relied on a linear programming representation that minimizes the difference in the sum of the charge of the reactants and the sum of the charge on the products. The complete formulation is provided in the documentation at MetRxn.

Step 6: Iterative reaction reconciliation. Reactions with one (or more) unresolved reactants and/or products were string compared against the entire resolved collection of reactions. This step was successively executed as newly resolved metabolites and reactions could enable the resolution of previously unresolved ones. After the first pass 164 metabolites were resolved, while subsequent passes (up to 18 for some models) helped resolved a total of 8,720 entries. Reactions with significant (but not complete) overlapping sets of reactants/products are additionally sent to the curator GUI including phonetic information. Briefly, the phonetic tokens of synonyms with known structures were compared against the ones without any associated structure. The algorithm suppresses keywords/tokens depicting stereo information such as *cis*, *trans*, L-, D-, alpha, beta, gamma, and numerical entries because they change the phonetic signature of the synonym under investigation. In addition, the algorithm ignores non-chemistry related words (e.g., use, for, experiment) that are found in some metabolite names. Certain tokens such as "-ic acid" and "-ate" are treated as equivalent. PubChem and Chemspider sources were accessed through the GUI so that the curator gets as much information as possible to identify the data correctly. Phonetic matches provided clues for resolving over 159 metabolites. The iterative application of string and phonetic comparison algorithms resolved as many as 8,879 metabolites after 18 rounds of reconciliation.

Upon completion of this workflow, all genome-scale models are reformatted into a computations-ready form and Flux Balance Analysis [60] is performed on both the source model and the standardized model in MetRxn to ascertain the ability of the model to produce biomass before and after standardization. We performed the calculations using GAMS version 12.6. MetRxn is accessible through a web interface that indirectly generates MySQL queries. In order to facilitate analysis and use of the data, a number of tools are provided as part of MetRxn.

Data export and display

MetRxn supports a number of export capabilities. In general, any list that is displayed contains live links to the metabolite or reaction entities. These lists can consist of an entire model, data from a comparison, or query results. All items can be exported to SMBL format. In addition, the public MySQL database will be made available upon request. Because of licensing limitations, the BRENDA database cannot be exported and is not part of the public MySQL database. However, we plan to provide Java source code that allows for the integration of a local copy of the public MySQL database with the BRENDA database (provided upon request).

Source comparisons and visualization

In addition to listing the content (number of metabolites, reactions, etc.) of the selected data source(s), MetRxn contains tools for comparing two or more models and visualizing the results. These associations can be for metabolites or reactions. During these comparisons compartment information and reversibility are suppressed. Comparison tables are generated by comparing the associations between the selected data source(s) using the canonical structures.

MetRxn Scope

An initial repository of reaction (i.e., 154,399) and metabolite (i.e., 322,936) entries were downloaded from 8 databases and 44 genome-scale metabolic models. We compiled a non-redundant list of 42,540 metabolites and 35,474 reactions (after consolidating duplicate entries) containing full atomistic and bond connectivity detail. Another 6,100 metabolites and 11,401 reactions have partial atomistic detail typically containing generic side-chains (R) and/or an unspecified number of polymer repeat units. Finally, 5,436 metabolites in metabolic models and 8,000 metabolites in databases are retained with no atomistic detail. In some cases lack of atomistic detail reflects complete lack of identity specificity (e.g., electron donor) whereas in other cases even though the chemical species is fully defined, atomistic level description is not warranted (e.g., gene product of dsbC protein disulfide isomerase II (reduced)). Figure B.1.3 shows the distribution of metabolite resolution across models and databases in MetRxn. In general, metabolites without fully-specified structures tend to participate in a relatively small number of reactions.

The workflow followed in the creation of the MetRxn knowledgebase identified a number of inconsistencies. For instance, the same metabolite name may map to molecules with different numbers of repeat units (e.g., lecithin) or completely different structures (e.g., AMP could refer to either adenosine monosphate or ampicillin). Notably, even for the most well-curated metabolic model, *E. coli i*AF1260 [49], we found minor errors or omissions (a total of 17) arising from inconsistencies or incompleteness of representation in the data culled from other sources. For example, the metabolite abbreviation arbtn-fe3 was mistakenly associated with the KEGG ID and structure of aerobactin instead of ferric-aerobactin. The number of inconsistencies is dramatically increased for less-curated metabolic models. We used a variety of procedures to disambiguate the identity of metabolites lacking structural information ranging from reaction matching to phonetic searches. For example, in the *Corynebacterium glutamicum* model [61], 7,8-aminopelargonic acid (DAPA) has no associated structural information. Reaction matching found the same reaction in the *E. coli i*AF1260 model.

C. glutamicum DAPA + ATP + CO2 <=> DTBIOTIN + ADP + PI

*i*AF1260 [c] : atp + co2 + dann --> adp + dtbt + (3) h + pi

which implies that 7,8-aminopelargonic acid (DAPA) is identical to 7,8-Diaminononanoate (dann). Examination of pelargonic acid and nonanoate reveals that they were indeed known synonyms. In many cases, we were also able to assign stereo-specific information to metabolite entries in models (e.g., stipulate the L-lysine isomer for lysine). We made use of an iterative approach that allowed us to map structures from models with explicit links to structures (e.g. to KEGG or CAS numbers) to models that only provided metabolite names. Furthermore, by using a phonetic algorithm that uses tokens for equivalent strings in metabolite names (e.g., '-ic acid' and '-ate' are equivalent) we were able to resolve more than an additional 159 metabolites. For example, phonetic searches flagged cis-4-coumarate and COUMARATE in the *Acinetobacter baylyi* model [62] as potentially identical compounds. Additional checks revealed that indeed both metabolites should map to the same structure. A more complex matching example involved 1-(5'-Phosphoribosyl)-4-(N-succinocarboxamide)-5-aminoimidazole from the *Bacillus subtilis*

model [63] and 1-(5'-Phosphoribosyl)-5-amino-4-(N-succinocarboxamide)-imidazole from the *Aspergillus nidulans* model [64]. We note that the phonetic algorithm only makes suggestions and orders the possible matches for the curator. Next, we detail three examples that provide an insight into the type of tasks that MetRxn can facilitate.

B.1.3 Utility and Discussion

1. Charge and elementally balanced metabolic models

The standardized description of metabolites and balanced reactions afforded by MetRxn enables the expedient repair of existing models for metabolite naming inconsistencies and reaction balancing errors. Here we highlight one such metabolic model repair for *Acinetobacter baylyi* *i*Abaylyi^{v4} [62]. We identified that 189 out of 880 reactions are not elementally or charge balanced. Most of the reactions with charge balance errors involved a missed proton in reactions involving cofactor pairs such as NAD/NADH. For example, a proton had to be added to the reactants side in the reaction (R,R)-Butanediol-dehydrogenase in which butanediol reacts with NAD to form acetoin. In addition, the stoichiometric coefficient of water in GTP cyclohydrolase I was erroneously set at -2 which resulted in an imbalance in oxygen atoms. The re-balancing analysis changed the coefficient to -1 (as listed in BRENDA) and added a proton to the list of reactants (absent from BRENDA) in order to also balance charges.

We performed flux balance analysis (FBA) on both the published and MetRxn-based rebalanced version of the *Acinetobacter baylyi* model using the uptake constraints listed in [62] to assess the effect of re-balancing reaction entries on FBA results. We found that the maximum biomass using the glucose/ammonia uptake environment decreased by 9% primarily due to the increased energetic costs associated with maintaining the proton gradient. This result demonstrates the significant effect that lack of reaction balancing may cause in FBA calculations. Overall, we found that nearly two-thirds of the models had at least one unbalanced reaction, with over 2,400 entities across all models that were either charge or elementally imbalanced. Frequently, the same reaction was imbalanced in multiple models (each occurrence was counted separately).

2. Contrasting existing metabolic models

At the onset of creating MetRxn, we conducted a brief preliminary study to quantify the extent/severity of naming inconsistencies by contrasting the reaction information contained in an initial collection of 34 of the most popular genome-scale models spanning 21 bacterial, 10 eukaryotic and three archaeal organisms. Across all branches of life, most metabolic processes are largely conserved (e.g., glycolysis, pentose phosphate pathway, amino acid biosynthesis, etc.) therefore we expected to uncover a large core of common reactions shared by all models. Surprisingly, we found that only three reactions (i.e., phosphoglycerate mutase, phosphoglycerate kinase, and CO_2 transport) were directly recognized as common across those 34 models using a simple string match comparison. Even when examining models for only a few bacterial organisms (*Bacillus subtilis*, *Escherichia coli*, *Mycobacterium tuberculosis*, *Mycoplasma genitalium*, and *Salmonella* Typhimurium) simple text searches recognized only 40 common reactions (out of a possible 262, which is the size of the *M. genitalium* model). The reason for this glaring inconsistency is that differing metabolite naming conventions, compartment designations, stoichiometric ratios, reversibility, and water/proton balancing issues prevents the automated recognition of genuinely shared reactions across models. Using the glucose-6-phosphate dehydrogenase reaction as a representative example, Table 1 reveals some of the reasons for failing to automatically recognize common reactions across selected models [30, 49, 52, 53, 64-81]. As many as nine different representations of the same reaction exist due to incomplete elemental and charge balancing, alternate cofactor usage among different organisms, and lack of universal metabolite naming conventions. We have found that this level of discord between models is representative for most metabolic reactions. This lack of consistency renders direct pathway

comparisons across models meaningless and the aggregation of reaction information from multiple models precarious. This deficiency motivated the development of MetRxn. Given standardization in metabolite naming and elementally / charge balanced reaction entries MetRxn allows for the identification of shared reactions as well as differences between any two metabolic models (assuming that all the metabolites in the compared reaction entries have full atomistic information). When making the comparison of those same metabolic models, MetRxn found an additional 15 reactions in common (for a total of 55 — a 38% increase) and that 142 reactions are shared by *B. subtilis*, *E. coli* and *Salmonella* Typhimurium.

The Web interface of MetRxn allows for any number of models to be simultaneously compared. As a demonstration of this capability we selected to contrast the metabolic content of two clostridia models: *Clostridium acetobutylicum* [82] and *Clostridium thermocellum* [83]. Figure B.1.4 shows the results in the form of a Venn diagram. Some of the differences between the clostridia species are not surprising arising due to their differing lifestyles (*C. acetobutylicum* contains solventogenesis pathways and a CoB12 pathway, whereas *C. thermocellum* contains cellulosome reactions). However, we found many differences that appear to reflect different conventions adopted when the two models were generated rather than genuine differences in metabolism. In particular, in the *C. thermocellum* model [83] charged/uncharged tRNA metabolites are explicitly tracked whereas they are not included in the *C. acetobutylicum* model [82]. Surprisingly, both clostridia models are more similar, at the metabolite level, to the *Bacillus subtilis* iBsu1103 model [63] rather than to each other (see Figure B.1.4). Charged/uncharged tRNA metabolites account for most of the increased overlap between *C. thermocellum* and *B. subtilis*. Most of the reaction overlaps are in the amino acids biosynthesis pathways, carbohydrate metabolism, and nucleoside metabolism. It is important to note that 48 reactions in *C. acetobutylicum*, 67 reactions in *C. thermocellum*, and 120 reactions in *B. subtilis* lack full atomistic information (see Figure B.1.3) and thus were excluded from any comparisons. It is possible that additional shared reactions between the two models can be deduced by further examining comparisons between not fully structurally specified metabolite entries. The string / phonetic comparison algorithms described under Step 6 along with assisted curation could be adapted for this task.

3. Using MetRxn to Bio-Prospect for Novel Production Routes

A "Grand Challenge" in biotechnological production is the identification of novel production routes that allow for the conversion of inexpensive resources (e.g., various sugars) into useful products (e.g., succinate, artemisinin) and bio-fuels (e.g., ethanol, butanol, biodiesel etc.). Selected production routes must exhibit high yields, avoid thermodynamic barriers, bypass toxic intermediates and circumvent existing intellectual property restrictions. Historically, the incorporation of heterologous pathways relied largely on human intuition and literature review followed by experimentation [84, 85]. Currently, rapidly expanding compilations of biotransformations such as KEGG [86] and BRENDA [87] are increasingly being prospected using search algorithms to identify biosynthetic routes to important product molecules. Several optimization and graph-based methods have been employed to computationally assemble novel biochemical routes from these sources. OptStrain [88] used a mixed-integer linear optimization representation to identify the minimal number of reactions to be added (i.e. knock-ins) into a genome-scale metabolic model to enable the production of the new molecule. However the combinatorial nature of the problem poses a significant challenge to the OptStrain methodology as the number of reaction database entries increase from a few to tens of thousands. At the expense of not enforcing stoichiometric balances, graph-based algorithms have inherently better-scaling properties for exhaustively identifying all min-path reaction entries that link a source with a target metabolite. Hatzimanikatis *et. al.* [89] introduced a graph-based heuristic approach (BNICE) to identify all possible biosynthetic routes from a given substrate to a target chemical by hypothesized enzymatic reaction rules. In addition, the BNICE framework was used to identify

novel metabolic pathways for the synthesis of 3-hydroxypropionate in *E. coli* [90]. Based on a similar approach, a new scoring algorithm [91] was introduced to evaluate and compare novel pathways generated using enzyme-reaction rules. In addition, several techniques such as PathMiner [92], PathComp [93], Pathway Tools [94, 95], MetaRoute [96], PathFinder [97] and UM-BBD Pathway Prediction System [98] have been used to search databases for bioconversion routes.

We recently published [99] a graph-based algorithm that used reaction information from BRENDA and KEGG to exhaustively identify all connected paths from a source to a target metabolite using a customized min-path algorithm [100]. We first demonstrated the min-path procedure by identifying all synthesis routes for 1-butanol from pyruvate using a database of 9,921 reactions and 17,013 metabolites manually extracted from both BRENDA and KEGG. Here, we re-visited the same task using the full list of reactions and metabolites present in MetRxn to assess the discovery potential of using MetRxn. Figure B.1.5 illustrates all identified pathways from pyruvate to 1-butanol before MetRxn (29, shown in blue) and the ones discovered after using MetRxn (112, shown in green). As many as 83 new avenues for 1-butanol production were revealed as a consequence of using the expanded and standardized MetRxn resource. In addition, the search algorithm recovered known [101-105] synthesis routes using *E. coli* for the production of 1-butanol (shown in orange). The first pathway involves the fermentative transformation of pyruvate and acetyl-CoA to 1-butanol using enzymes from *C. acetobutylicum* [15]. The second pathway uses ketoacid precursors [101]. This example demonstrates how the biotransformations stored in MetRxn can be used to traverse a multitude of production routes for targeted bioproducts.

B.1.4 Conclusions

MetRxn enables the standardization, correction and utilization of rapidly growing metabolic information for over 76,000 metabolites participating in 72,000 reactions (including unresolved entries). The library of standardized and balanced reactions streamlines the process of reconstructing organism-specific metabolism and opens the way for identifying new paths for metabolic flux redirection. Moreover, the standardization of published genome-scale models enables the rapidly growing community of researchers who make use of metabolic information to understand metabolism at an organism-level and re-deploy it for various biotechnological objectives. By removing standardization and data heterogeneity bottlenecks the pace of knowledge creation and discovery from users of this resource will be accelerated. MetRxn is constructed in a way that allows for quick updating and tracking of changes that occur in the primary databases, as well as available parsing tools that allow for rapid import of new genome-scale metabolic models as they become available. By having exports in SBML, MetRxn's output can be directly interfaced with software packages such as the COBRA toolbox.

During the construction of the initial release of MetRxn, we managed to associate structures for over 8,800 metabolites and re-balanced more than 2,400 reaction instances across 44 metabolic models. This enables the genuine comparison of metabolic content between metabolic models. Preliminary results reinforce that that discrepancies between metabolic models echo not only genuine differences in metabolism but also assumptions and workflow followed by the model creator(s). Going forward, we will continue to expand MetRxn to include more genome-scale metabolic models and add additional tools to aid in their analysis. Because we anticipate that the scope and number of models will rapidly expand, we plan to invite and encourage the community to offer comments about metabolite and reaction information as well as provide feedback on MetRxn itself.

B.1.4 Availability and Requirements

MetRxn is available at http://metrxn.che.psu.edu. Its use is freely available for all non-commercial activity.

1) Naming Inconsistencies

2-Oxoglutarate + L-Alanine <=> Pyruvate + L-Glutamate

KEGG	C00026 + C00041 <=> C00022 + C00025
BRENDA	alpha-ketoglutarate + L-alanine <=> L-glutamate + pyruvate
*Escherichia coli i*AF1260	[c] : akg + ala-L --> glu-L + pyr
Acinetobacter baylyi iAbaylyi	1 GLT + 1 PYRUVATE <-> 1 2-KETOGLUTARATE + 1 L-ALPHA-ALANINE
*Leishmania major i*AC560	[m] : akg + ala-L -> glu-L + pyr
Mannheimia succiniciproducens	PYR + GLU --> AKG + ALA

2) Elemental and charge imbalances

Balanced

KEGG	(R)-Lactate + NAD+ <=> Pyruvate + NADH + H+
*Escherichia coli i*AF1260	[c] : lac-D + nad --> h + nadh + pyr

Unbalanced

Acinetobacter baylyi iAbaylyi	1 D-LACTATE + 1 NAD <=> 1 NADH + 1 PYRUVATE

3) Errors/incompleteness/ambiguity in structural information

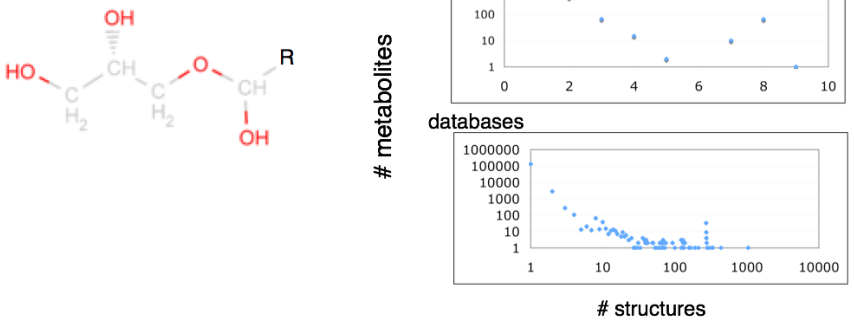

Figure B.1.1: Typical incompatibilities and inconsistencies in genome-scale models and databases. Roadblocks to using genome-scale models and databases include ambiguities and differences in naming conventions, lack of balanced reactions, and incompleteness of structural information.

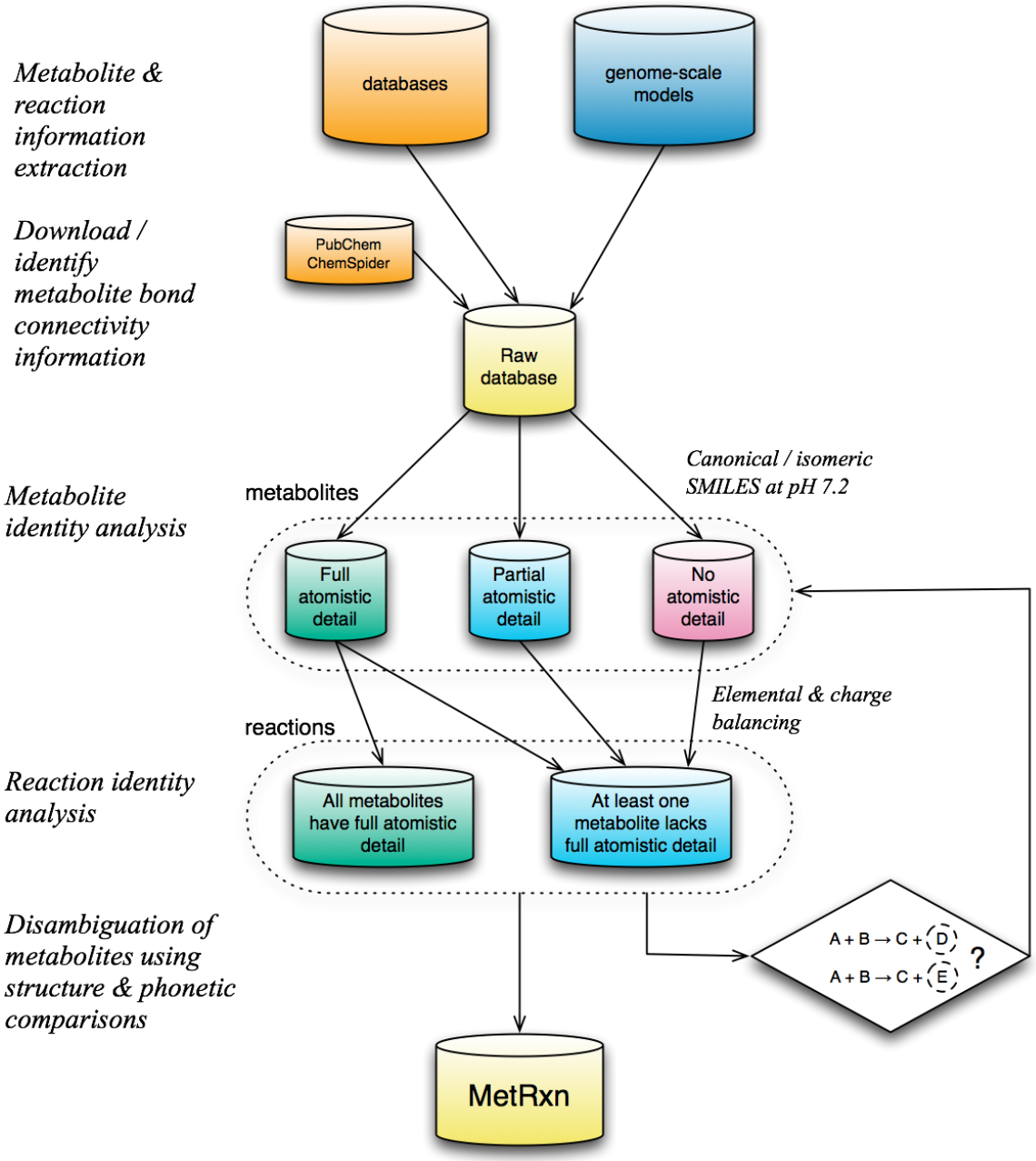

Figure B.1.2: Flowchart outlining the construction of MetRxn. After download of primary sources of data from databases and models, we integrated metabolite and reaction data, followed by calculation and reconciliation of structural information. By identifying overlaps between metabolite and reaction information, we generated elemental and charge balancing of reactions. The procedure for developing MetRxn was iterative with subsequent passes making use of previous associations to resolve remaining ambiguities.

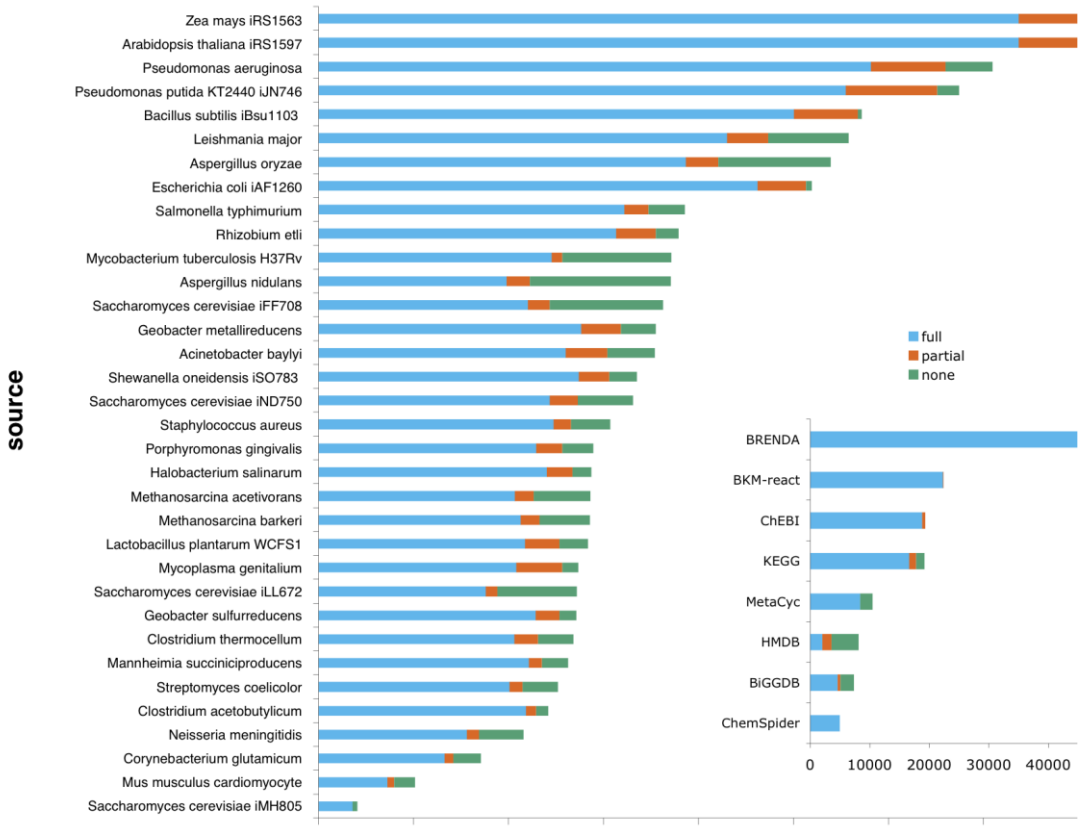

Figure B.1.3: Various levels of structural information was available for models (main) and databases (inset). For every model, the majority of metabolites had full atomistic detail (blue). The smaller number of metabolites with partial atomistic detail (orange) such as genetic side chains, or with no atomistic detail (green) such as gene products, participated in few reactions.

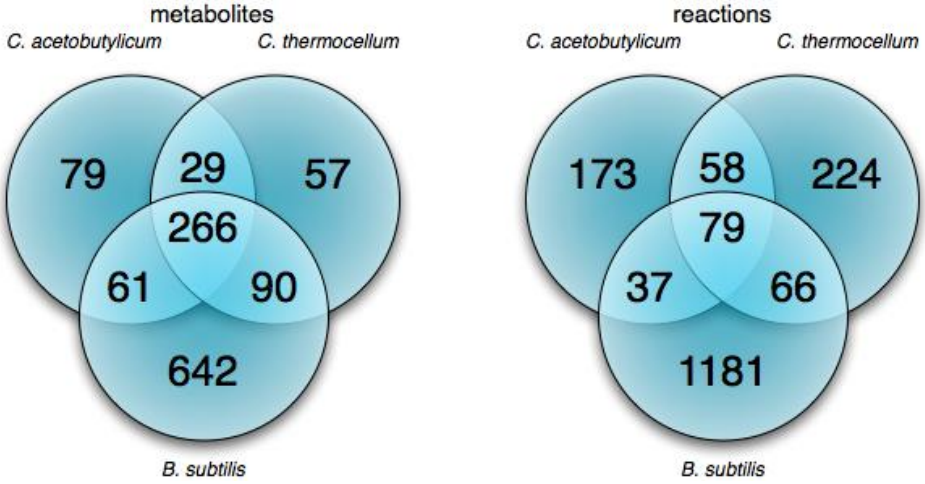

Figure B.1.4: Comparison of metabolite and reaction overlaps for *C. acetobutylicum* and *C. thermocellum* (A). Although the organisms are same genus, the models of these two species had significant numbers of unique metabolites (left) and reactions (right). Additional comparisons revealed that there was more similarity in metabolite usage with a model of B. subtilis than with each other. In part, these overlaps were driven by the explicit accounting for charged tRNA species in *C. thermocellum* and *B. subtilis* models, which was also reflected in the reaction overlaps through reactions involving these metabolites.

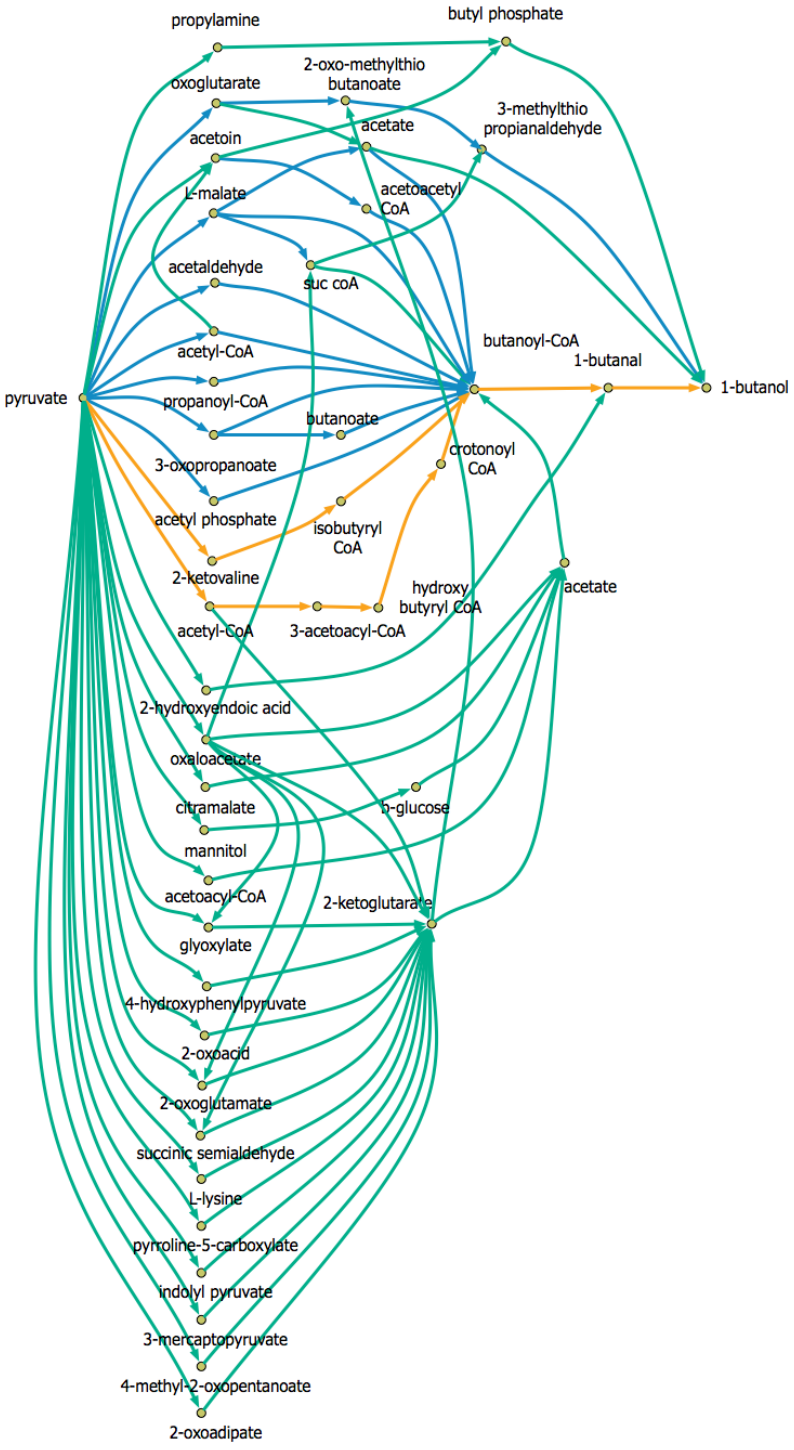

Figure B.1.5: Pathways from pyruvate to 1-butanol. Using the MetRxn knowledgebase, we identified a large number of new pathways (green) as well as previously established ones (orange) and those identified found in a previous study (blue).

B.2. OptCom: A Multi-Level Optimization Framework for the Metabolic Modeling and Analysis of Microbial Communities

The work in this section has been published [106].

B.2.1. Introduction

Solitary species are rarely found in natural environments as most microorganisms tend to function in concert in integrative and interactive units, (i.e., communities). Natural microbial ecosystems drive global biogeochemical cycling of energy and carbon [107] and are involved in applications ranging from production of biofuels [108, 109], biodegradation and natural attenuation of pollutants [110-112], bacterially mediated wastewater treatment [113, 114] and many other biotechnology-related processes [115, 116]. The species within these ecosystems communicate through unidirectional or bidirectional exchange of biochemical cues. The interactions among the participants in a microbial community can be such that one or more population(s) benefit from the association (e.g., through cooperation), some are negatively affected, (e.g., by competing for limiting resources), or more often than not a combination of both. These inter-species interactions and their temporal changes in response to environmental stimuli are known to significantly affect the structure and function of microbial communities and play a pivotal role in species evolution [117-122].

Recent advances in the use of high-throughput sequencing and whole-community analysis techniques such as meta-genomics and meta-transcriptomics promise to revolutionize the availability of genomic information [122-124]. Despite the growing availability of this high-throughput data, we still know very little about the metabolic contributions of individual microbial players within an ecological niche and the extent and directionality of metabolic interactions among them. This calls for development of efficient modeling frameworks to elucidate less understood aspects of metabolism in microbial communities. Spurred by recent advances in reconstruction and analysis of metabolic networks of individual microorganisms, a number of metabolic models of simple microbial consortia have been developed. Efforts in this direction started with the development of metabolic model for a mutualistic two-species microbial community [19]. The metabolic network of each microorganism was treated as a separate compartment in analogy to eukaryotic metabolic models [125, 126]. A third compartment was also added through which the two organisms can interact by exchanging metabolites. The same approach was employed for the metabolic modeling of another syntrophic association between *Clostridium butyricum* and *Methanosarcina mazei* [127]. Lewis *et al* [128] have also described a workflow for large-scale metabolic modeling of interactions between various cell lines in the human brain using compartments to represent different cells. Similarly, Bordbar *et al* [129] developed a multi-tissue type metabolic model for analysis of whole-body systems physiology. Alternatively, others proceeded to identify and model synthetic interactions among different mutants of the same species using genome-scale metabolic models. For example, Tzamali *et al* [130] computationally identified potential communities of non-lethal *E. coli* mutants using a graph-theoretic approach and analyzed them by extending dynamic flux balance analysis model of Varma and Palsson [60]. The same researchers have recently extended their study to describe the co-growth of different *E. coli* mutants on various carbon sources in a batch culture [131]. Wintermute and Silver [132] identified mutualistic relationships in pairs of auxotroph *E. coli* mutants. Each pair was modeled using an extended form of the minimization of metabolic adjustment (MOMA) hypothesis [133]. More recently, the concept of inducing synthetic microbial ecosystems not by genetic modifications but rather with environmental perturbations such as changing the growth medium was introduced [134].

All these studies aimed primarily at modeling communities where one or both species benefit from the association while none is negatively affected. The first study to characterize a negative interaction between two microorganisms using genome-scale metabolic models was published by Zhuang *et al* [135] where similar to [130, 131] an extension of the dynamic flux balance analysis [136] was employed to model the competition between *Rhodoferax ferrireducens* and *Geobacter sulfurreducens* in an anoxic subsurface environment. The same procedure was also employed in a study that characterized the metabolic interactions in a co-culture of *Clostridium acetobutylicum* and *Clostridium cellulolyticum* [137]. A wide range of methods beyond flux balance analysis have been used to model microbial communities [138-149]. For example, Taffs *et al* [150] proposed three different approaches based on elementary mode analysis to model a microbial community containing three interacting guilds. Other studies have drawn from evolutionary game theory, nonlinear dynamics and the theory of stochastic processes to model ecological systems [143, 144, 147].

Despite these efforts, all existing methods for the flux balance analysis of microbial communities are based on optimization problems with a single objective function (related to individual species), which cannot always capture the multi-level nature of decision-making in microbial communities. For example, the flux balance analysis model described in [19] is applicable only to syntrophic associations, where the growth of both species is coupled through the transfer of a key metabolite. The dynamic flux balance analysis models introduced by Zhuang *et al* [135] and Tzamali *et al* [130, 131] rely on solving separate FBA problems for each individual species within each time interval. In all cases these methods cannot trade off the optimization of fitness of individual species versus the fitness function of the entire community. Therefore, there is still a need to develop an efficient modeling procedure to address this issue and to analyze and characterize microbial communities of increasing size with any combination of positive and/or negative interactions.

Here, we introduce OptCom, a comprehensive flux balance analysis framework for microbial communities, which relies on a multi-level optimization description. In contrast to earlier approaches that rely on a single objective function, OptCom's multi-level/objective structure enables properly assessing trade-offs between individual vs. community level fitness criteria. This modeling framework is general enough to capture any type of interactions (positive, negative or combination of both) for any number of species (or guilds) involved. In addition, OptCom is able to explain *in vivo* observations in terms of the levels of optimality of growth for each participant of the community. We first analyze a simple and well-determined microbial community involving a syntrophic association between *D. vulgaris* and *M. maripaludis* [19] to demonstrate the ability of OptCom in recapitulating known interactions. Next, OptCom is employed to model the more complex ecological system of the phototrophic microbial mats of Octopus and Mushroom Springs of Yellowstone National Park and compare our results with those obtained using elementary mode analysis [150]. OptCom identifies the level of sub-optimal growth of one of the guilds (SYN) in this community to benefit other community members and/or the entire population. Finally, we use OptCom to elucidate the extent and direction of inter-species metabolite transfers for a model microbial community [151], identifying the proportion of metabolic resources apportioned to different community members and predicting the relative contribution of hydrogen and ethanol as electron donors in the community. Addition of a new member to this community is also examined in this study.

B.2.2. Methods

OptCom postulates a separate biomass maximization problem for each species as inner problems. The inner problems capture species-level fitness driving forces exemplified through the maximization of individual species' biomass production. If preferable, alternate objective function (e.g., MOMA [133]) could be utilized in the inner stage to represent the cellular fitness

criteria. Inter-species interactions are modeled with appropriate constraints in the outer problem representing the exchange of metabolites among different species. The inner problems are subsequently linked with the outer stage through inter-organism flow constraints and optimality criteria so as a community-level (e.g., overall community biomass) objective function is optimized. Figure B.2.1A schematically illustrates the proposed concept. OptCom is solved using the solution methods previously developed for bilevel programs [21, 88, 152, 153] (see Text S1 for details of the optimization formulation and solution). Note that since OptCom yields a (non-covex) bilinear optimization problem, all case studies presented in this paper were solved using the BARON solver [154], accessed through GAMS, to global optimality.

It is important to note that OptCom can be readily modified to account for the case when one or more organisms show a form of cooperative behavior that benefits the whole population, but comes at the expense of growing at rates slower than the maximum possible [121, 155]. To quantify the deviation of community members from their optimal behavior, we introduce a metric called *optimality level* for each species k (i.e., c^k). The optimality level for each one of the microorganisms is quantified using a variation of OptCom which we refer to as *descriptive*. Descriptive OptCom incorporates all available experimental data for the entire community (e.g., community biomass composition) as constraints in the outer problem and all data related to individual species as constraints in the respective inner problems while allowing the biomass flux of individual species to fall below (or rise above) the maxima ($\mathrm{vopt}^k_{\mathrm{biomass}}$) of the inner problems (see Figure B.2.1B). We note that here the optimum biomass flux for each species ($\mathrm{vopt}^k_{\mathrm{biomass}}$) is community-specific as it is computed in the context of all microorganisms striving to grow at their maximum rate (using the formulation given in Figure B.2.1A). An optimality level of less than one for a microorganism k implies that it grows sub-optimally at a rate equal to $100c^k$ % of the maximum ($\mathrm{vopt}^k_{\mathrm{biomass}}$) to optimize a community-level fitness criterion while matching experimental observations. Alternatively, an optimality level of one implies that microorganism k grows exactly optimally at a rate equal to $\mathrm{vopt}^k_{\mathrm{biomass}}$ whereas a value greater than one indicates that it achieves a higher biomass production level than the community-specific maximum (i.e., super-optimality) by depleting resources from one or more other community members. It is worth noting that super-optimality is achievable for a species only at the expense of sub-optimal behavior of at least one other member in the community. The identified combination of sub- and/or super-optimal behaviors of individual species is driven by the maximization of a community-level criterion (e.g., maximize the total community biomass).

OptCom can capture various types of interactions among members of a microbial community. Symbiotic interactions between two (or more) populations can be such that one or more species benefit from the association (i.e., *positive* interaction), are negatively affected (i.e., *negative* interactions), or combination of both. Mutualism, synergism and commensalism are examples of positive interactions, whereas parasitism and competition are examples of negative interactions. A pictorial representation of how these interactions can be captured within OptCom by appropriately restricting inter-organism metabolic flows is provided in Figure B.2.2 (see Text S1 for implementation details).

B.2.3. Results

Modeling a mutualistic microbial community

We first explore the capability of OptCom to model and analyze a relatively simple and well-characterized syntrophic association between two microorganisms, namely *Desulfovibrio vulgaris* Hildenborough and *Methanococcus maripaludis*. Syntrophy is a mutualistic relationship between two microorganisms, which together degrade an otherwise indigestible organic substrate. A prominent example of syntrophic interactions is interspecies hydrogen transfer, where the

hydrogen produced by one of the species has to be consumed by the other to stimulate the growth of both microorganisms [156-159]. In these communities degradation of a substrate by fermenting bacteria is energetically unfavorable as it carries out a reaction, which is endergonic under standard conditions. However, if this fermenting bacteria is coupled with a hydrogen scavenging partner such as methanogenic bacteria, the organic compound degrading reaction can proceed [160]. Methanogens use hydrogen and energy gained from the first reaction and reduce CO_2 to methane [158, 160].

Here we focus on such a syntrophic association between *Desulfovibrio vulgaris* Hildenborough and *Methano- coccus maripaludis* S2, for which genomes-scale metabolic models as well as experimental growth data for the co-culture are available [19]. With lactate as the sole carbon source and in the absence of a suitable electron acceptor for the sulfate reducer, *M. maripaludis* provides favorable thermodynamic conditions for the growth of *D. vulgaris* by consuming hydrogen and maintaining its partial pressure low. Stoylar *et al* [19] modeled this microbial community as a multi-compartment metabolic network and employed FBA to identify community-level fluxes by maximizing the weighted sum of the biomass fluxes of two microorganisms.

Comparing the OptCom predictions with experimental results

First, we examined whether our model is capable of predicting the relative abundance of species (i.e., composition) in the community by maximizing the community biomass as the outer problem objective function. Each microorganism was allowed to maximize its own biomass yield in the inner problems. Consistent with Stoylar *et al* [19], the lactate uptake rate was set to 48 μM/h and formate and hydrogen accumulation were set to zero, so as all formate and hydrogen produced by *D. vulgaris* is utilized by *M. maripaludis*. Lower and upper bounds on all other reactions (except for the uptake and export fluxes of the shared metabolites) were taken from [19]. The ratio of the biomass yields for *D. vulgaris* and *M. maripaludis* was predicted to be 2.28 based on our simulations. This is consistent with *in vivo* observation that *D. vulgaris* dominates in the co-culture by a ratio of at least 2:1 [19]. Throughout this and the following studies we assume that the biomass flux for each species is proportional to its biomass abundance in the community.

We next explore how well OptCom performs in predicting various metabolic activities across different stages of syntrophic growth. To this end, we applied OptCom for each time interval and compared the model predictions for acetate, methane and carbon dioxide evolution rates as well as total biomass production rates with experimental measurements [19]. A separate run was performed for each time interval where lactate uptake and hydrogen evolution rates were fixed at their experimentally determined values in that interval [19]. The results of this comparison are illustrated in Figure B.2.3. We can see that OptCom predictions are generally in good agreement with experimental data especially for the acetate and methane production rates. The predicted CO_2 evolution rate, however, is lower in all time intervals (except for 62-76 hr) than the measured values. Between 62 hr and 76 hr the experimental data show that the CO_2 evolution rate is close to zero, which may indicate that all CO_2 produced by *D. vulgaris* is consumed by *M. maripaludis* [19]. In addition, OptCom predicts an increase in the biomass production of the whole community over time with increasing lactate uptake rate as expected, although, all of predicted yields are higher than experimental measurements. This inconsistency could be due to missing regulatory information, incorrect modeling of ATP utilization and maintenance energy requirements and/or the presence of futile cycles in the metabolic models of one or both species. It is worth noting that all predictions by Stolyar's multi-compartment approach are also very close to the results obtained by OptCom. This is because in this syntrophic microbial community the growth of both species is coupled and uniquely dependent on the exchange of hydrogen and/or formate. This allows for a single fitness function to describe the behavior of the entire community.

The role of hydrogen and formate in interspecies electron transfer

Hydrogen and formate are primary shuttle compounds for interspecies electron transfer. There are two enzymes in *D. vulgaris* that are involved in production of hydrogen and formate namely pyruvate oxidoreductase and pyruvate-fomrate lyase [19, 161]. While both of these enzymes convert pyruvate to acetyl-CoA, the former produces reduced ferredoxin, which is then used for hydrogen production, whereas the latter produces formate, which can be secreted to the medium. For an uptake rate of 10 μmol/hr, OptCom predicts that a total of 18.6 μmol/ hr of electron transfer in the form of hydrogen and/or formate transfer are required to achieve the maximum growth for both species and community. To investigate the relative contribution of formate and hydrogen in interspecies electron transfer, we examined what portion of the total required electron transfer could be carried by hydrogen or formate while maintaining the maximum biomass yield for both species. This analysis showed that hydrogen could be used as the sole electron carrier to support the maximum growth for both microorganisms even if no formate is secreted by *D. vulgaris*. Formate, on the other hand, could only account for up to 26% (4.9 μmol/ hr) of the total electron transfer to maintain the biomass productions at their maximum. In addition, OptCom results show that formate exchange rates of more than 5.5 μmol/hr (~30%) are not able to support growth for any of the two species. Using OptCom we find that *D. vulgaris* is unable to produce sufficient formate to meet the minimum electron transfer required to maintain the redox balance in the absence of hydrogen.

When hydrogen production by *D. vulgaris* is constrained to be at most 13.7 μmol/hr (i.e., the rest of 4.9 μmol/hr electron transfer is assumed to be carried out by formate if possible), OptCom predictions show that in a co-culture consisting of *D. vulgaris* and a mutant of *M. maripaludis* the growth rate of both *D. vulgaris* and *M. maripaludis* is reduced by 26%. The simulation results also show that no fomrate is produced by *D. vulgaris* in this case, which was expected, as it cannot be consumed by the *M. maripaludis* mutant. Despite no formate production by *D. vulgaris*, OptCom reveals that the flux through pyruvate formate lyase is higher compared to the community having the wild-type strains. Further investigation of the *in silico* flux distributions shows that the entire amount of formate produced by the pyruvate formate lyase reaction is directed towards CO_2 production. This in turn results in an increased consumption of CO_2 by the *M. maripaludis* mutant and consequently a lower accumulation of CO_2 in the extracellular environment compared to the community with the wild-type strains. The predictions by OptCom for the community with mutant of *M. maripaludis* are in agreement with experimental results by Stolyar *et al* [19] who established a syntrophic association between *D. vulgaris* and the *M. maripaludis* mutant MM709 lacking the two annotated formate dehydrogenase enzymes. It was observed that this co-culture is able to grow, confirming that hydrogen alone can support the syntrophic growth of both species. Nevertheless, the growth rate, biomass yield and lactate uptake rates were lower compared to the syntrophic growth between the wild-type strains [19]. Notably, OptCom predictions suggest that if the wild-type *D. vulgaris* in Stolyar's experiment is replaced with a mutant lacking pyruvate-formate lyase, so as all electron equivalent is produced in the form of hydrogen, then the co-culture should be able to restore growth to that of wild-type species community as hydrogen alone can carry all required electron equivalents.

Assessing optimality levels in a phototrophic microbial community

Here we examine the applicability of OptCom for modeling a more complex microbial community containing three interacting guilds, the phototrophic microbial mats of Octopus and Mushroom Springs of Yellowstone National Park (Wyoming, USA) [162]. The inhabitants of this community include unicellular cyanobacteria related to *Synechococcus* spp (SYN), filamentous

anoxygenic phototrophs (FAP) related to *Chloroflexus* and *Roseiflexus* spp and sulfate-reducing bacteria (SRB) as well as other prokaryotes supported by the products of the photosynthetic bacteria [150, 162]. Diel (day-night) variations in metabolic activities of members of this community were observed before [163-165]. During the day when the mat is oxygenated cyanobacteria appear to be the main carbon fixer, consuming CO_2 and producing storage products such as polyglucose as well as O_2 as a by-product of photosynthesis. High levels of O_2 relative to CO_2 stimulate the production of glycolate. Glycolate is then used as a carbon and energy source by other community members such as photoheterotrophic FAP. At night, the mat becomes anoxic and cyanobacteria start to ferment the stored polyglucose into small organic acids such as acetate, propionate and H_2. FAP can incorporate fermentation products photoheterotrophically while SRB oxidizes the fermentation products under anaerobic condition and produces sulfide [162, 166-168]. A schematic diagram representing the interactions in this community is given in [150].

This microbial community has been previously modeled and analyzed by Taffs *et al* [150] using a representative microorganism for each guild: Oxygenic photoautotrophs related to *Synechococcus* spp were chosen to represent the mat's primary carbon and nitrogen fixers. FAP from the family *Chloroflexaceae*, were selected to represent metabolically versatile photoheterotrophs that capture light energy as phosphodiester bonds but require external reducing equivalents and carbon sources other than CO_2. A SRB guild representative whose metabolic behavior was based on several well-studied sulfate-reducing bacteria was also included in the community model description [150]. The metabolic networks representing central carbon and energy metabolism for each guild were then constructed and three different modeling approaches based on the elementary mode analysis were employed to elucidate sustainable physiological properties of this community [150]. Here, we focus only on daylight metabolism (for which more experimental data is available) to assess the efficacy of OptCom in describing carbon and energy flows and the biomass ratio between guilds.

Analysis of the daylight metabolism

The relative abundance of various species in a microbial community (i.e., composition) is of significant ecological importance. The ratio of cyanobacterial (SYN) to FAP biovolumes in a Mushroom Spring mat was determined experimentally to be 1.6:1 [169]. It was assumed that biomass formation rates and biovolume of species in the community are directly related [150]. In another study the biomass ratio in the top 1 mm of Octopus and Mushroom Spring mats was estimated to range from 1.5:1 to 3.5:1 based on the relative abundances of metagenomic reads [150]. We used OptCom to model this community postulating that each guild strives to maximize its biomass and examined if the biomass ratio of SYN/FAP can be correctly predicted. We chose as the outer problem objective function to maximize the total community biomass (i.e., SYN biomass + FAP biomass + SRB biomass). During the day O_2 competes with CO_2 for the rubisco active site, leading to production of glycolate (O_2 + ribulose − 5 − P + ATP → glycolate + triose phophate + ADP) instead of additional reduced carbon (CO_2 + ribulose−5−P + ATP → 2 triose phophate + ADP) [150]. The flux ratio of these two reactions (O_2/CO_2) was measured for the Octopus and Mushroom Spring microbial mats and reported to vary approximately between 0.03 and 0.07 [150, 170]. We incorporated this information into our modeling framework by fixing the flux ratio of these reactions at different values between 0.03 and 0.07 (using a constraint in the inner problem of SYN). Lower and upper bounds on all reactions (except for the uptake and export fluxes of the shared metabolites) were taken from [150]. Under these conditions, the SYN/FAP biomass ratio was predicted to range from 7.94 (for O_2/CO_2 flux ratio = 0.07) to 20.26 (O_2/CO_2 flux ratio = 0.03), which are significantly higher than the experimentally determined values of 1.5 to 3.5. This suggests that the reason for the discrepancy in prediction may be that the SYN guild does not maximize its biomass. Therefore, we decided to test this hypothesis by using the descriptive mode of the OptCom procedure (see Figure B.2.1B) and establish the

optimality level of SYN and other members of this community. To this end, we added a constraint to the outer problem to fix the SYN/FAP biomass ratio at different values in the experimentally observed range (1.5 to 3.5). The objective function of the outer problem was assumed to be maximization of the total community biomass. We determined the optimality levels across different values of SYN/FAP biomass and O_2/CO_2 flux ratios in their experimentally determined ranges (see Figure B.2.4). OptCom finds that the observed SYN/FAP biomass ratios are consistent with SYN guild growing sub-optimally at 61-82% of its community-specific maximum with lower values corresponding to higher O_2/CO_2 flux ratios (see Figure B.2.4A). On the other hand, FAP guild appears to benefit from this sub-optimal behavior of SYN by growing at rates, which are approximately 4.5 to 8.5 times higher than its community-specific maximum (see Figure B.2.4B).

SYN grows sub-optimally in this community to benefit other community members (e.g., FAP) and optimize a community-level fitness criterion (e.g., maximize the total community biomass). We investigated the effect of sub-optimal growth of the SYN guild on the total community biomass production across different values of SYN/FAP biomass and O_2/CO_2 flux ratios (see Figure B.2.4C). As illustrated in Figure B.2.4C, at higher O_2/CO_2 flux ratios, the total community biomass is higher compared to the case when SYN grows optimally. The metabolic reason for this lower growth of SYN is that fixing more carbon (manifested by 3-7 times more predicted glycolate and acetate production) to supply other guilds and increase the overall community biomass imposes extra energy demands on the SYN guild. In contrast, for low O_2/CO_2 flux ratios the maximum community biomass when SYN grows sub-optimally is lower compared with when it grows optimally (i.e., both dashed lines lie below the solid line in Figure B.2.4C). A possible reason for this discrepancy is that the experimental measurements for SYN/FAP biomass ratio were performed when the O_2/CO_2 flux ratio was high. This could also be a consequence of the experimental underestimation of glycolate production due to consumption of radio-labeled photosynthate during incubation as stated in [150]. Alternatively, SYN may grow sub-optimally so that it can divert some resources towards polysaccharide production to fuel night-time maintenance energy and morning nitrogen fixation. This is another type of a cooperative behavior by SYN.

Notably, two different cases were considered by Taffs *et al* [150] using the elementary modes and compartmentalized approach: a selfish criterion where each guild attempts to maximize its own biomass and an altruistic criterion where the guilds strive to maximize the total community biomass. It was concluded that predictions using the first criterion are in better agreement with experimental data. OptCom, on the other hand reveals that a trade-off between these two criteria appears to be driving the metabolism in this community. While some guilds strive to maximize their own growth, others (e.g., SYN) grow sub-optimally to maximize the biomass of entire community or benefit the nighttime metabolism, or a combination of both, depending on O_2/CO_2 flux ratio and environmental conditions.

Elucidating trophic and electron accepting interactions in sub-surface anaerobic environments

In a recent study, Miller *et al* [151] established a model microbial community to better understand the trophic interactions in sub-surface anaerobic environments. This community was composed of three species including *Clostridium cellulolyticum*, *Desulfovibrio vulgaris* Hildenborough, and *Geobacter sulfurreducens*. Cellobiose was provided as the sole carbon and energy source for *C. cellulolyticum* whereas the growth of *D. vulgaris* and *G. sulfurreducens* were dependent on the fermentation by-products produced by *C. cellulolyticum*. *D. vulgaris* and *G. sulfurreducens* were supplemented with sulfate and fumarate, respectively, as electron-acceptors to avoid electron acceptor competition [151]. The experimental measurements for the biomass composition of the community showed that, as expected, *C. cellulolyticum* was the dominant member in the co-

culture and confirmed the presence of *D. vulgaris* and *G. sulfurreducens*. It was, however, not possible to quantify experimentally the flow of shared metabolites among the community members as their concentrations were below the detection limits. Therefore, the authors proposed an approximate model of the carbon and electron flow based on some measurements of the three species community at steady-state, pure culture chemostat experiments and data from the literature [151].

Here, we model this microbial community by making use of the corresponding bacterial metabolic models and employ OptCom to elucidate the inter-species interactions. The metabolic models of *C. cellulolyticum* (i.e., *i*FS431) and *G. sulfurreducens* were reconstructed by Salimi *et al* [137] and Mahadevan *et al* [171], respectively. A basic metabolic model of *D. vulgaris* containing 86 reactions was introduced by Stolyar et al [19], however, this model had only a compact representation of the central metabolism. For example, the model was not able to support growth in the presence of acetate or ethanol as the sole carbon source. Therefore, we expanded this model by adding new reactions from a first draft reconstructed model in the Model Seed [54] and the KEGG database [86] using the GrowMatch procedure [152] (see Text S1 for details). The updated model of *D. vulgaris* consists of 145 reactions and is capable of supporting growth on acetate as well as ethanol. This model is available in the supplementary material (Table S1).

Fumarate consumption by G. sulfurreducens

FBA simulations showed that the metabolic model for *G. sulfurreducens* [171] is not able to capture the experimental observation that the amount of fumarate consumed is higher than the amount of succinate produced. In addition, the model predicts that no malate is produced under the examined conditions. An inspection of the metabolic model of *G. sulfurreducens* revealed that the only included uptake pathway for fumarate is through mutual dicarboxylic acid transporter (fumarate[e] + succinate[c] $\leftrightarrow$ fumarate[c] + succinate[e]) implying that the amount of succinate produced must be equal to the amount of fumarate consumed. Interestingly, in support of the observations by Miller *et al* [151], a recent study [172] has confirmed that the fumarate consumption rate by *G. sulfurreducens* is higher than the succinate production rate and demonstrated using ^{13}C-based metabolic flux analysis that fumarate can be used as an additional carbon source through the TCA cycle where it is converted to malate by fumarase, and oxaloacetate via malate dehydrogenase. These findings suggest that the *dcu* gene family (responsible for the uptake of dicarboxylates such as fumarate) in *G. sulfurreducens* may have a dual function, i.e., they can act both mutually (with exchange of another compound such as succinate) or independently (i.e., protonated), similarly to those in *E. coli* [173]. This was verified by performing a bi-directional BLAST analysis that revealed high sequence similarity between the *dcu* gene families in *G. sulfurreducens* and *E. coli*. It is worth noting that addition of an alternative succinate transporter to the model could also have been another way of explaining the experimental data, however this hypothesis was not supported by the BLAST analysis. Therefore, in the absence of any other experimental data, we decided to add a protonated transport reaction for fumarate to the model. In our simulations we restricted the flux of this reaction to 15.5% of the fumarate transfer by dicaboxylic acid transporter based on the metabolic flux data under electron acceptor limited conditions [172].

Uncovering the inter-species metabolite transfers in the community

While the relative molar abundance of each species was measured experimentally by Miller *et al* [151], the metabolite flows across community members were untraceable. We thus chose to use OptCom to gain insight into inter-species metabolite trafficking. To this end, we employed the descriptive mode of OptCom (see Figure B.2.1B) first to establish the optimality levels of species participating in this community, by fixing the biomass composition of the community at the values obtained experimentally by adding constraints to the outer problem. The objective function

of the outer problem was maximization of the total community biomass. Descriptive OptCom revealed that the experimentally determined biomass composition in this community was consistent with optimal growth for all microorganisms (i.e., optimality level of one for all species involved). Upon verifying that biomass maximization was driving metabolism in this community, we used OptCom to make predictions about inter-organism flow rates with a basis of 1 mole/gDW.hr of cellobiose uptake by *C. cellulolyticum* so that we can directly compare our results with the estimates in Miller *et al* [151]. The lower bound and upper bounds on all reactions (except for the uptake and export fluxes of the shared metabolites) were taken from the publications of the respective metabolic models [19, 137, 171]. Because *D. vulgaris* has a much more efficient enzymatic process for hydrogen consumption than *G. sulfurreducens*, we initially allowed *G. sulfurredcens* to take up only a small portion (between 1 to 10%) of the total hydrogen produced by *C. cellulolyticum.* However, the total predicted acetate and CO_2 accumulation in the extracellular environment deviated significantly from the experimental observations by Miller *et al* [151]. Therefore, we decided to perform the remaining simulations assuming that *D. vulgaris* consumes all hydrogen produced by *C. cellulolyticum* (even though this may not be the only way of reconciling model predictions and the experimental data). OptCom found that under these conditions 1 mol/gDW.hr of cellobiose leads to 2.48 moles/gDW.hr of acetate and 3.22 moles/gDW.hr of CO_2 in the extracellular environment which agree well with 2.7 and 3.3 moles/gDW.hr of acetate and CO_2, respectively, observed in the supernatant of the bioreactor (per mole of cellobiose) by Miller *et al* [151]. We note, however, that the predicted level of acetate production by *C. cellulolyticum* metabolic model (1.65 mol/gDW.hr) is lower than what was estimated in Miller's model (2.9 mol/gDW.hr). In general, however, the predicted allocation of metabolic resources to different members of the community by OptCom is in good agreements with estimations in Miller [151] (see Figure B.2.5). For example, OptCom suggests that about 13% of the acetate produced by *C. cellulolyticum* is directed towards *G. sulfurreducens,* which is very close to the 15.5% value estimated in [151].

OptCom results also show that hydrogen and ethanol produced by *C. cellulolyticum* can be completely utilized by *D. vulgaris* to reduce sulfate to hydrogen sulfide. A rough estimate for the ratio of hydrogen to ethanol, which serve as electron donors for *D. vulgaris*, is given in by Miller *et al* [151] (H_2/Ethanol = 20) based on the pure culture data under similar conditions. The simulations with OptCom using genome-scale metabolic models of the community members, however, indicate a much higher contribution of ethanol in inter-species electron transfer (H_2/Ethanol = 2.34). We performed a flux variability analysis to see if this ratio can change under the examined condition, while maintaining the maximum community biomass, but no changes in this ratio were possible. This suggests that under the observed experimental condition, a H_2/Ethanol ratio of 2.34 is needed to support the maximum growth for each species as well as for the community as a whole. While acetate serves as the only carbon substrate for both *G. sulfurreducens* and *D. vulgaris,* it was not possible to determine experimentally if *D. vulgaris* directly uses the available acetate in the medium released by *C. cellulolyticum* or it derives acetate from ethanol. OptCom results support the latter scenario (see Figure B.2.5). This is more likely to happen because acetate is already available internally to *D. vulgaris* from the cytosolic oxidation of ethanol. OptCom also identifies that 77.6% of the converted ethanol to acetate is secreted to the medium by *D. vulgaris*, while the rest is incorporated into biomass (see Figure B.2.5). This is in good agreement with the estimate by Miller *et al* [151] suggesting that *D. vulgaris* does not consume any acetate produced by *C. cellulolyticum* and that it exports 62.5% of the assimilated ethanol to the medium as acetate. Elucidation of the metabolic interactions among the members of this community was achieved by OptCom after verifying that all species appear to grow optimally based on the *in vivo* observations for the community biomass composition.

Addition of a new member to the microbial community

As mentioned earlier, 2.48 moles/gDW.hr of acetate was predicted to be available in the extracellular environment (per mole of cellobiose consumed) which could be utilized by other trophic anaerobic bacteria [151]. Therefore, an acetate utilizing methanogen such as *Methanosarcina* species, which are known to be avid consumers of acetate, can be envisioned as an additional member of this community. We chose *Methanosarcina barkeri* for this analysis as its metabolic model has been reconstructed by Feist *et al* [74]. Another inner problem was added to the OptCom to account for addition of *M. barkeri* to this community. Consistent with other community members the objective function for this inner problem was to maximize the biomass flux of *M. barkeri*, whereas the objective function of the outer problem was to maximize the total community biomass. The acetate uptake rates by *G. sulfurreducens* and *D. vulgaris* were fixed at the values obtained by OptCom for the tri-culture. *D. vulgaris* and *M. barkeri* were suggested to compete in anoxic environments for hydrogen [174], however, we assumed that all H_2 produced by *C. cellulolyticum* is consumed by *D. vulgaris*, as it has been reported to have much more favorable kinetic parameters for H_2 metabolism than methanogens [175-177]. In addition, it was demonstrated that *Methanosarcina* species can not only consume but also produce hydrogen when growing on organic substrates such as acetate [178, 179]. Therefore, we allowed *D. vulgaris* to consume the hydrogen produced by *M. barkeri* (if any) in addition to that produced by *C. cellulolyticum*.

The biomass flux of *M. barkeri* is strongly dependent on the value of growth-associated maintenance (GAM), which was found to be a function of the proton translocation efficiency of the Ech hydrogenase reaction [74]. The range of GAM values for 0.2-2 protons translocated/$2e^-$ that result in a growth yield consistent with *in vivo* observations was computed by Feist *et al* [74]. Here, we examined the variability in growth yields and relative abundance of *M. barkeri* in the tetra-culture community across different GAM values associated with 0.2-2 protons translocated/$2e^-$. This analysis showed that *M. barkeri* is capable of consuming the entire 2.48 moles of acetate produced by *C. cellulolyticum* and *D. vulgari.* Depending on the GAM value and the proton translocation efficiency, *M. barkeri* was predicted to constitute 2.5 to 10.4% of the total community biomass (assuming that the biomass fluxes are proportional directly with the abundance levels of species in the community) with the other three members growing at rates similar to the ones obtained for the tri-culture. *C. cellulolyticum* still dominates the co-culture as before with biomass fractions ranging from 69.6 to 75.7% (depending on *M. barkeri*'s biomass flux). The methane evolution rate by *M. barkeri* was predicted by OptCom to range from 2.36 to 2.45 moles/gDM.hr. It is important to note that previous studies have reported that the internal carbon and electron flow of *M. barkeri* could be altered by *D. vulgaris* in a co-culture grown on an organic substrate such as acetate, [180]: It was suggested that *D. vulgaris* strives to keep the partial pressure of hydrogen low enough to shift the catabolic redox system of methanogen so that more H_2 is produced by *M. barkeri* (compared to pure cultures) and more acetate is oxidized to CO_2 instead of methane [180]. Even though we allowed *D. vulgaris* to take up all hydrogen produced by *M. barkeri* (in addition to that produced by *C. cellulolyticum*), no such shift in methanogenesis was observed for the tetra-culture according to the OptCom predictions. A possible reason might be that enough hydrogen (as well as ethanol) is already available to *D. vulgaris* from *C. cellulolyticum*, obviating the need to alter methanogenesis in order to gain the reducing equivalents. This hypothesis is supported by the experimental observation that if excess H_2 is added to the co-culture of *M. barkeri* and *D. vulgaris*, it is completely consumed by *D. vulgaris* and the acetate catabolism by *M. barkeri* is no longer affected [180].

Even though 3.22 moles/gDW.hr of CO_2 produced by *C. cellulolyticum* and *G. sulfurreducens* is available in the medium, OptCom predicts that it remains completely unused in the tetra-culture. This was expected as growth of *M. barkeri* on CO_2 relies on presence of hydrogen, which we assumed that it was consumed completely by *D. valgaris*. In order to examine if *M. barkeri* is indeed capable of utilizing the available CO_2 as a carbon source (in

addition to acetate), we temporarily allowed *M. barkeri* to take up the hydrogen produced by *C. cellulolyticum*. For this case, OptCom revealed that if the entire hydrogen produced by *C. cellulolyticum* is available to *M. barkeri*, it can support growth on CO_2 only for proton translocation efficiencies of less than one/$2e^-$. Notably, for proton translocation efficiencies of more than one, even though no CO_2 is assimilated by *M. barkeri*, OptCom shows that the availability of hydrogen will lead to an increase in the methane production by about 26-28%.

B.2.4. Discussion

Here, we introduced OptCom, a comprehensive computational framework for the flux balance analysis of microbial communities using genome-scale metabolic models. We demonstrated that OptCom can be used for assessing the optimality level of growth for different members in a microbial community (i.e., Descriptive mode) and subsequently making predictions regarding metabolic trafficking (i.e., Predictive mode) given the identified optimality levels. Unlike earlier FBA-based modeling approaches that rely on a single objective function to describe the entire community [19, 134] or separate FBA problems for each microorganism [130, 131, 135, 137], OptCom integrates both species- and community-level fitness criteria into a multi-level/objective framework. This multi-level description allows for properly quantifying the trade-offs between selfish and altruistic driving forces in a microbial ecosystem. Species and community level fitness functions are quantified by maximizing the biomass formation for the respective entity. We note, however, that the physiology of microbial communities is highly context and environment dependent and a universal community-specific fitness criterion does not exist. Studies similar to those conducted for mono-cultures that examine and compare various presumed hypotheses on cellular objective function [181-186] or algorithms that identify/test a relevant objective function using experimental flux data [187, 188] are needed in the context of multi-species systems.

An important goal of studying natural and synthetic microbial communities is their targeted manipulation towards important biotechnological goals (e.g., cellulose degradation, ethanol production, etc.). This has motivated researchers to construct simple synthetic microbial ecosystems, which are amenable to genetic and engineering interventions, for biotechnology- and bioenergy-related applications. As an example, Bizukojc *et al* [127], have proposed a co-culture composed of *Clostridium butyricum* and *Methanosarcina mazei* to relieve the inhibition of fermentation products and increase production of 1,3-propanediol (PDO) by *Clostridium butyricum*. Mixed cultures have been also established for overproduction of polyhydroxyalkanoates (PHA) [189, 190] and ethanol [191-195]. For example, *Clostridium thermocellum*, which is used for ethanol production, has been found to be capable of utilizing hexoses, but not pentose sugars generated from breakdown of cellulose and hemicellulose [195]. Therefore, cultivation of *C. thermocellum* with other thermophilic anaerobic bacteria capable of utilizing hexoses as well as pentose to produce ethanol (e.g., *Clostridium thermosaccharolyticum* and *Thermoanaerobacter ethanolicus*) has been previously examined *in vivo* [191-195]. The multi-objective and multi-level structure of the OptCom procedure, introduced here, can help assess the metabolic capabilities of such synthetic ecosystems. Taking a step further, OptCom can be readily modified to identify the minimal number of direct interventions (i.e., knock-up/down/outs) to the community leading to the elevated production of a desired compound (e.g., by considering the overproduction of desired compound as the outer problem objective function), thus extending the applicability of strain design tools such as OptKnock [21], OptStrain [88], OptReg [24] and OptForce [196]. It is worth noting that a key bottleneck to the modeling and analysis of microbial communities is the paucity of genome-scale models for all participants in a complex microbial community. Overcoming this barrier would require the development of high-throughput metabolic reconstruction tools such as the Model Seed [54] resource. Given that microbial communities change with time (e.g., day/night cycle) and also location (e.g., nutrient gradients), approaches that would be able to capture temporal and spatial varying inter-species

metabolic interactions are needed. For example, the separate FBA problems for each individual species in the dynamic flux balance analysis methods of Zhuang *et al* [135] and Tzamali *et al* [130, 131] can be integrated with OptCom to account for inter-species interactions and community-level fitness driving forces within each time interval.

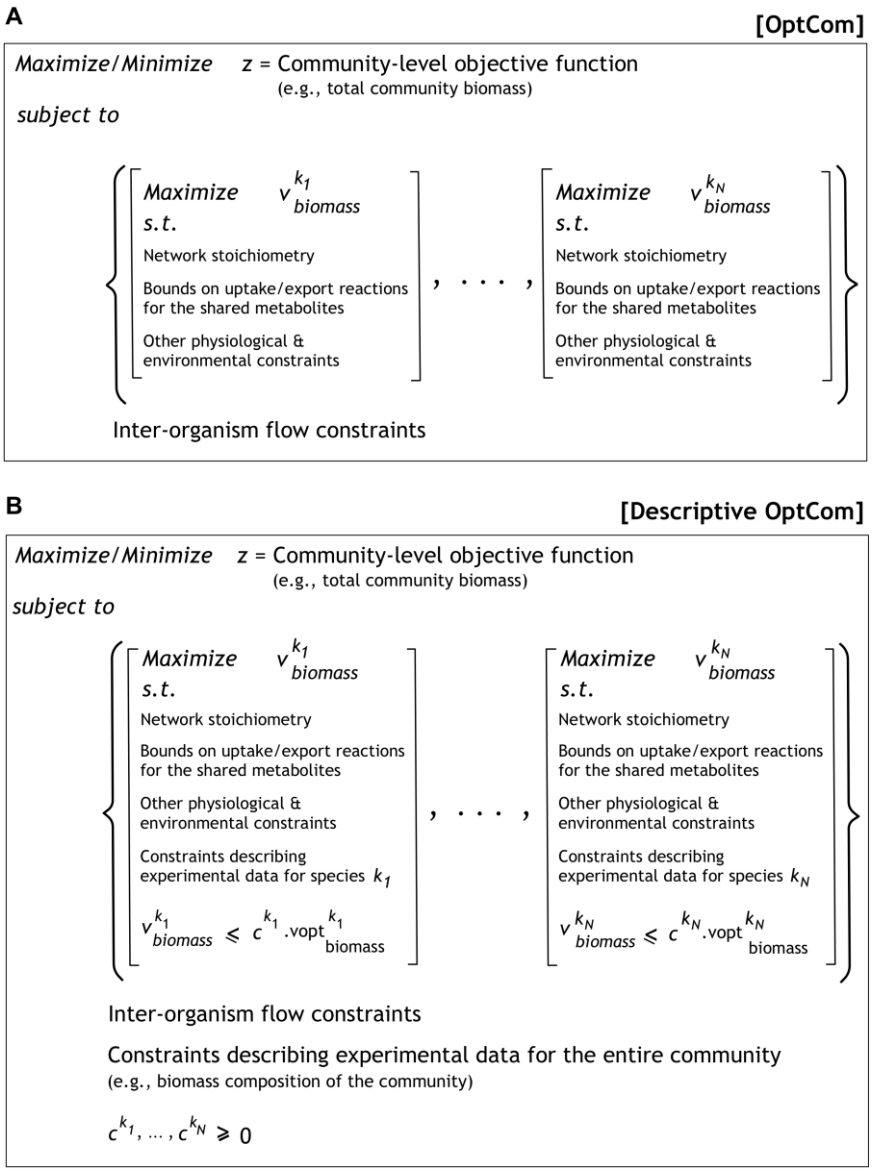

Figure B.2.1. Schematic illustration of OptCom. (A) The multi-level optimization structure of the OptCom. A separate biomass maximization problem is defined for each species as inner problems. These inner problems are then integrated in the outer stage through the inter-organism flow constraint to optimize a community-level objective function. (B) Structure of the Descriptive OptCom to determine the optimality level of each species (c^k), given a set of experimental data. The available experimental data for the entire community and the individual species are described using constraints in the outer and inner problems, respectively, whereas, sub- or super-optimal behavior of each microorganism is captured by using a constraint for the respective inner problem.

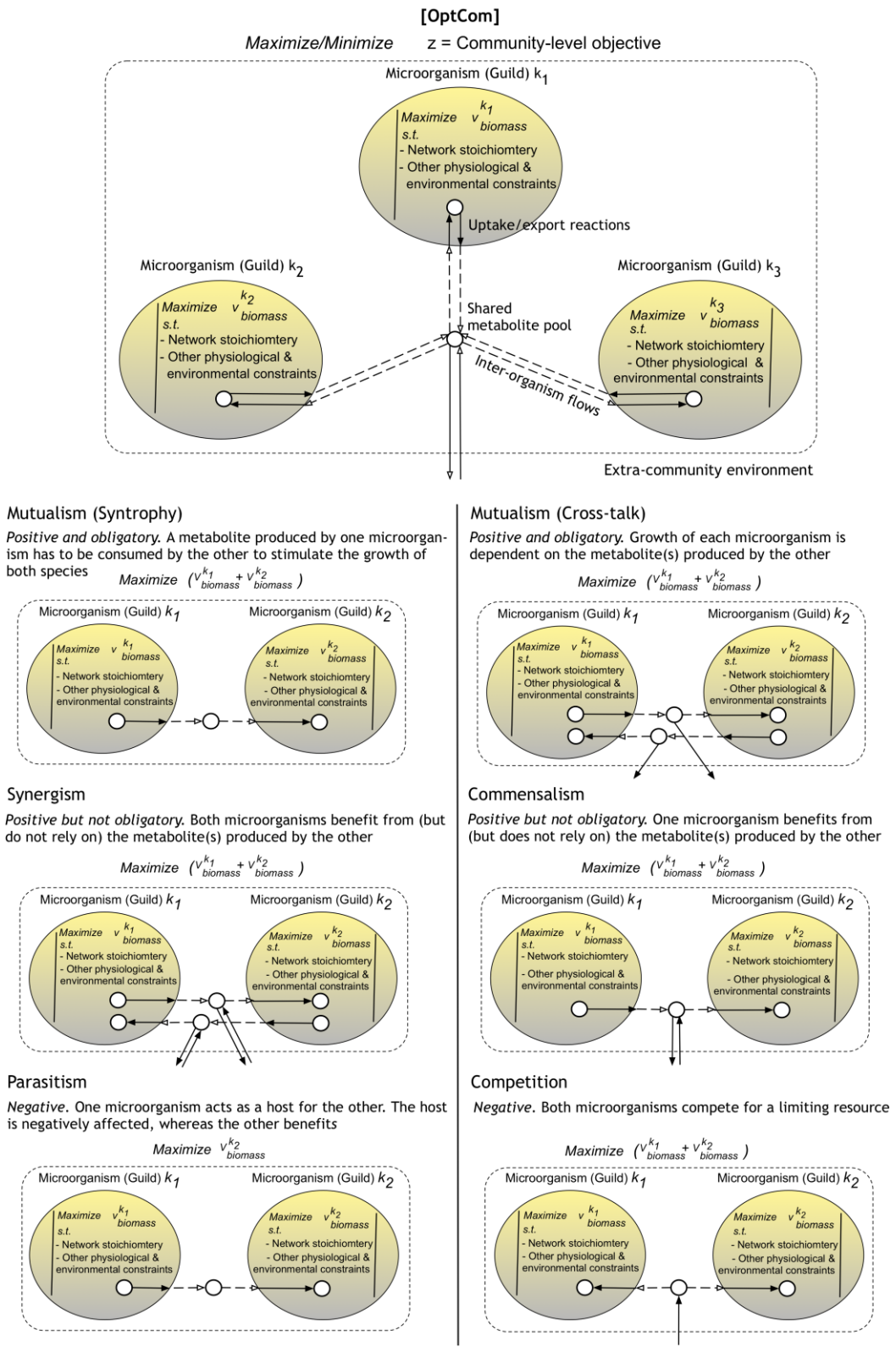

Figure B.2.2. Pictorial illustration of the customized OptCom for various types of interactions. OptCom (top panel) can be readily customized for each type of interaction through properly adjusting the inter-organism flow constraints as demonstrated for a typical microbial community composed of two interacting members.

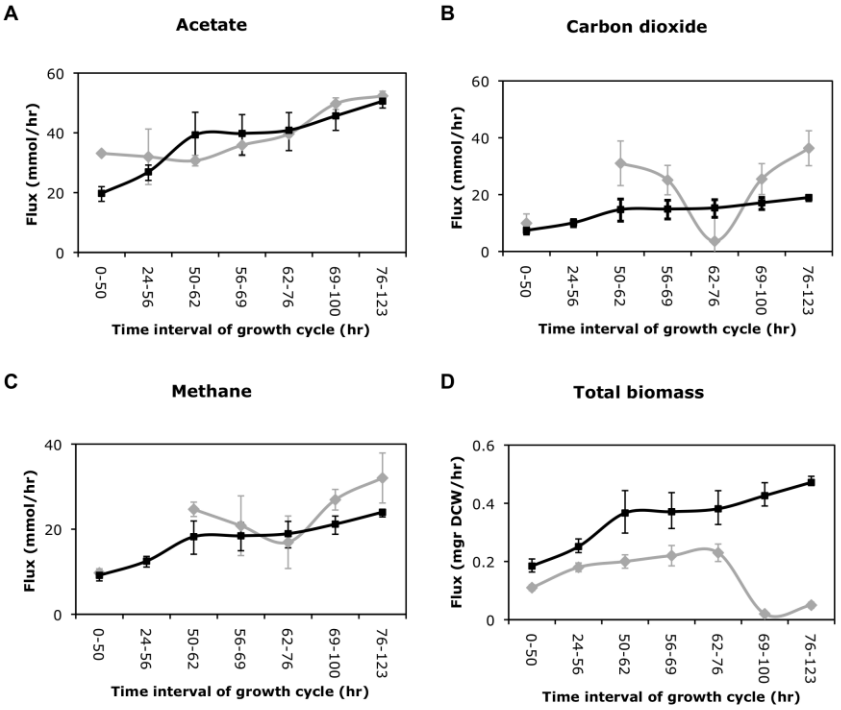

Figure B.2.3 Comparison of the predicted metabolic activities during the syntrophic growth with experimental data. Experimentally determined (gray diamond) and predicted production fluxes by OptCom (black square) for (A) acetate, (B) carbon dioxide (C) methane and (D) total community biomass in the syntrophic growth of *D. vulgaris* and *M. maripaludis*. All experimental data were obtained through personal communications with authors of [19]. A separate simulation was performed for each time interval wherein lactate uptake and hydrogen evolution rates were fixed at their experimentally determined values for that interval. Error bars for experimental values indicate the bounds of 95% confidence intervals [19]. The error bars for OptCom predictions were calculated by performing the simulations on the upper and lower bounds of the 95% confidence intervals for measured lactate and hydrogen flux rates.

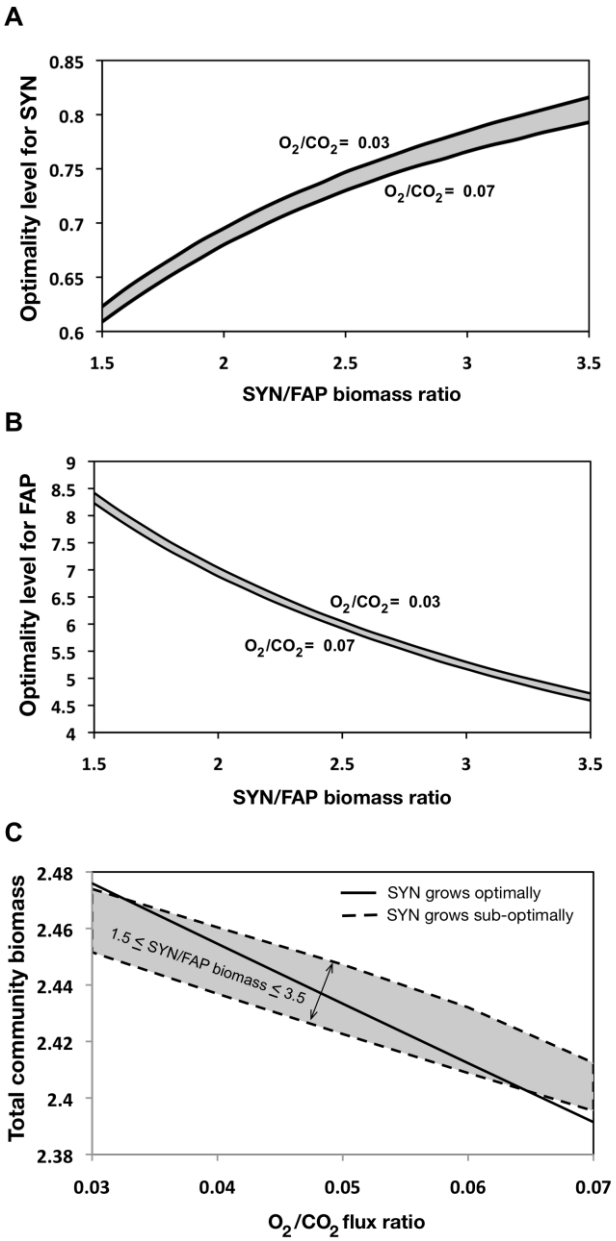

Figure B.2.4. Optimality levels for the SYN and FAP guilds and their effect on the total community biomass. Optimality levels for (A) SYN and (B) FAP as a function of the SYN/FAP biomass ratio across different values of the O_2/CO_2 flux ratio (C) Comparison of the predicted total community biomass (1/h) for the case when SYN grows sub-optimally and when it grows optimally. Note that, to compute the total community biomass when SYN grows optimally only O_2/CO_2 flux ratio was fixed at values in the experimentally determined range (i.e., 0.03 to 0.07), whereas for all other cases, in addition to O_2/CO_2 flux ratio, SYN/FAP biomass ratio was also fixed at values measured experimentally (i.e., 1.5 to 3.5). Lower and upper dashed lines in (C) represent the maximum and minimum predicted community biomass (when SYN grows sub-optimally) across various SYN/FAP biomass ratios.

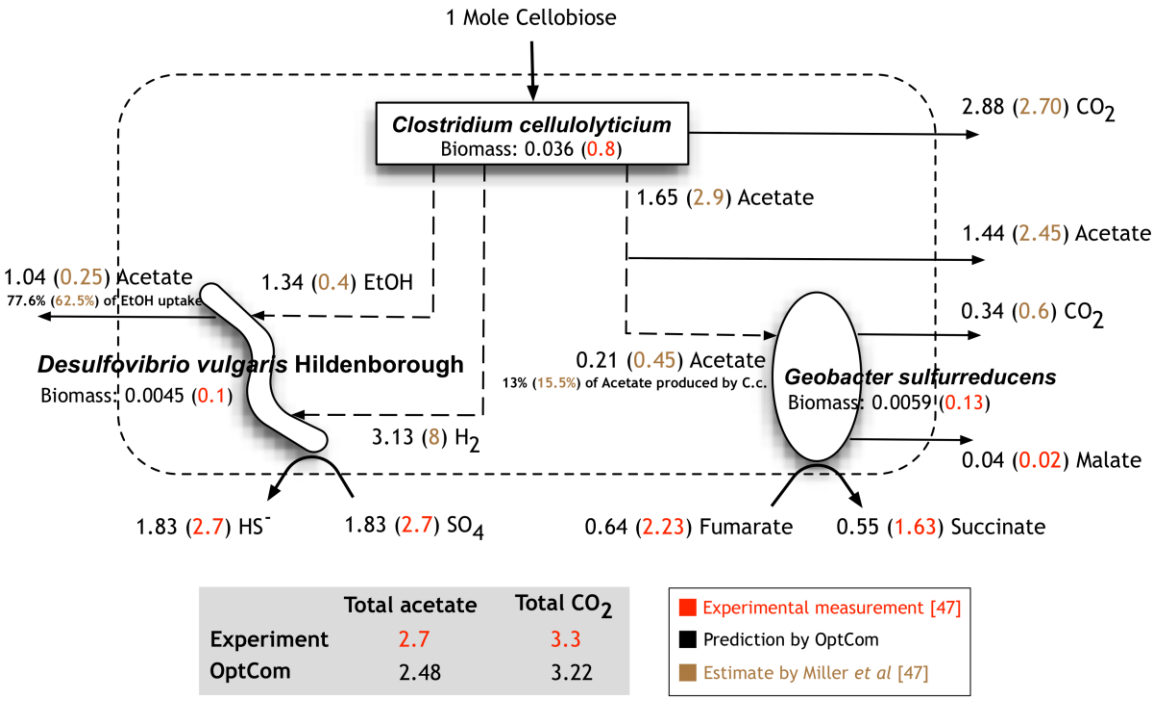

Figure B.2.5. Comparison of the predicted fluxes by OptCom with estimates in the proposed model of [151]. The total predicted acetate and CO_2 production rates by OptCom are in good agreement with experimental measurements by Miller *et al* [151]. Note that it was not possible to determine experimentally how much of the total acetate or CO_2 available in the supernatant of the bioreactor is produced by which microorganism (the values provided by Miller *et al* [151] for the acetate and CO_2 production by each species as well as all inter-organism flow rates are estimates and not experimental measurements). The values associated with the biomass of each microorganism represent fluxes (1/h) for OptCom predictions and concentrations (M) for experimental measurements [151].

B.3. *Zea mays* iRS1563: A comprehensive genome-scale metabolic reconstruction of maize metabolism

The work in this section has been published [197].

B.3.1 Introduction

Zea mays, commonly known as maize or corn, is a plant organism of paramount importance as a food crop, biofuel production platform and a model for studying plant genetics [198]. Maize accounts for 31% of the world production of cereals occupying almost one-fifth of the worldwide land dedicated for cereal production [199]. Maize cultivation led to 12 billion bushels of grain in the USA alone in 2008 worth $47 billion [200]. Maize is the second largest crop, after soybean, used for biotech applications [199]. In addition to its importance as a food crop, 3.4 billion gallons of ethanol was produced from maize in 2004 [200]. Maize derived ethanol accounts for 99% of all biofuels produced in the United States [200]. However, currently nearly all of this bioethanol is produced from corn seed [201]. Ongoing efforts are focused on developing and commercializing technologies that will allow for the efficient utilization of plant fiber or cellulosic materials (e.g. maize stover and cereal straws) for biofuel production. Maize is the most studied species among all grasses with respect to cell wall lignification and digestibility, which are critical for the efficient production of cellulosic biofuels [202]. A thorough evaluation of the metabolic capabilities of maize would be an important resource to address challenges associated with its dual role as a food (e.g., starch storage) and biofuel crop (e.g., cell wall deconstruction).

This decade we witnessed significant advancements towards mapping plant genes to metabolic functions culminating with the complete genome sequencing and partial annotation of a number of plant species, namely, *Arabidopsis thaliana* [203], *Oryza Sativa* [204, 205], *Sorghum bicolor* [206], *Zea mays* [207] and *Theobroma cacao* [208]. Nevertheless, attempts to engineer plant metabolism for desired overproductions have been met with only limited success [209]. Genetic modifications seldom bring about the expected/desired effect in plant metabolism primarily due to the built-in metabolic redundancy circumventing the imposed genetic changes [210, 211]. This necessitates the development of genome-wide comprehensive metabolic reconstructions capable of taking account of the complete inventory of metabolic transformations of a given plant organism.

Genome-scale metabolic reconstructions are available for an increasing number of organisms [212, 213]. At least 40 bacterial, 2 archaeal and 15 eukaryotic reconstructions are available to-date [209, 212, 214, 215] while many others are under development. Recently Poolman et al (2009) and Dal'Molin et al (2010) independently constructed the first two genome-scale metabolic reconstructions for a plant organism (i.e., *Arabidopsis thaliana)*. The model by Dal'Molin *et al* identifies the set of essential reactions, accounts for the classical photorespiratory cycle and highlights the significant differences between photosynthetic and non-photosynthetic metabolism. The model by Poolman *et al* includes ATP demand constraints for biomass production and maintenance and suggests strategies for the construction of metabolic modules as a consequence of variation in ATP requirement. Both models make a significant step forward towards assessing the metabolic capabilities of plants establishing production routes for key biomass precursors and major pathways of Arabidopsis primary metabolism. In addition, two recent efforts involved the reconstruction of plant models with an emphasis on specific physiological conditions or tissue types [216, 217]. Model C4GEM [217] focused on C4 plants such as maize, sugarcane and sorghum and investigated flux distributions in mesophyll and bundle sheath cells during C4 photosynthesis. Grafahrend-Belau *et al* developed a metabolic network of only primary metabolism in barley seeds and studied grain yield and metabolic fluxes under a variety of oxygen availability scenarios and genetic manipulations [216]. Pilalis *et al.* reconstructed a multi-compartmental model of the central metabolism of *Brassica napus* (Rapeseed) and simulated seed growth during the stage of oil accumulation and subsequently studied network properties of seed metabolism via Flux Balance Analysis, Principal Component Analysis and reaction deletion studies [218].

In this section, we describe the construction of a genome-scale in silico model of maize metabolism (i.e., *Zea mays i*RS1563). This is, to the best of our knowledge, the first attempt of globally characterizing the metabolic capabilities (both primary and secondary metabolism) using a compartmentalized photosynthetic model of an important crop and energy plant species. The development of a genome-scale model for maize is a significant challenge due to its genome size which is 14 times larger [207] than that of *Arabidopsis thaliana* (157 million base pairs) [219]. The constructed model contains 1,563 genes and 1,825 metabolites participating in 1,985 reactions from both primary and secondary metabolism of maize. For 42% of the reaction entries direct literature evidence in addition to homology criteria for their inclusion to the model was identified. We found that as many as 676 reactions and 441 metabolites are unique to *Zea mays i*RS1563 in comparison to the AraGEM model by Dal'Molin *et al*. We chose the AraGEM model as a basis of comparisons as at the onset of this study it was the most comprehensive genome-scale compartmentalized model of a plant species capable of recapitulating basic plant physiological states. In order to deduce the genuine differences between maize and Arabidopsis irrespective of annotation chronology we also reconstructed an up-to-date model of Arabidopsis, *A. thaliana i*RS1597. *A. thaliana i*RS1597 contains 1597 genes, 1798 reactions and 1820 metabolites. In comparison to *A. thaliana i*RS1597, *Zea mays i*RS1563 has 445 new reactions and 369 new metabolites. Notably, 893 reactions and 674 metabolites are included in *Zea mays*

*i*RS1563 that are absent from the maize C4GEM model. All reactions present in Zea mays iRS1563 are elementally and charged balanced and localized into six compartments including cytoplasm, mitochondrion, plastid, peroxisome, vacuole and extracellular space. Provisions for accounting that photosynthesis in maize (i.e., a C4 plant) occurs in two separate cell types (i.e., mesophyll cell and bundle sheath cell) are included in the model. GPR associations are delineated from the available functional annotation information and homology prediction accounting for monofunctional, multifunctional and multimeric proteins, isozymes and protein complexes. A biomass equation is established that quantifies the relative abundance of different constituents of dry plant cell biomass. Biomass production under three different physiological states (i.e., photosynthesis, photorespiration and respiration) is demonstrated and the model is tested against experimental data for two naturally occurring maize mutants (i.e., bm1 and bm3).

B.3.2. Results

The metabolic model reconstruction process follows three major steps: (1) Reconstruction of draft model via automated homology searches for the identification of native biotransformations; (2) Generation of a computations-ready model after defining biomass equation and system boundary and establishing GPR; (3) Model refinement via GapFind and GapFill [220] to unblock biomass precursors as well as reconnect unreachable metabolites. Upon construction of the model, key features such as physiological constraints, network connectivity, light reactions, carbon fixation and secondary metabolism and uniqueness compared to AraGEM and maize C4GEM are described. In addition, model predictions are contrasted against experimental observations.

Construction of Auto & Draft models

The B73 maize genome [207] has 32,540 genes and 53,764 transcripts in the Filtered Gene Set (FGS). Out of 32,540 genes, 30,599 (93%) are evidence-based [221], while the remaining 2,141 (7%) are predicted by the Fgenesh program [222]. 13,726 genes (42% of total) do not have any functional annotation information or are identified as proteins with no or hypothetical/putative functions. Of the remainder, 1,361 (7%) genes encode proteins that do not participate in specific metabolic transformations but rather are involved in transcription, signal transduction, DNA repair, DNA binding, DNA/RNA polymerization, protein folding and adhesion. Because the B73 maize genome is not completely annotated we first established Gene-Protein-Reaction (GPR) mappings for the AraGEM genome-scale model of *A. thaliana* [209] to be used as a proxy. Using these GPRs as a point of comparison we next identified Arabidopsis gene orthologs in maize and transferred the corresponding GPRs via the AUTOGRAPH method [39]. This step was followed by annotation of the remainder maize genes by bidirectional protein BLAST (i.e., BLASTp) searches against the NCBI non-redundant (nr) database. Out of a total of 1,567 metabolic or transport reactions of AraGEM, GPRs were established for 1,254 reactions via 1,467 genes and 653 enzymes by making use of information from several online databases such as AraCyc, KEGG, Uniprot and Brenda (see File S1). Bidirectional BLASTp searches for each one of the 1,467 genes included in AraGEM model were carried out against the B73 maize genome using a stringent cutoff value of 10^{-30}. This fully automated process generated an initial model, termed as *'Automodel'*, containing 946 genes and 1,365 unique metabolites participating in 1,186 reactions (see Table 1 and File S2) exclusively derived from AraGEM. Out of 1,186 reactions, 32 are inter-organelle transport reactions for which homologs were found in maize.

Genes not included in the automodel were scrutinized further by comparing them against the NCBI non-redundant protein database using the same BLASTp cut-off. This increased the model size to 1,485 genes and 1,703 unique metabolites involved in 1,667 reactions by pulling functionalities absent in AraGEM. This is referred to as the *'Draft model'* (see Table 1 and Files S2 and S3). As described in Table 2, orthologous genes were found in *Oryza Sativa* (Rice), *Arabidopsis thaliana* (Arabidopsis), *Sorghum bicolor* (Sorghum) and less frequently in other plant species such as wheat, tobacco, spinach, soya bean, etc. (See File S3). Notably, 802

orthologous genes from *A. thaliana* were added in the model *Zea mays i*RS1563 that were absent from AraGEM primarily due to recent annotation updates. Reactions associated with these genes were subsequently extracted from on-line databases such as KEGG and BRENDA. Table 2 shows the total number of reactions as well as the number of new reactions included in the draft model. Seven reactions having KEGG reaction IDs R00379, R00381, R06023, R06049, R06082, R06138 and R06209 were excluded since they involve generic groups and were not elementally fully defined. Figure 1 shows the distribution of the newly added reactions in the draft model based on their orthologous gene of origin.

Generation of computations-ready model

A computations-ready model requires a fully characterized biomass equation, assignment of metabolites to reactions, establishment of GPR associations, localization of reactions in compartment(s), and inclusion of intra- and extracellular transport reactions [223].

(i) *Establishing a fully characterized biomass equation:* A biomass equation that drains all necessary precursors present in maize was derived (see File S4 and Table 3). We used the biomass composition of young and vegetative maize plants as measured by Penningd *et al.* and expressed on a dry weight basis [224]. The amino acid and lignin composition were derived based on the data from [225, 226]. The composition of hemicellulose was approximated using data for Orchard Grass [227], another monocot grass species, as no corresponding information was found for maize. Based on these compositions we also defined aggregate reactions such as 'Amino acid synthesis', 'Protein synthesis', 'Carbohydrate synthesis', 'Hemicellulose synthesis', 'Lignin synthesis', 'Lipid synthesis', 'Material synthesis', 'Nitrogenous compound synthesis', 'Nucleic acid synthesis' and 'Organic acid synthesis' to produce necessary biomass precursors (i.e., amino acids, protein, carbohydrates, hemicellulose, lignin, lipids, materials, nitrogenous compounds, nucleic acids and organic acids respectively). The biomass equation also contains a non-growth associated ATP maintenance as in the latest Arabidopsis model AraGEM [209].

(ii) *Assignments of genes, reactions, metabolites and compartments.* All metabolic and inter-organelle transport reactions in the draft model have full gene associations. During this step all reactions were elementally balanced and metabolites were assigned appropriate protonation states corresponding to a physiological pH of 7.2. We included an additional 86 reactions to the model without enzyme association information based on direct literature evidence [209]. For example, reactions with KEGG IDs R08053, R08054 and R08055 involved in chlorophyll metabolism are included in the model. Reaction localization information for maize can in some cases be found in database PPDB (a plant proteome database of maize and Arabidopsis) [228]. Because only limited reaction localization information exists for maize, we adopted the compartment or organelle reaction location of the corresponding orthologous gene/enzyme in Arabidopsis using the Arabidopsis Subcellular Database, SUBA [229] and also PPDB [228]. As in AraGEM, reactions for which no such information is available we assumed that they are present only in the cytoplasm.

(iii) *Identification of system boundary*. The entire reaction network (i.e., system boundary) was distributed across five different intracellular organelles enveloped by the cytoplasmic membrane. Exchange reactions were added in the model to ensure that gaseous metabolites (i.e., carbon dioxide and oxygen), inorganic nutrient metabolites (i.e., nitrate, ammonia, hydrogen sulfide, sulfate, phosphate, potassium and chloride), sugar metabolites (i.e., glucose, fructose, maltose and sucrose), water and photons could enter and leave the system whenever necessary depending on the physiological state. As shown in Table 4, constraints on these exchange reactions as well as reactions involved with enzyme RuBisCO (Ribulose-1, 5-bisphosphate carboxylase oxygenase) were established to define three different physiological states (i.e., photosynthesis, photorespiration and respiration) by allowing the selective uptake/release of certain metabolites.

Even though photorespiration is limited in C4 plants (i.e., maize, sorghum, etc.), literature evidence [230-232] alludes that it is still present. Therefore, we made sure that the model is capable of simulating this condition.

The stoichiometric matrix of the draft model (see Table 1) contains 1,901 rows (i.e., total metabolites after taking account of their compartmental appearance) and 1,682 columns (i.e., metabolic reactions, inter-organelle transport reactions and exchange reactions). 970 reactions have one-to-one GPR associations whereas 712 map to more than one gene. 532 reactions map to both isozymes and protein complexes while 4 of them map to only protein complexes, 36 to only isozymes, and 140 to only multimeric proteins.

Network connectivity analysis and restoration

The draft metabolic model inherently contained gaps, unreachable metabolites, omitted transport mechanisms and missing biomass components. We used the procedures termed GapFind and GapFill [233] to correct for these pathologies. We first concentrated on resolving problems with the participation of components in the biomass equation followed by network connectivity.

We found that 723 out of the 1,683 reactions in the draft model could not carry any flux (i.e., blocked reactions) under any of the relevant three physiological states (e.g. photosynthesis (PS), photorespiration (PR) and respiration (R)). As a result, these blocked reactions prevented the formation of some of biomass precursors. GapFind [233] revealed that only 21 out of 64 biomass components could be synthesized using the draft model. GapFill [233] was applied for bridging the gaps through the addition of metabolic and inter-organelle transport reactions and the relaxing of irreversible of existing reactions in the model. GapFill suggested the addition of 94 metabolic and 35 inter-organelle transport reactions in the model to unblock the production of all 64 biomass components. These putative additions to the model were tested by performing an additional round of BLASTp searches for the corresponding genes against the maize genome. We found that 54 (out of 93) metabolic reactions could be assigned to maize gene(s) if the expectation value cut-off for BLASTp was lowered to 10^{-5}. In light of the critical need of restoring biomass formation the less stringent cut-off for inclusion was accepted for these genes. Addition of these reactions ensured the production of biomass under all relevant physiological states validating the use of the term *'Functional'* for the updated model (see Table 1).

Upon ensuring biomass formation GapFind was also applied to assess network connectivity and 715 blocked metabolites were found in the functional model. By applying GapFill connectivity of 322 (45%) blocked metabolites was restored through the addition of 159 metabolic and 3 inter-organelle transport reactions. Table 5 shows the distribution of blocked metabolites into four intracellular organelles before and after applying GapFill. BLASTp searches allowed us to assign 31 (20% of GapFill suggestions) metabolic reactions with specific maize genes (File S2). Biological evidence of the occurrence of such additional reactions in maize or other plant species was sought whenever possible. For example, as shown in Figure 2 phenylacetaldehyde appears to be a "no-consumption" [233] metabolite in the functional model as no reaction can consume it. Using GapFill we found a homolog in maize (i.e., BLASTp score of 10^{-24}) and also literature evidence [234] that *Arabidopsis thaliana* has a aldehyde dehydrogenase activity that catalyzes the conversion of phenylacetaldehyde to phenylacetic acid. Hence, by adding this chemical transformation to *Zea mays i*RS1563 a consumption pathway for phenylacetaldehyde is established. After adding these reactions to the functional model and following charge and elemental balancing and GPR association checking the *'Final' Zea mays i*RS1563 model (see Table 1) is derived.

***Zea mays i*RS1563 model**

The *Zea mays i*RS1563 metabolic reconstruction contains 1,825 unique metabolites and 1,985 reactions associated with 1,563 genes and 876 proteins. Of these reactions 1,898 are metabolic reactions, 70 are inter-organelle transport reactions and 15 are exchange reactions between intra- and extracellular environments. GPR associations are established for all entries (see Table 1). Notably, we identified that the fraction of multifunctional proteins (19% of the total number of proteins) in *Zea mays i*RS1563 is similar to the ratio found in *E. coli* [235]. *Zea mays i*RS1563 accounts for the metabolic functions for all three physiological states. Photosynthetic as well as photorespiration metabolism was modelled by including light mediated ATP and NADPH production via separate charged balanced reactions in the electron transfer system of the thylakoid membrane [236]. Furthermore, the ratio of fluxes for the carboxylation and oxidation reactions associated with enzyme RuBisCO was kept at 1:0 thus ensuring complete carbon fixation during photosynthesis. This ratio was shifted to 3:1 during photorespiration to model simultaneous carbon fixation and oxidation [237]. Because sucrose is the main growth substrate during respiration for higher plants [238], the aforementioned reactions were inactivated and the exchange reaction for sucrose uptake was activated. Under all these three conditions, inorganic nutrients required for plant growth, e.g. sulfate, nitrate, ammonia, hydrogen sulfide, phosphate, potassium and chloride, were allowed to be freely taken up from the environment via extracellular exchange reactions.

The participation of *Zea mays i*RS1563 metabolites across different compartments is shown in Figure 3. The five intracellular organelles differ notably in terms of mutual connectivity, metabolite uniqueness and number of metabolites. As shown in Figure 3a, approximately 90% of these metabolites are unique to cytoplasm. In addition, cytoplasm contains all metabolites shared between any two organelles because any metabolite needs to be transported through cytoplasm in order to be exchanged between organelles. Among the remaining metabolites, cytoplasm shares the highest number with the plastid (i.e., 63) where photosynthesis and photorespiration occur. It also shares a significant number of metabolites with mitochondrion (i.e., 27) and peroxisome (i.e., 22) that are involved in energy production and fatty acid biosynthesis, respectively. Figure 3b shows the distribution of other non-cytoplasmic *Zea mays i*RS1563 metabolites in terms of how many organelles they participate.

Light reactions, carbon fixation and secondary metabolism

In plants photosynthesis reactions include light dependent and light independent or carbon fixation reactions [239]. *Zea mays i*RS1563 includes charged balanced light reactions culled from a number of literature sources [236, 240-242]. The overall photosynthesis reaction cascade produces two NADPH, three ATP and one O_2 whenever nine photons are absorbed and fourteen H^+ are transferred via the electron-transport system. This defines the following overall balance equations:

$$12\ H^+[c] + 2\ H_2O[c] + 2\ NADP[c] + 9\ hvi[c] \rightarrow 14\ H^+[p] + 2\ NADPH[c] + O_2[c] + 9\ hvo[c]$$

$$3\ ADP[c] + 14\ H^+[p] + 3\ Pi[c] \rightarrow 3\ ATP[c] + 14\ H^+[c]$$

Here, [c] and [p] represent cytoplasm and plastid and hvi and hvo signify input and output photons respectively. Carbon fixation in maize (C_4 plant) is more complex compared to Arabidopsis or other C_3 plants [239]. *Zea mays i*RS1563 captures these differences by accounting for (i) direct carboxylation of phosphoenol pyruvate and CO_2 fixation to form C_4 acids such as oxaloacetic acid [ATP: oxaloacetate carboxy-lyase *(ocl)*] and malic acid [Oxaloacetate: NADPH hydrogenase *(oha)*] in mesophyll cells, (ii) transport of malic acid from mesophyll cell to bundle-sheath cells, (iii) decarboxylation of malic acid [Malate:NADP+ oxidoreductase *(mor)*] in bundle-sheath cells to produce pyruvic acid and CO_2, which enters the Calvin cycle, (iv) transport of pyruvic acid from bundle-sheath cells to mesophyll cells, and (v) production of phosphoenol pyruvic (i.e., C_3) acid [ATP:pyruvate,phosphate phosphotransferase *(ppt)*] from pyruvic acid

[239]. Figure 4, pictorially shows the localization of reactions and organelles between mesophyll and bundle sheath cells. In addition, to differences in carbon fixation reactions, the peroxisome activity is primarily present in bundle-sheath cells and largely absent from mesophyll cells [243]. Based on this localization information a standalone metabolic model can be developed for the photosynthetic tissue of maize. Because RuBisCO that operates in the Calvin cycle cannot come in direct contact with atmospheric oxygen during day time (see Figure 4), photorespiration is restricted providing an advantage for survival in hot and arid environments for maize and other C_4 plants. This comes at the expense of higher (ATP) requirements as C_4 carbon fixation involves additional steps [239].

In addition to photosynthesis, secondary metabolism plays a key role in the physiology of maize. For example, phenylpropanoid metabolism produces monolignols (i.e., *p*-coumaroyl alcohol, coniferyl alcohol and sinapyl alcohol) that are used in the generation of three major lignin subunits H-lignin, G-lignin and S-lignin, respectively [244]. Many of these enzymes such as hydroxycinnamoyl transferase (HCT), ferulate 5-hydroxylase (F5H) and caffeic acid 3-*O*-methyltranferase (COMT) along with their associated reactions are unique to C_4 plants and are not present in the lignin biosynthesis pathways of *A. thaliana* [244]. HCT is involved in the early stages of lignin biosynthesis by controlling the flux from *p*-coumaroyl-CoA towards caffeoyl-CoA while F5H and COMT regulate fluxes from coniferaldehyde and coniferyl alcohol to sinapaldehyde and sinapyl alcohol, respectively [244]. *Zea mays i*RS1563 contains all these enzymes and associated reactions thus providing a comprehensive lignin biosynthesis pathway for a C_4 plant.

In addition to phenylpropanoid metabolism, *Zea mays i*RS1563 provides a detailed description of flavonoid biosynthesis pathways. Flavonoids are pigments occurring in plant as secondary metabolites and mostly function in the recruitment of pollinators and/or seed dispersers [245]. For example, maize is known to produce 3-deoxyanthocyanins, which are a specialized class of flavonoids [246, 247]. *Zea mays i*RS1563 contains the dihydroflavonol 4-reductase (DFR) enzyme that catalyzes the reaction for flavan-4-ols biosynthesis that channels flux towards 3-deoxyanthocyanins production [247]. The model also accounts for isoflavone 7-O-glucosyltransferase (IF7GT) and associated reactions that are involved in the production of necessary intermediates for pterocarpin phytoalexin conjugates such as medicarpin 3-O-glucoside-6'-O-malonate (MeGM) and maackain 3-O-glucoside-6'-O-malonate (MaGM) involved in plant defense against fungal elicitation [248].

Comparing *Zea mays i*RS1563 with *Arabidopsis thaliana* and maize C4GEM models

Figure 5a compares the total number of genes, reactions and metabolites between *Zea mays i*RS1563 and the *A. thaliana* AraGEM genome-scale-models [209]. Approximately, only 61% of genes in *Zea mays i*RS1563 are present in AraGEM. This yields a surprisingly low degree of matching between these two models of 64% and 76%, respectively in terms of reactions and metabolites. In the interest of elucidating the true differences between maize and Arabidopsis irrespective of annotation chronology we constructed a more up-to-date genome-scale model for Arabidopsis by appending onto AraGEM newly annotated genes as well as full GPR annotations. We refer to this updated model containing 1,597 genes, 1,798 reactions and 1,820 metabolites as *A. thaliana i*RS1597 (see File S1). The newly added 228 reactions (absent from AraGEM) are involved in various pathways in primary (i.e., glycolysis, TCA, fatty acid and amino acid biosynthesis, starch and sucrose metabolism) and secondary (i.e., biosynthesis of steroid, ubiquionone, streptomycin, thiamin, riboflavin, terpenoid, brassinosteroid, phenylpropanoid, etc.) metabolism of Arabidopsis.

A direct comparison of *Zea mays i*RS1563 with *A. thaliana i*RS1597 reveals, as expected, an increased degree of matching of 72%, 76% and 80% in terms of genes, reactions and

metabolites, respectively (see Figure 5b). We find that 445 reactions are unique to maize with no counterpart in *A. thaliana*. Secondary plant metabolism including flavonoid, mono- and diterpenoid, brassinosteroid, phenylpropanoid, anthocyanin, zeatin biosynthesis, riboflavin and caffeine metabolism account for 185 of the maize-specific reactions. In addition, a variety of primary metabolism reactions dispersed throughout central metabolism, photosynthesis, amino acid and fatty acid biosynthesis account for the remaining 260 reactions. This comparison implies that about one third of the differences between *Zea mays i*RS1563 and AraGEM are caused by the incompleteness of AraGEM model especially in terms of secondary metabolism while the remaining two third reflect genuine differences between C_3 (i.e., Arabidopsis) and C_4 (i.e., maize) plant metabolism.

Figure 5c shows a similar comparison between *Zea mays i*RS1563 and maize C4GEM genome-scale-models. Degrees of matching between these two models are 39%, 53% and 63% in terms of genes, reactions and metabolites, respectively. This surprisingly low degree of matching is caused primarily due to the fact that maize C4GEM includes only metabolites and reactions in leaves during photosynthesis. Therefore, there are 893 reactions in *Zea mays i*RS1563 absent from maize C4GEM. 343 of these reactions describe secondary plant metabolism such as brassinosteroid, phenylpropanoid, carotenoid, flavonoid, mono- and diterpenoid, and glucosinolate metabolism. The remaining 550 reactions are found in a wide range of primary metabolism pathways such as central metabolism, photosynthesis, benjoate degradtion, starch and sucrose metabolism, lipid metabolism, nitrogen metabolism amino acid and fatty acid biosynthesis. Conversely, 116 (out of 149) new reactions in maize C4GEM have untraceable EC numbers and gene loci.

*Zea mays i*RS 1563 model testing

*Zea mays i*RS1563 allows for the production of biomass under all three different physiological states (see Files S5 and S6 for detailed information of the model). Due to limited photorespiration C4 plants usually have higher photosynthetic efficiency [239]. Under higher light intensity and photosynthetic condition, *Zea mays i*RS1563 produces 0.0008 mole biomass/mole CO2 whereas *A. thaliana i*RS1597 yields 0.0006 mole biomass/mole CO2. Thus, the model predictions match with findings reported in literature [239]. We also investigated the model's ability to predict the effect of suppressing genes in the lignin biosynthesis pathway observed in naturally occurring *brown midrib* (*bm*) maize mutants (i.e., *bm1, bm2, bm3* and *bm4*) [244, 249-251]. These maize mutants are Mendelian recessives that are characterized by brown vascular tissue in leaves and stems due to a changed lignin content and/or composition [252]. The specific genetic background for two of these mutants (*bm1* and *bm3*) was elucidated based on the analysis of cell wall composition [251]. Mutants *bm1* and *bm3* were found to have disrupted enzymatic activity for cinnamyl alcohol dehydrogenase (CAD) and caffeic acid 3-*O*-methyltranferase (COMT). Both of these enzymes are involved in the last stages of the monolignol pathway [251] that controls lignin synthesis and composition (i.e., the ratio of three major subunits, H-lignin, G-lignin and S-lignin) [253].

We simulated mutants *bm1* and *bm3* using *Zea mays i*RS1563 under photosynthetic conditions by restricting the flux of the reactions catalyzed by enzymes CAD and COMT to 10% of the wild-type values. It is expected that the disruption of the activity for these genes will directly affect lignin content and composition (see File S7 to find literature data used for *bm1* and *bm3* mutants). We were interested to see whether the *Zea mays i*RS1563 metabolic model will be able to correctly propagate this disruption across the metabolic pathways and correctly predict the effect on other key metabolites. Table 6 contrasts experimental results by (Marita et al (2003), Vanholme et al (2008) and Sattler et al (2010)) with *in silico* predictions for the maximum theoretical yield of lignins, sugars and crude protein in terms of whether they increased, decreased, or remained the same in the mutant strains. Out of 21 compared components *Zea mays*

*i*RS1563 correctly predicted the direction (or absence) of change for 17 cases.

In Figure 6 we highlight two cases that describe the availability of glucose and galactose to cell wall for mutants *bm1* and *bm3*, respectively. 'Carbohydrate synthesis' and 'Hemicellulose synthesis' are aggregate reactions that describe the utilization ratios of sugar molecules such as arabinose, fructose, galactose, glucose ribose, mannose, sucrose, and xylose for the production of carbohydrate and hemicellulose present in the plant cell wall. For simplicity, we have simulated the model under the photosynthetic condition where CO_2 can be uptaken with a maximum allowable rate of 1000 mM/gDW-h along with photons in excess. In Figure 6a, wild-type and *bm1* mutant flux values for reactions involving glucose as reactant including 'Carbohydrate synthesis', 'Hemicellulose synthesis', 'Alpha,alpha-trehalose glucohydrolase' [R00010], 'Sucrose glucohydrolase' [R00801], 'Sn-Glycerol-3-phosphate: D-glucose 6-phosphotransferase' [R00850] and 'Cellobiose glucohydrolase' [R00306], are highlighted. For the wild-type case, the maximum theoretical yield of glucose is predicted to be 1.66 moles/mole of CO_2 but it is reduced to 0.93 moles/moles of CO_2 for the *bm1* mutant. The reduced capability of the *bm1* mutant to direct flux towards 'Carbohydrate synthesis' and 'Hemicellulose synthesis' implies that less glucose is available for the formation of cell wall components which is consistent with the experimental finding of Table 6.

Figure 6b contrasts the wild-type and *bm3* mutant maximum theoretical yields for all reactions involving galactose including 'Hemicellulose synthesis', 'ATP: D-galactose 1-phosphotransferase' [R01092] and 'Galactosylglycerol galactohydrolase' [R01104], '3-O-alpha-D-Galactosyl-1D-myo-inositol galactohydrolase' [R01194] and 'alpha-galactosidase' [R03634]. A reduction of the maximum theoretical yield of galactose from 0.81 to 0.65 moles/mole of CO_2 for the *bm3* mutant is observed. In addition, the maximum theoretical yield for reaction 'Hemicellulose synthesis' decreases by 4-fold compared to wild-type in line with the experimental finding. However, the experimentally observed increase of glucose availability in mutant *bm3* and xylose availability for both *bm1* and *bm3* mutants are in contrast with the model predictions (see Table 6). As reported by Guillaumie et al (2007) several gene expression levels were changed during *bm1* and *bm3* mutations implying that additional regulatory constraints may be needed to capture these changes.

B.3.3. Discussion

Maize, apart from its central role a food crop, is also a promising plant biomass target for cellulosic biofuels production. Plant cell wall cellulose, hemicellulose and lignin polymers are major contributors of plant biomass [244, 254]. Therefore, controlling the amount and composition of cell wall polymers is important in developing cellulosic maize for biofuel production. In cell wall, lignin provides rigidity by forming a matrix where cellulose and hemicellulose are imbedded via cross-linking bonds [249, 255]. This makes digestion of cellulose and hemicellulose by microbial enzymes (i.e., cellulases) difficult during dilignification, one of the critical steps in cellulosic biofuel production [256]. Many genetic modification strategies have been explored to improve maize food crop and/or biofuel characteristics. For example, cellulosic biomass yield improvements have been pursued before by altering the lignin content and composition [257, 258], genetically manipulating the cellulose biosynthetic pathway [259] and over-expressing the gene encoding phosphoenolpyruvate carboxylase (PEPC) to improve CO_2 fixation rate [260]. At the same time, grain yield enhancements have been attempted by up-regulating ADP-glucose pyrophosphorylase (AGP) that catalyzes the rate limiting step in starch synthesis [261].

Unfortunately, existing genetic engineering strategies to reduce lignin content are problematic as lignin reductions are usually achieved at the expense of plant viability and fitness [256]. It is becoming widely accepted that focusing on a single pathway at a time without quantitatively assessing the system-wide implications of the genetic disruptions may be

responsible for not preserving the agronomic properties of the plant. By accounting for both primary and some secondary metabolism pathways of maize, *Zea mays i*RS1563 can be used to explore *in silico* the effect of genetic modifications aimed at plant cell wall modification and/or starch storage on the overall metabolic state of the plant (e.g., biomass precursor availability, cofactor balancing, redox state, etc.). Moving a step further, the use of computational strain optimization techniques [196, 262] can be customized for engineering plant metabolism. By taking full inventory of plant metabolism optimal gene modifications could be pursued for a variety of targets in coordination with experimental techniques. These may include (i) increase cellulose and hemicellulose production, (ii) starch yield, (iii) tolerance against biotic stress (e.g., fungal elicitation), or (iv) disruption of the production of lignin subunits (H/G/S) while enhancing the production of easily digestible lignin precursor (e.g., rosmarinic acid, conferyl ferulate, tyramine conjugates, etc).

In this section, we introduced the first comprehensive genome-scale metabolic model (*Zea mays i*RS1563) for maize metabolism. The model meets (or exceeds) the quality and completeness criteria set out [263, 264] for genome-scale reconstructions. In analogy to the human genome-scale model Recon 1 [265], *Zea mays i*RS1563 can be viewed as a mathematically structured database enabling systematic studies of maize metabolism.185 of unique to maize reactions accounting for a fraction of secondary metabolism were delineated. As a by product of this effort a more up-to-date version of AraGEM [209] was constructed including GPR associations. Comparisons between *Zea mays i*RS1563 and maize C4GEM also revealed the detail in description of primary and secondary metabolism. Model predictions of *Zea mays i*RS1563 for two widely occurring maize Mendelian mutants were tested against experimental observations with very good agreement in the direction of changes. By making use of high throughput enzymatic assays, proteomic and transcriptomic data across different parts of the maize plant, *Zea mays i*RS1563 could serve as the starting point for the development of tissue-specific maize models [217, 266, 267]. Furthermore, *Zea mays i*RS1563 could also serve as the stepping stone for the development of genome-scale models for other important C_4 plants such as Sorghum and switch grass.

B.3.4. Materials and Methods

A number of recent publications [212, 223, 263] have outlined the general steps necessary for the metabolic reconstruction process. In the following section, we highlight the specific methods used in the reconstruction of *Zea mays i*RS1563 and subsequent model simulations in more detail.

Model reconstruction

The maizesequence database [207] provided the filtered gene set (FGS) which has been generated from the working gene set upon removing pseudogenes and low confidence hypothetical models. The FGS of B73 maize genome (release 4a.53) was downloaded from maizesequence database on February 17, 2010. Once maize genes were obtained, we used sequence comparison tools [268] such as stand-alone BLAST (version 2.2.22, NIH) and BLAST+ (version 2.2.22, NIH) for performing homology comparisons. Marvin (version 5.3.3, ChemAxon Kft) was used to calculate the average micro-species charge to determine the net charge of individual metabolites at pH 7.2 assumed for all organelles. In the final step of the model reconstruction, we implemented GapFind and GapFill [233] for analyzing and subsequently restoring metabolic network connectivity.

Model simulations

Flux balance analysis (FBA) [269] was employed both in model validation and model testing phases. *Zea mays i*RS1563 was evaluated in terms of biomass production under three standard

physiological scenarios: photosynthesis, photorespiration, and respiration. Flux distributions for each one of these states were approximated using FBA:

Maximize $v_{Biomass}$

Subject to

$$\sum_{j=1}^{m} S_{ij} v_j = 0 \;\forall i \in 1,\ldots..,\ \mathrm{n} \qquad (1)$$

$$v_{j,\min} \le v_j \le v_{j,\max} \;\forall j \in 1,\ldots..,\mathrm{m} \qquad (2)$$

Here, S_{ij} is the stoichiometric coefficient of metabolite i in reaction j and v_j is the flux value of reaction j. Parameters $v_{j,min}$ and $v_{j,max}$ denote the minimum and maximum allowable fluxes for reaction j, respectively. As mentioned in Table 4, the three physiological states were represented via modifying the relevant minimum or maximum allowable fluxes and the following constraints:

$$v_{oxi} = 0 \qquad (3)$$

$$v_{carboxi} \ge 3 v_{oxi} \qquad (4)$$

$$v_{carboxi} = 0 \qquad (5)$$

where $v_{Biomass}$ is the flux of biomass reaction and v_{oxi} and $v_{carboxi}$ are the fluxes of carboxylation and oxidation reactions associated with enzyme RUBISCO. For photosynthesis and photorespiration, constraints (3) and (4) were respectively included in the linear model, whereas for respiration both constraints (3) and (5) were included.

Once the model was validated, it was further tested for two maize mutants (i.e., *bm1* and *bm3*) under the photosynthetic condition. The following two constraints were included individually in the linear model to represent the mutants:

$$v_{bm1} \le w \times WF_{bm1} \qquad (6)$$

$$v_{bm3} \le w \times WF_{bm3} \qquad (7)$$

Here, w represents the percent of residual activity of 10%. v_{bm1} and v_{bm3} are the fluxes of reactions catalyzed by CAD and COMT, respectively and WF_{bm1} and WF_{bm3} are the corresponding wild-type flux values under the photosynthetic condition.

CPLEX solver (version 12.1, IBM ILOG) was used in the GAMS (version 23.3.3, GAMS Development Corporation) environment for implementing GapFind and GapFill [233] and solving the aforementioned optimization models. All computations were carried out on Intel Xeon E5450 Quad-Core 3.0 GH and Intel Xeon E5472 Quad-Core 3.0 GH processors that are the part of the lionxj cluster (Intel Xeon E type processors and 96 GB memory) of High Performance Computing Group of The Pennsylvania State University.

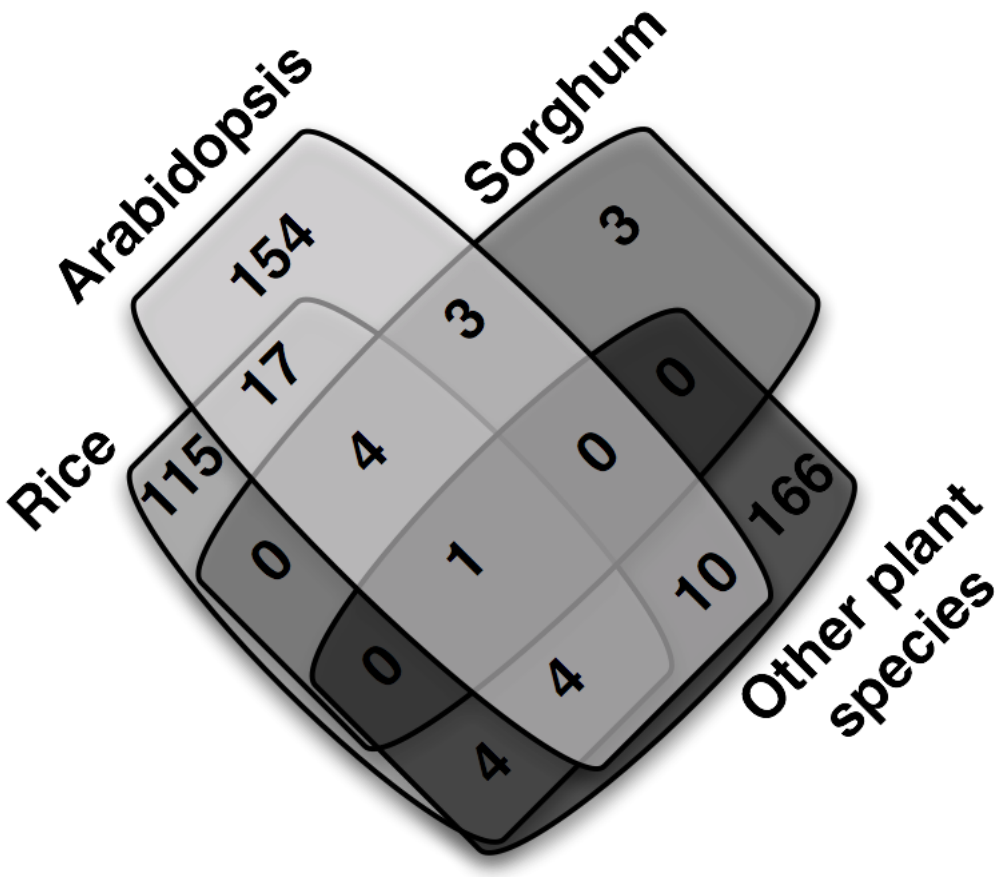

Figure 1. Species origin of newly added reactions in the draft model.

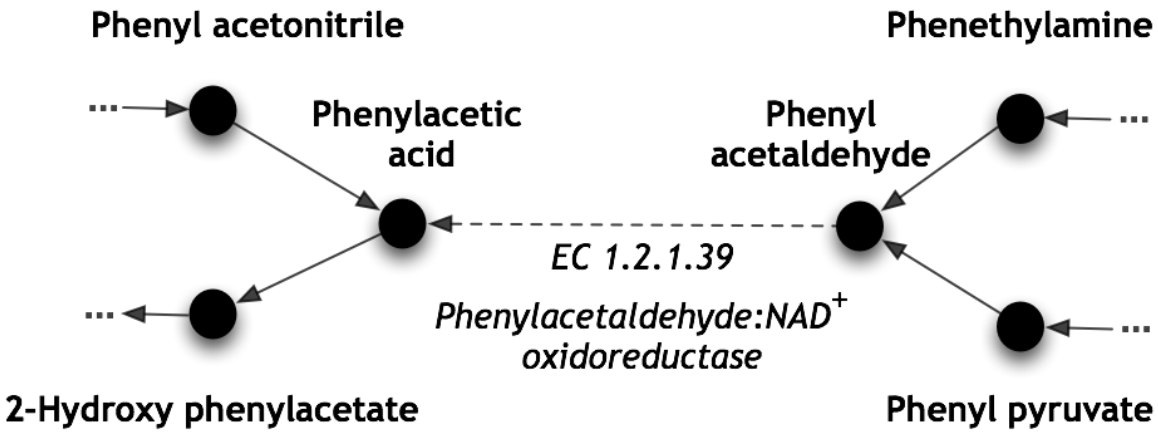

Figure 2: Example of connectivity restoration for phenylacetaldehyde.

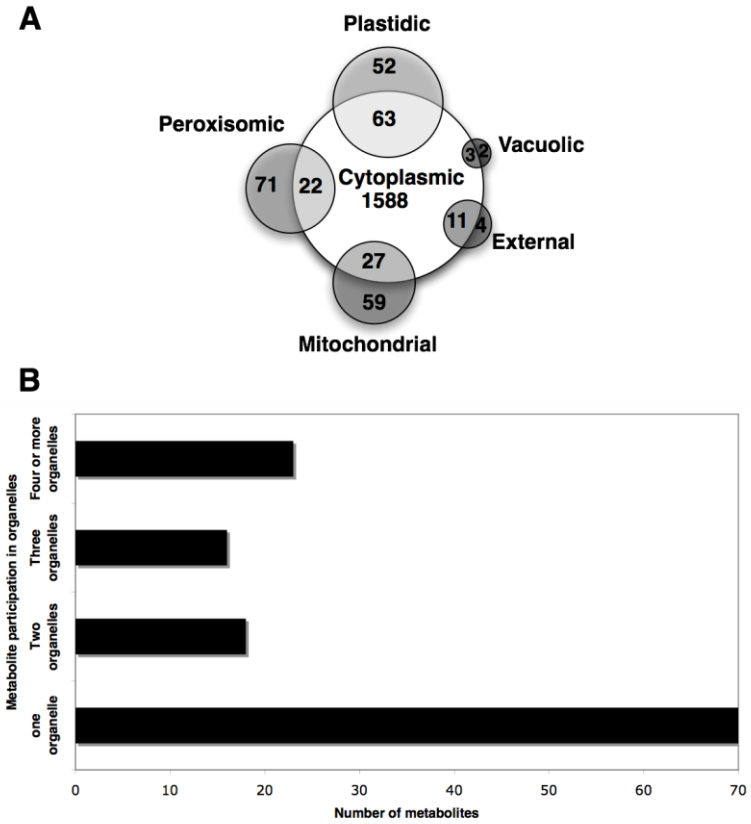

Figure 3: Distribution of metabolites based on their number of appearance in different organelles. (a) cytoplasmic *Zea mays i*RS1563 metabolites in cytoplasm and other organelles, and, (b) non-cytoplasmic *Zea mays i*RS1563 metabolite-organelle participation.

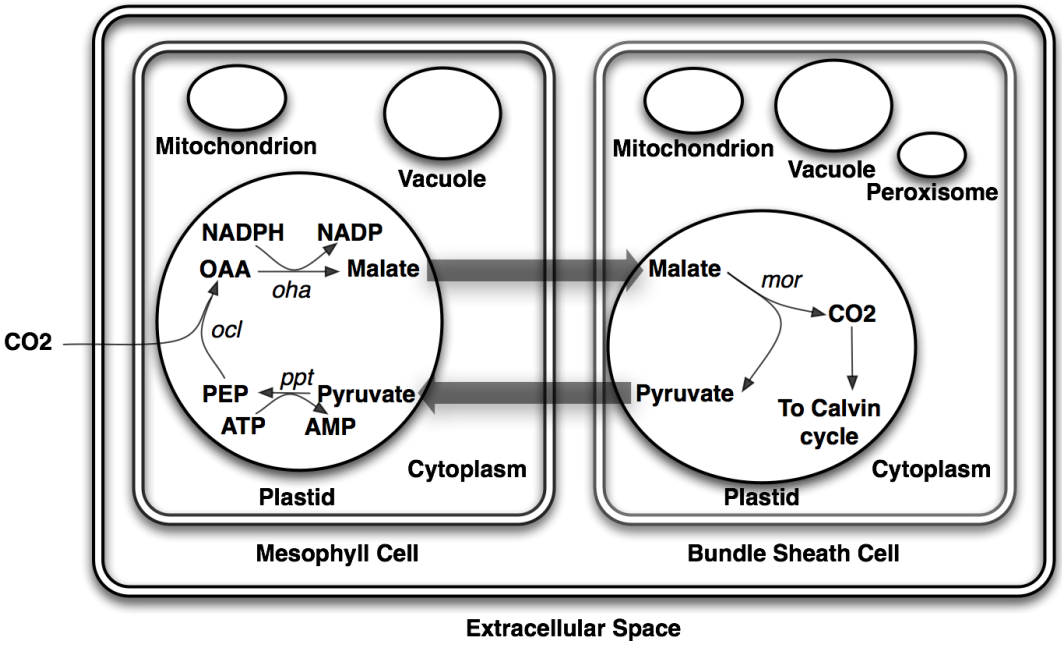

Figure 4: Compartment and localization information for *Zea mays iRS* 1563. Mitochondrion and vacuole compartments are present in both cell types whereas peroxisome is only present in bundle-sheath cell [40]. Plastidic reactions are distributed between mesophyll and bundle-sheath cells.

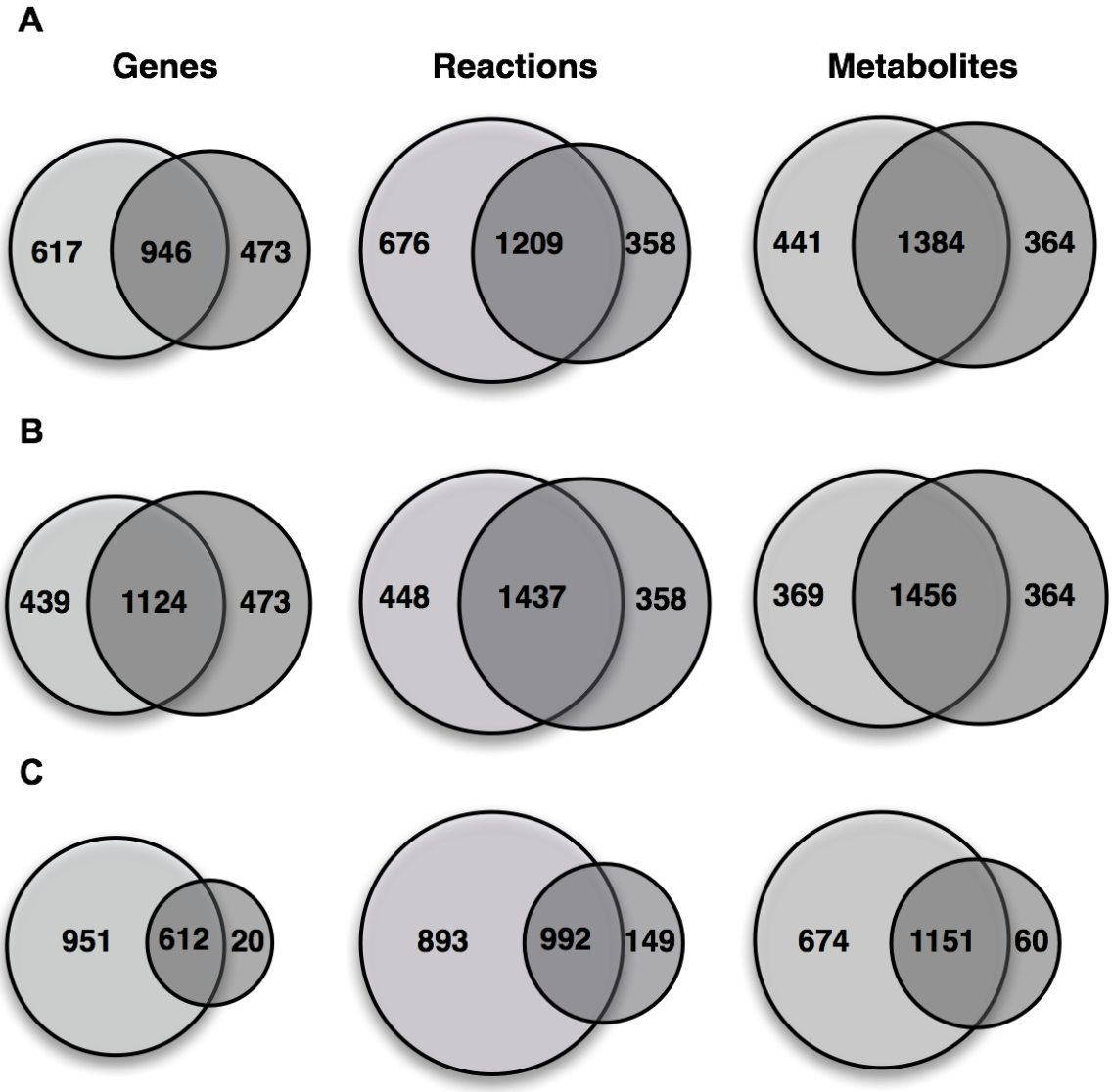

Figure 5: Venn diagram for genes, reactions and metabolites. (a) between *Zea mays i*RS1563 and AraGEM, (b) between *Zea mays i*RS1563 and *Arabidopsis thaliana i*RS1597, and (c) between *Zea mays i*RS1563 and maize C4GEM.

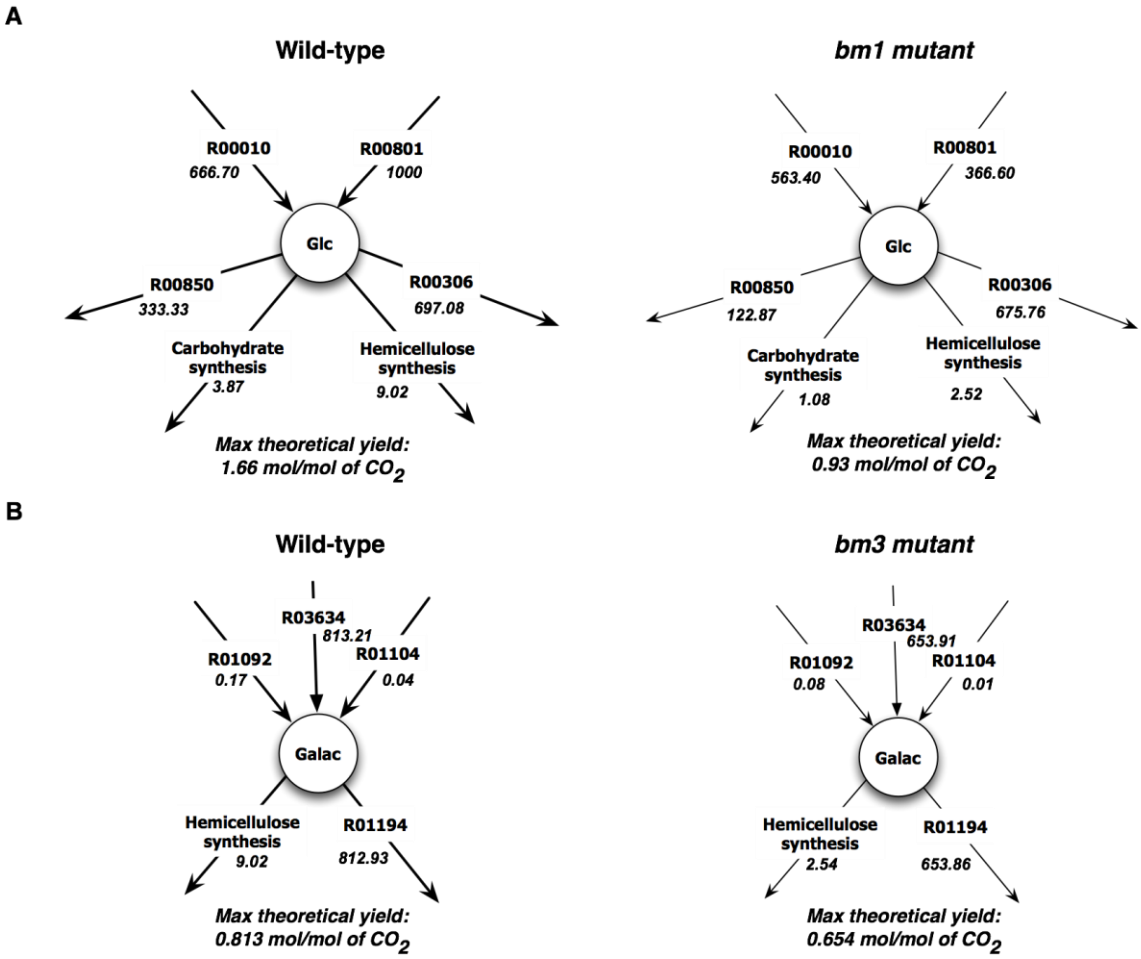

Figure 6: Maximum theoretical yields of (a) glucose and (b) galactose for wild-type vs *bm1* mutant and wild-type vs *bm3* mutant, respectively. Here the numeric values represent reaction fluxes and have the unit of mM/gDW-h.

B.4. Metabolic reconstruction of the archaeon methanogen *Methanosarcina Acetivorans*

The work in this section has been published [270].

B.4.1 Introduction

Genome-scale metabolic models (for recent reviews, see [37] and [271]]) are increasingly becoming available for an expanding range of organisms. There exists at least forty completed bacterial and thirteen eukaryotic metabolic reconstructions with many more under development [37]. In the past decade, several studies [272] have demonstrated a variety of uses ranging from strain optimization [21, 273, 274] pathogen drug target identification [275, 276], bacterial community metabolic interactions [19] and identification of human disease biomarkers [277]. In contrast to the extensive interest devoted towards bacterial and eukaryotic metabolism reconstruction, efforts to construct archaeal metabolic models have been noticeably limited [53, 74]. This is partly due to the current relative paucity of -omics datasets available for species from the *Archaea* domain. This dearth of data, however, is likely to change in the near future as recent

interest in methanogenic *archaea* has led to several sequencing efforts [278-280], as well as transcriptomic and proteomic analyses [281-285]. Furthermore, it is increasingly becoming apparent that archaeal metabolism has significant implications to the earth's climate [286] thus motivating the need to globally assess their true metabolic capabilities by reconstructing high quality metabolic models.

Methanogens (def., methane-producing) constitute the largest group of the *Archaea* domain of life [287]. Methanogens are terminal organisms in anaerobic microbial food chains (i.e., consortia) that break down organic matter to methane in diverse anaerobic environments in a process that helps regulate the global carbon flux [288] . The process plays a surprisingly significant role in global warming accounting for about one billion tons of the annual methane production [286, 289]. Recently, Cheng and coworkers used a consortia of methanogens to convert electricity into methane thereby paving the way for converting electric current generated by renewable energy sources into renewable biofuels [290]. On the evolutionary front, methanogens are amongst the most ancient form of life on earth and their role as the progenitors of the first eukaryotic cell has been suggested under two separate hypotheses [291, 292]. In addition, unique biochemical properties such as broad substrate specificity, participation of novel coenzymes in the methanogenesis pathways and the presence of unique lipids in their cell wall set these organisms apart from the bacterial and eukaryotic branches of life [293]. Therefore, the reconstruction of archaeal methanogen metabolic models could help paint a more complete picture of life's metabolic processes.

Feist and coworkers first developed a genome-scale model (named *i*AF692) [74] for the fresh-water methanogen, *Methanosarcina barkeri* using a draft version of its genome. In this work, we reconstruct a genome-scale metabolic model for the marine methanogen, *Methanosarcina acetivorans*. *M. acetivorans* is an aceticlastic methanogen that was first isolated from marine microbial communities that degrade kelp into methane [294]. At over 5.7 million base pairs [280], it has the largest reported genome of all sequenced *Archaea* (about 20% larger than the *M. barkeri* genome) alluding to an expanded metabolic repertoire. This repertoire includes unique methanogenic pathways, broad substrate specificity than other methanogens and a large number of duplicate genes [280, 283-285, 295, 296]. Recent studies have shown that *M. acetivorans* uniquely exhibits both methanogenic and acetotrophic growth on carbon monoxide [296]. All these unique metabolic characteristics and planet-wide carbon balance impact [286, 289] provide the motivation to globally assess its metabolic capabilities.

Draft metabolic reconstructions generated using homology-based comparisons unavoidably contain some omissions and misclassifications. These errors are manifested either as unreachable metabolites or as *in silico* predictions that are in contrast with observed *in vivo* behavior [152, 220]. In response to these challenges, Suthers et al., proposed a computational workflow to generate and curate the metabolic models and applied it to the metabolic reconstruction of *Mycoplasma genitalium* [223]. The proposed workflow makes use of two separate model correction procedures. GapFind and GapFill identify and subsequently restore connectivity to unreachable metabolites [220] and GrowMatch that reconciles *in silico* growth predictions with *in vivo* growth data [152]. In this work, we streamline this workflow for the generation of an archaeal metabolic model and customize it to the available types of data.

We first generated a draft reconstruction of *M. acetivorans* using homology comparisons with the published model [74] of the fresh-water methanogen, *M. barkeri*. We then deployed a modified version of the workflow presented in Suthers et al., by combining the GapFind/ GapFill and GrowMatch steps of the procedure [223], which was included in the first year progress report of this grant. The completed model accounts for 1007 genes, 835 reactions and 790 metabolites. The model also predicted substrate specific phenotypes of *M. acetivorans* and captured unique bioenergetics exhibited by the organism across different conditions.

B.4.2 Results

The metabolic model reconstruction workflow consists of four steps. Step 1 refers to the draft model generation using bidirectional BLAST homology (BBH) and database/literature searches. Step 2 involves model modifications to ensure biomass formation for growth under all known substrates. Step 3 applies GrowMatch [152] to restore growth prediction inconsistencies and Step 4 applies GapFind and GapFill [220] to restore connectivity.

Step 1: Generating Draft model

BBH searches for each of the 692 genes included in the *i*AF692 model were conducted against the latest genome sequence of *Methanosarcina acetivorans C2A* strain [280]. At this stage of the reconstruction process, we included only open reading frames (ORFs) that have e-values (in both directions) of at most 10^{-30}. This process yields an initial conservative model for *M. acetivorans* that has 820 genes. Based on the primary TIGR annotation of *M. acetivorans* [297] this accounts for 18.1% (820/4540) of all ORFs in the *M. acetivorans* genome. The mapping of the metabolic genotypes between these two very closely related organisms is surprisingly complex. Specifically, 369 one-to-one mappings, 1,113 one-to-many mappings and 1,050 many-to-many mappings (*M. barkeri* to *M. acetivorans*) were generated. The large number of one-to-many and many-to-many mappings is consistent with the incidence of a high number of gene duplicates in the *M. acetivorans* genome [280] [298].

We use multiple sources to annotate the remaining 3,720 ORFs in the genome. Specifically, we assigned metabolic annotation to seven genes based on the information available at an organism-specific annotation effort for *M. acetivorans* [299], 51 genes based on SEED annotations [300] and 110 genes based on TIGR annotations. Interestingly out of these 168 genes as many as 68m code for isozymes. Predicted or hypothetical proteins account for the remaining 2,411 ORFs not included in the model after the annotation step. Approximately 46% of all genes in *M. acetivorans* (upon excluding hypotheticals and predicted proteins) were present in the draft metabolic model. The methanogenesis pathways in the *M. acetivorans* model were modified to account for known differences documented in the literature. Specifically, we added reactions carbonic anhydrase (abbreviation in *i*VS1007: CAM), multiple resistance/pH regulation Na+ /H+ antiporter (abbreviation in *i*VS1007: MRP) and an electron transfer complex which oxidizes ferredoxin and exports sodium ions (abbreviation in *i*VS1007: RNF). We removed the ECH and VHO hydrogenases The added reactions are involved in ATP synthesis and replace the activities of ECH and VHO hydrogenase, which are observed in H_2/CO_2-utilizing fresh-water methanogens [295]. In contrast with other archaeal models [53, 74], we delineated methyltransferase specificity [298, 301] for different substrates observed in *M. acetivorans* .

We generated the Gene-Protein-Reaction mappings for the *M. acetivorans* model using as a starting point the *i*AF692 model based on the AUTOGRAPH method developed by Notebaard and coworkers [39]. All exchange reactions and non-gene associated intracellular reactions available in the *i*AF692 model were also appended to the model, as we did not find any evidence to the contrary [see Methods]. Upon conclusion of Step 1, a draft model with 988 genes, 820 reactions and 792 metabolites was generated.

Step 2: Model correction to enable biomass formation

We determine the metabolic capabilities of the assembled draft model to grow on known methanogenic substrates by first specifying the biomass equation and then specifying the composition of the minimal medium. The first requirement is addressed by assuming that the set of components that make up the biomass equation in *M. acetivorans* is identical to that used in the *i*AF692 model. However, we changed the stoichiometric coefficients of the nucleotide components (datp, dgtp, dctp, dttp, ctp and gtp) to reflect the difference in the G/C contents of the

two organisms. The utilization of the same biomass component set is supported by experimental data on the minimal medium (Ferry et al., unpublished data). The minimal growth medium six additional vitamins and trace elements (pyridoxine-HCL, sodium molybdate, thioctic acid, nitrilo tri acetic acid and boric acid) over the one used in *i*AF692 [74]. We chose to exclude them from our model as no metabolic role for them was identified based on literature searches or gleaned by the model.

Equipped with the biomass composition and the minimal medium, we tested the capability of the draft model to enable growth on the following known methanogenic substrates: carbon monoxide, acetate, methanol and monomethylamine, dimethylamine and trimethylamine [294]. The draft model did not exhibit growth on any of these substrates motivating the use of GapFind [220] to identify the biomass precursor metabolites that could not be produced using these substrates in a minimal medium. GapFind revealed that the same precursor metabolite Adenosylcobalamin-HBI could not be produced for all substrate choices in the draft model. We used GapFill [220] to restore flow through this metabolite. This was achieved under all aforementioned substrate conditions by adding an exporter for the cofactor, tetrahydrosarcinapterin. No evidence was found in the literature for the presence of a tetrahydrosarcinapterin exporter. However, it is possible there exists an enzyme outside the cell wall that utilizes the cofactor as a substrate.

Step 3: Evaluating and improving model performance using GrowMatch

After ensuring *in silico* growth on a defined medium across different substrates, we further examined the model by testing for growth prediction agreement with experimental data across different genetic/environmental perturbations. Using literature searches, we assembled a dataset consisting of *in vivo* growth data for 66 different conditions (See Table B.4.2). As shown in Table B.4.2, growth data was assembled for 29 genetic perturbations for growth on methanol, thirteen on acetate, seven on carbon monoxide as carbon and energy source, and 22 on methylamines as carbon substrates. Not surprisingly, most of these gene deletions are in methanogenesis pathways (Table B.4.2) indicative of the significant attention this pathway has received before.

In line with previous approaches [302] the growth cutoff for classifying a mutant as a "Growth" or a "No-Growth" mutant was chosen to be $1/3^{rd}$ of average growth across the dataset. Using this cutoff and the terminology introduced in the GrowMatch procedure [152] we arrive at 43 GG (*in silico* and *in vivo* "**G**rowth") fifteen GNG (*in silico* "**G**rowth" and *in vivo* "**N**o-**G**rowth") and eight NGNG (*in silico* and *in vivo* "**N**o-**G**rowth") cases. Notably, the incidence of only GNG model/experimental discrepancies indicates that the draft model tends to over-predict the metabolic capabilities of the organism when in error. A closer examination reveals that in 32 out of 43 GG cases the deleted genes correspond to isozymes while the remaining eight correspond to deletions of methyltransferases. In all these cases both the model and the experiment agree that the deleted genes are non-essential. Of the nine GNG cases that could be resolved, eight were resolved by conditionally suppressing one additional reaction and one was resolved by carrying out a single global suppression.

M. acetivorans. As shown in Figure B.4.1(A), the genes encoding for Methyl Coenzyme reductase (the reaction that forms methane) under growth on Carbon Monoxide are non-essential *in silico* and essential *in vivo* [303]. GrowMatch suggests suppressing either the reaction catalyzed by acetate kinase (ACKr) or phosphotransacetylase (PTAr) to restore consistency to this mutant. These hypotheses are consistent with the bioenergetics when *M. acetivorans* grows on CO as the sole energy source [296]. Both the acetogenic (acetate forming) and methanogenic (methane forming) branches of the methanogenesis pathway are energy yielding. Flux in the methanogenic branch results in a proton and sodium ion gradient which is then used to synthesize

ATP catalyzed by the proton translocating ATP synthase. Alternatively, flux through the acetogenic branch results in ATP synthesis using substrate level phosphorylation when acetyl phosphate is converted to acetate by acetate kinase. When Mcr is deleted there is no mechanism to recycle HS-CoM for another round of methylation and the Mtr-catalyzed methyl transfer reaction coupled to generation of the sodium gradient is also deactivated thereby diverting CH_3-THSPT towards synthesis of acetate and ATP. Therefore suppressing ACKr (or equivalently PTAr) in a mutant lacking Methyl coenzyme reductase (and consequently, the methane forming branched pathway) ensures that both energy yielding pathways are deactivated thereby halting growth.

In the second case (Figure B.4.1(B)), deleting ATP synthase results in a GNG mutant when the organism grows on methanol as the sole carbon and energy source [284]. This deletion negates proton- coupled generation of energy *via* methanogenesis leaving substrate level generation of energy *via* acetogenesis. GrowMatch suggests restoring consistency to this mutant by suppressing the sodium proton antiporter (abbreviation in *i*VS1007: Nat3_1). Suppressing this reaction in this mutant metabolic network deactivates flux in the sodium-dependent reaction methyl-THSPT:coenzyme M methyltransferase (abbreviation in *i*VS1007: MTSPCMMT) which results in no flux in the acetogenesis pathway (Figure B.4.1B)).

Step 4: Network connectivity analysis and restoration

After evaluating and improving the model using *in vivo* gene deletion data, we used the automated procedures GapFind and GapFill [220] to identify and rectify any remaining network connectivity inconsistencies. Using GapFind, we identify 95 metabolites (i.e., 12.2% of all metabolites in model) that cannot be produced for any choice of carbon substrate. Not surprisingly, none of the 95 no production metabolites were present in the methanogenesis pathway alluding to the completeness of its reconstruction. Interestingly, of the 161 metabolites present in the *M. acetivorans* model but absent in *i*AF692, 101 can be produced whereas 60 have blocked production routes. Notably, GapFind revealed that 35 out of these 95 metabolites could also not be produced in the *i*AF692 model of *M. barkeri.*

Flow restoration to all 95 metabolites was attempted using GapFill by adding reactions from KEGG [86]. In this step, we restored consistency to only 21 of the 95 no production metabolites. Flow through two of these 21 metabolites was restored by treating two existing reactions (cob(I)alamin-HBI adenosyltransferase and hydroxyethylthiazole kinase) as reversible. Flow through the remaining nineteen metabolites was restored by adding three reactions from the *i*AF692 model and thirteen reactions from the KEGG database. In accordance with the prescribed systematic cutoffs (see Methods section) reactions are added only when they have e-value lower than 10^{-10} against the *M. acetivorans* genome.

Model characteristics for *i*VS1007

Table B.4.2 contrasts the model statistics for the *i*VS1007 model against previously constructed archaeal models. The *i*VS1007 model is characterized by a large number of entries with high confidence scores due to the stringent cutoffs prescribed at each step. Furthermore, the inclusion of seven regulatory constraints that allow for substrate specific activation of methyltransferases and the addition of reactions using multiple pieces of evidence are unique features of this model. Finally, in contrast to the remaining models, the *i*VS1007 model documents global and conditional suppressions based on systematic evaluation of model predictions with *in vivo* growth data and network gap correction.

We compared flux values through the methane forming reaction catalyzed by Methyl Coenzyme Reductase and the biomass equation to ascertain the extent of coupling between flux in the methanogenesis pathway and *in silico* growth rates. We identified the range of methane

production flux by maximizing and minimizing flux through the MCR reaction for different values of biomass formation. Conversely, we identified the range of biomass production for different values of required methane production. Figure B.4.2 shows these plots for growth on methanol, acetate and carbon monoxide.

As shown in Figure B.4.2 (A) and (B), a positive biomass flux implies a non-zero MCR flux for growth in methanol and acetate but not the reverse. Using the terminology introduced in [26], this implies that the flux in biomass reaction is *directionally coupled* to the flux in MCR. This is consistent with the indispensability of the methanogenic branch when *M. acetivorans* grows on acetate and methanol [295, 303]. Moreover, the maximum biomass formation is reached at when the flux through MCR is fixed at 74% of its maximum value for growth on methanol and 86% for growth on methanol. At maximum biomass production, the ratio of biomass to methane production is 0.016 GDW/mmol and 0.005 GDW/mmol for growth on methanol and acetate, respectively. This higher biomass yield is qualitatively consistent with the higher energetic yield per mole of methanol observed for *M. acetivorans* [304].

Figure B.4.2(C) illustrates the predictions of the *i*VS1007 model for growth on carbon monoxide as the sole carbon and energy substrate. The model prediction that the methanogenic branch is dispensable when *M. acetivorans* grows on carbon monoxide is consistent with the mechanism proposed in [291, 296]. Notably, the maximum biomass production is achieved at 58% of the maximum flux in the MCR reaction and the ratio of the two fluxes is 0.033 GDW/mmol. It has been previously established that the acetogenic and methanogenic branches of the pathway are energy yielding when *M. acetivorans* grows on carbon monoxide [296]. Using the coupling analysis described above, we find that the acetogenic and methanogenic branches are *not* coupled This supports the independence of the energy yielding branches for growth on carbon monoxide.

B.4.4 Summary

Metabolic reconstruction technology has been used extensively to document the metabolic fingerprints of organisms in the *Bacteria* and the *Eukarya* domains [305]. Here, we take advantage of the increased availability of experimental and -omics datasets for archaeal organisms to build the metabolic model, called *i*VS1007, of the archaeon with the largest known genome, *Methanosarcina acetivorans*. The *i*VS1007 model is constructed using a systematic procedure that enables sequential evaluation and improvement of model capabilities. The model consists of 835 reactions, 790 metabolites and 1007 genes; the latter accounting for 45% of all ORFs in *M. acetivorans* with putative annotations [297]. The completed model has 716 metabolites (91%) that can be produced and it has a high agreement of 91% against *in vivo* growth data across environmental and genetic perturbations with specificity of 74% (i.e., percent of correctly identified essential genes) and selectivity of 86% (i.e., percent of correctly identified non-essential genes). Additionally, the model recapitulates substrate-specific energetic characteristics such as ATP synthase indispensability for growth on acetate/methanol and its dispensability for growth on carbon monoxide.

The number of reactions included in the draft model under Step 1 is quite sensitive to the adopted BLAST cutoff. The number of reaction entries increases to 1,090 when the cutoff is relaxed to 10^{-20} from the 820 entries for the adopted cutoff of 10^{-30}. This more stringent cutoff was chosen to ensure that the draft model did not contain any falsely added reactions. We have found that it is much easier to find and add missing functionalities than correctly identifying and removing erroneous ones. Interestingly, all but one reaction in the methanogenesis pathway known to occur in *M. acetivorans* were included in the draft model using the most stringent cutoff. Reaction ECH Hydrogenase which is known to occur in *M. barkeri* but not in *M. acetivorans* was excluded from the draft model.

This constructed *i*VS1007 model represents the most comprehensive up-to-date effort to catalogue methanogenic metabolism. Given the attention methanogenic consortia have received and the growing amount of metagenomic data [306], this model can be used to assess the biological impact on carbon balance of methanogenic communities. This organism-specific compilation of the metabolic repertoire of *M. acetivorans* can serve as the framework for fusing additional experimental data on methanogens as they become available.

B.4.4 Materials and Methods

B.4.4.1 Generation of initial model

We generate the initial model for *M. acetivorans* by taking advantage of an existing genome-scale metabolic model for the marine methanogen *M. barkeri* (*i*AF692). The *i*AF692 model is based on a draft version the *M. barkeri fusaro* genome [74]. We first mapped the genes from *i*AF692 onto the current genome-sequence of *M. barkeri* to restore consistency with the most up-to-date genomic information. For every gene in the *i*AF692 model, we retrieved the corresponding protein sequence (personal communication with Adam Feist of UCSD) and conducted bidirectional BLAST (BBH) (BLASTp [307]) searches against the current genome sequence of *M. barkeri*. This mapping is available in the submitted paper.

The draft reconstruction for *M. acetivorans* is generated by conducting bidirectional BLAST (BLASTp) searches for each one of the 692 genes in *i*AF692 against its genome and including only those genes/protein/reaction associations with an e-value of better than 10^{-30}. We used multiple sources to annotate the remaining genes in *M. acetivorans*. Specifically, we incorporated in the following order updated annotations made available as part of an ongoing effort at the University of Maryland (carried out in the Sowers Lab at the Center for Marine Biotechnology), extracted from the SEED database [300], and ones available at TIGR [297].

Upon obtaining annotations for the remaining genes, we pinpointed metabolic genes by searching each annotation against the KEGG ligand [86] database and retrieving corresponding reactions. For annotations with no synonyms in the KEGG ligand database, we use their Enzyme Commission Number (if available) to search the Swiss- Prot database [308] and retrieve the metabolic reaction(s) (if at all) they are associated with. Finally, we also included reactions that are known to be present in *M. acetivorans* but absent in *M. barkeri* (e.g., reactions for CO metabolism. We use the AUTOGRAPH procedure developed by Notebaard et al., to generate the gene-protein-reaction (GPR) associations [39]. This procedure uses bidirectional BLAST hits (BBH) to generate GPR's for new metabolic reconstructions (*M. acetivorans* in our case) using the GPR's of related metabolic models (*M. barkeri*). We also added non-gene associated reactions and exchange reactions in *i*AF692 to the model unless we found evidence to contrary.

B.4.4.2 Model fidelity improvement using available data sources

Upon the generation of the draft model the next step involves the systematic elimination of network gaps using GapFind/GapFill [220] and growth prediction inconsistencies using GrowMatch [152]. These procedures are deployed in a synergistic manner to provide mutually corroborating evidence for model correction.

Step 1: We generate the draft model as discussed above.

Step 2: We test the ability of the model to grow on known substrates. If it doesnt, we use modified versions of GapFind and GapFIll respectively to identify biomass precursors that cannot be produced and ensure their production. We allow for addition of functionalities at this step only if the BLAST e-value is lower than 10^{-2}. Upon completion of this step all biomass components are available in *i*VS1007.

Step 3: We compare *in silico* biomass production in *i*VS1007 against available *in vivo* growth data across different environmental/genetic perturbations. Mutants are classified as Grow/Grow (GG), No-Grow/Grow (NGG), Grow/No-Grow (GNG) and No-Grow/No-Grow (NGNG) following the definitions proposed in [152]. GNG mutants are resolved by identifying global/conditional suppressions in the *i*VS1007 network using GrowMatch. NGG mutants are corrected by globally or conditionally adding reactions in *i*VS1007 using GrowMatch. These reactions are preferentially chosen from model *i*AF692 followed by additions from external databases such as KEGG [86] using a BBH e-value cutoff of 10^{-10}. Upon completion of this step, all *in silico*/ *in vivo* growth inconsistencies that could be corrected by either removing or adding reactions available in databases resolved.

Step 4: We next identified metabolites that cannot be produced or consumed using GapFind. Using GapFill, we restore connectivity to them by appending only reactions that have BBH e-values of less than 10^{-10}.

In addition, we mined for all published articles having the word "Acetivorans" anywhere in their content in the Web of Science and PubMed databases and download these articles using EndNote^{Web}. We used the mdfind command on a MacBook ^{TM} , search for articles that have loci-names of *M. acetivorans* genes included in the *i*VS1007 Model. This enables a relatively quick search for literature evidence supporting (or not) annotations in the *i*VS1007 Model. We update the model to resolve any incorrect annotations identified in this step and consolidate information from articles not included in the above search domain but have information regarding methanogenesis [301].

Figure 1B

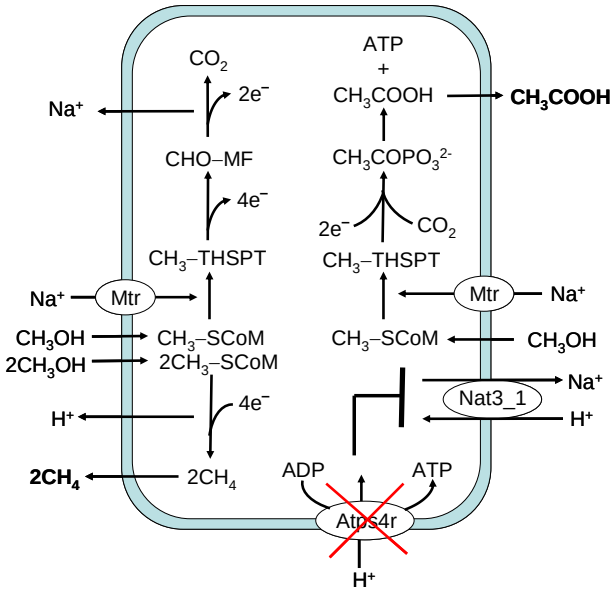

Figure B.4.1. A) GrowMatch resolution of the GNG mutant characterized by deleting Methyl Coenzyme Reductase with carbon monoxide as the carbon source. B) GrowMatch's resolution of the GNG mutant characterized by deleting ATP Synthase with methanol as the carbon source.

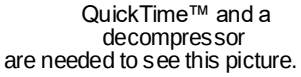

Figure B.4.2 S Flux coupling analysis between flux in Methyl coenzyme reductase and biomass for cellular growth on A) methanol, B) acetate, C) carbon monoxide. All values of fluxes are in mmol/gDW hr^{-1} and are normalized to the respective substrate uptake rates fixed at 1000 mmol/gDW hr.

Table B.4.1 *In vivo* gene deletion data used evaluate and improve iVS1007's predictive capabilities.

	Substrate					
Gene deletions	**acetate**	**carbon monoxide**	**methanol**	**MMA**	**DMA**	**TMA**
ackR	NGNG	**GNG**	GG	-	-	-
ATP synthase	NGNG		**GNG**	-	-	-
DMTsD	GG	GG	GG	-	-	GG
mtsD+mtsF	GG	GG	GG	-	-	GG
mtsD+mtsH	GG	GG	GG	-	-	GG
mtsF	GG	GG	GG	-	-	GG
mtsH	GG	GG	GG	-	-	GG
mtsF+mtsH	GG	GG	GG	-	-	GG
lysK	-	-	GG	-	GG	GG
lysS	-	-	GG	GG	GG	GG
MCR	NGNG	**GNG**	NGNG	**GNG**	**GNG**	**GNG**
mtaA1	-	-	NGNG		-	-
mtaA1 + MT1	**GNG**	-	-	-	-	-
mtaA2	-	-	GG	-	-	-
mtaCB1	-	-	GG	-	-	-
mtaCB1 + mtaCB2	-	-	GG	-	-	-
mtaCB1 + mtaCB2 + mtaCB3	-	-	**GNG**	-	-	-
mtaCB2	-	-	GG	-	-	-
mtaCB3	-	-	GG	-	-	-
mtbA	-	-	-	NGNG	NGNG	-
mtbA	-	-	-	-	-	GG
ppylR	-	-	**GNG**	-	-	-
ppylT	GG	-	GG	**GNG**	**GNG**	**GNG**
ptaR	NGNG	**GNG**	**GNG**	-	-	-
Rnf complex	**GNG**	-	-	-	-	-

Table B.4.2 Comparison between *i*VS1007 and other available Archaeal models

	Methanosarcina acetivorans	***Methanosarcina barkeri***	***Halo***
Genome size	5.7 Mb	4.8 Mb	
ORF's	1007	5072	
Metabolic genes	1007	692	
Unique proteins	1007	542	
Isozymes	140	31	
Multidomain proteins	145	65	
Reactions	835	619	
gene-associated	726	509	
non gene-associated	109	110	
transport reactions		88	
Metabolites	790	558	
Gaps		35	
Consistency with growth data	91%	69%	-

C. Specific Aim 2: Automated Generation of Genome-Scale Isotope Mapping Models

C.1 Automated generation of complete atom mappings for MFA using genome-scale metabolic reconstructions

The work in this section was recently published [309].

C. 1.1 Introduction

Metabolic flux analysis (MFA) [310] has emerged as a critical tool to understand the physiological state of a cell [311-313]. Using isotopically labeled substrates with different labeling patterns, experimental techniques such as NMR [314, 315] and GC-MS [316] are used to measure the amounts of different isotope forms of select metabolites. The fluxes in a metabolic network are directly coupled to the relative isotopic abundances of different metabolites through a system of nonlinear algebraic equations [317]. Details of the same can be found in literature in a recent review [318]. Briefly, these nonlinear equations are constructed using mapping matrices that trace the path of each atom and subsequently each isotopomer (isotope isomer) in a metabolic reaction. This information was initially represented using atom mapping matrices (AMM) [319] that track the transfer of carbon atoms from reactants to products. This concept was subsequently generalized in the form of isotopomer mapping matrices (IMM) [320] that enumerate all possible product isotopomers that can be created from each reactant isotopomer.

Two separate computational challenges arise during flux elucidation based on MFA. The first challenge involves the automated generation of isotope mapping matrices for genome-scale metabolic reconstructions while the second involves the efficient solution of the corresponding system of nonlinear equations for the unknown fluxes while accounting for measurement error. The challenge of flux elucidation has been previously addressed using a variety of computational techniques including the cumomer concept [321], theoretical bondomer [322], the elemental metabolic unit (EMU) framework [323] and FluxCalc [324]. However, the application of these methods has been restricted to models that were at least an order of magnitude smaller than genome-scale reconstructions as a consequence of the aforementioned challenges. Typical isotope mapping models contain 25-50 reactions [318], 76 reactions [325] or 238 reactions [324], which is the largest to-date model (developed in our group). A key shortcoming of using lumped metabolic abstractions to perform flux elucidation is that they may erroneously lead to the conclusion that the available GC or MS data is sufficient for unique flux elucidation [326]. The inferred metabolic fluxes may then inherently reflect the biases/assumptions built-in during the lumped metabolic map creation step. In addition, by utilizing a genome-scale model for simulation/strain design purposes and a separate lumped metabolic model for flux elucidation could complicate the seamless integration/transfer of results.

Motivated by these shortcomings, here we introduce a largely automated workflow for constructing isotope mapping matrices using as input full genome-scale metabolic reconstructions. This is a formidable challenge as it requires a detailed mapping of atom transitions for all reactions in a metabolic network and has so far remained organism-specific and labor intensive. Atom mapping matrices are obtained by tracing the origin and destination of atoms through each individual reaction in the metabolic network. In addition to tracing isotopically labeled carbon atoms (typically preferred in MFA experiments) we also trace the path of O, N, P, S atoms as well as of metal/non-metal ions. Tracing atoms from reactants to products requires the ability to topologically superimpose the structures of reactant and product molecules. This involves the identification of all “common” substructures between the two molecules.

Techniques relying on pattern recognition concepts from graph theory, which have been extensively employed in cheminformatics [327-329], can be used to topologically align and compare a reactant with a product molecule. These techniques essentially apply two mathematical operations on the molecular graphs of the two compounds to be aligned. A molecular graph is a mathematical representation in which nodes correspond to atoms and edges to molecular bonds. The first mathematical operation combines the two molecular graphs into a single association graph (AG). The second operation identifies the largest clique (i.e., connected graph) within the AG. In the maximum common edge subgraph (MCES) procedure, the edges of the AG incorporate bond-type information (single, double or triple bond) of the compared molecules. [330] adopted the MCES approach to match two randomly chosen (not part of a biochemical

reaction) structurally complex chemical compounds with reasonable accuracy. Unfortunately, the MCES method does not scale well for genome-scale level reaction compilations requiring prohibitive computational time [331].

The maximum common subgraph (MCS) approach [332], formulates the edges of the AG without considering the bond-type data of the two compounds involved. As a consequence, the MCS approach is more computationally efficient and thus more suitable for mapping atoms participating in a large number of reactions [333]. In addition, the accuracy of atom mappings produced through this procedure is quite high for most biochemical reactions. However, to date it has only been used to contrast pairs of compounds [333] or trace only carbons [334] within the KEGG/LIGAND database [335, 336]. Also, the atom transitions listed in KEGG are inadequate for flux analysis using MFA since alternative atom transitions are not explicitly listed when symmetric molecular sub-structures or symmetric molecules are present in the reaction. Alternatively, compound matching based on an algorithm that tallies the connectivity (i.e. number of atoms connected to a given atom) of atoms in the compared compounds [337], has been used to trace atoms across reactions [338, 339]. However, this procedure requires the manual reordering of metabolites in reactions and has scaling limitations (i.e., it cannot detect rings of size greater than ten such as heme) [340].

We chose to overcome these limitations and generate mappings for the latest metabolic reconstruction of *E. coli* [49] by first representing molecular chemical structures as graphs defined by a set of vertices (the atoms) connected by edges (the bonds). Subsequently, the MCS method [333] coupled with a modified branch and bound algorithm for clique finding [341] is customized to automatically generate genome scale atom mappings.

In the next section we describe in detail the generated isotope mapping model imPR90068 for the *E. coli* strain K-12, which spans 1,039 metabolites, 2,077 reactions and contains a total of 1.37 x 10^{157} isotopomers and 8.34 x 10^{93} ^{13}C isotopomers. Furthermore, we highlight the enhanced pathway resolution capability of imPR90068 with an emphasis on nucleotide salvage, cofactor and prosthetic group biosynthesis, glycerophospholipid metabolism and alternate carbon metabolism. We also provide guidelines on the application of the model for MFA using an EMU representation. Finally, we describe the general procedure developed for the largely automated generation of genome-scale atom mappings.

C.1.2 Results and Discussion

We first highlight the adopted molecular graph based description by reactant to product atom mapping for two separate example reactions. Next, the size and content statistics of imPR90068 are reviewed followed by the application of the model for flux elucidation.

C.1.2.1 Reaction to product atom mapping examples

The metabolic network *i*AF1260 contains 304 exchange reactions, 690 transport reactions and 1,387 metabolic reactions [49]. The atom mappings for the 690 transport reactions, which account for 12,325 of the traced atoms, were generated in a straightforward manner as the molecular graphs remain invariant upon transport. For example, the atom mappings for the arginine/agmatine antiport reaction, which is a reversible inner membrane transport reaction, are retained as arginine and agmatine as they simply transported from the cytosol to the periplasmic space without any bond modifications:

ARGAGMt7pp agm[c] + arg-L[p] <==> agm[p] + arg-L[c]

The atom mappings for the remaining 1,387 metabolic reactions, containing 77,619 of the mapped atoms, were created by iteratively applying for every reaction the proposed workflow (see Steps 1-4 in Figure C.1.1). During this process, five frequently occurring reaction motifs

were automatically identified and stored in a database (see Table C.1.1). The atom mappings of these five reaction motifs, which occur in 424 different reactions, were simply copied from the reaction motif library (Table C.1.1).

The following two example reactions illustrate the results obtained upon applying the four steps of the mapping procedure (see Figure C.1.1). The first reaction is histidinol-phosphatase, which is part of histidine metabolism and listed in *i*AF1260 as:

HISTP [c] : h2o + hisp --> histd + pi

In Step 1, the reaction is parsed into reactants, products and the reaction name (i.e. HISTP). In Step 2, atom and bond information of water, L-histidinol phosphate, L-histidinol and phosphate molecules are obtained from respective MDL mol files of *i*AF1260. These data are used to create the reactant and product graphs of histidinol-phosphatase reaction (Figure C.1.2). The reactant and product graphs contain fifteen nodes linked through fourteen edges. Each node is associated with two parameters, one representing the atom number and the other representing the atom type. Each edge is associated with three parameters, the pair of atoms they connect and the bond type (Figure C.1.2). The corresponding chemical structures of the metabolites represented in the reactant and product graphs are also shown in Figure C.1.2. In Step 3, the six carbon, one phosphorous, five oxygen and three nitrogen atoms of water and L-histidinol phosphate are traced to atoms in phosphate and L-histidinol through 24 alternate atom mappings (grey atom traces in Figure C.1.2). In Step 4, known reaction chemistry information is used to prevent the oxygen atom in the water molecule water from being traced to the phosphate moiety (dephosphorylation reaction chemistry). Due to the interchangeability of all oxygen atoms in the phosphate group, 24 separate atom mappings are generated and stored as one *reaction mapping* under the reaction name HISTP.

Symmetric molecules introduce a number of additional complications. They are illustrated using the taurine dioxygenase reaction:

TAUDO [c] : akg + o2 + taur --> aacald + co2 + h + so3 + succ

After parsing the reaction into the required components, MDL mol files of 2-oxoglutarate, oxygen, taurine, succinate, carbon dioxide, aminoacetaldehyde and sulphite are used to create the reactant and product graphs (Figure C.1.3). Due to the presence of symmetric molecules (i.e., succinate, carbon dioxide and oxygen) and interchangeable atoms within groups (i.e., sulphite), 96 alternate mappings are generated for the taurine dioxygenase reaction. These atom mappings, which trace seven carbon, one nitrogen, ten oxygen and one sulfur atoms between seven metabolites (Figure C.1.3), are stored as a reaction mapping under the reaction name, TAUDO. A manual curation of the generated 96 atom mappings reveals that there are no erroneous mappings implying that symmetry is properly handled by our mapping procedure.

C.1.2.2 Size statistics and content of imPR90068

The genome-scale mapping model imPR90068 generated for the *E. coli* encodes the complete list of reactions in *i*AF1260 (Feist et al., 2007) as a library of 2,077 reaction mappings (see supplemental information for the mapping files). Each reaction mapping contains multiple atom mappings that trace all reactant atoms to all product atoms in the respective reaction. The model contains a total of 20,872 alternate atom mappings that trace the fate of 90,068 atoms through a network of 2,077 reactions and 1,039 metabolites. These atom mappings trace the path of C, O, N, P, S atoms as well as Ag, As, Ca, Cd, Cl, Co, Cu, halogens, Fe, Hg, K, Mg, Mn, Na, Ni, Se, W, Zn ions. Detailed information on atoms traced is provided in Table C.1.2.

The classification of all 1,387 metabolic reactions in imPR90068 based on the number of alternative mappings (per reaction) is shown in Table C.1.4 and Figure C.1.4. The reaction

mappings of 734 reactions contain a single alternative, which implies that the atoms in these reactions are uniquely mapped from reactants to products. The majority of these 734 reactions with no mapping degeneracy are isomerization, displacement or substitution reactions typically containing less than three reacting species. The remaining reaction mappings are degenerate to various degrees and contain multiple alternative atom transitions from reactants to products due to symmetry(ies) present in the reaction operator (Table C.1.4). A general downward trend is observed in the number of reactions with increasing reaction mapping degeneracy with 578, 256, 155 reactions containing respectively 2-8, 9-128, 129-1024 alternative mappings (Figure C.1.4). A certain obvious clustering of reactions is observed at 17-32 alternatives and similarly at 257-512 alternatives. This happens due to the nature of reactive groups participating in individual reactions such as phosphate, which typically produces 24 alternatives or diphosphate, which typically results in 288 alternatives. Table C.1.4 also identifies which atom type (or combination of atoms) is responsible for the degeneracy in the mapping. The individual reactions containing a modest number of mappings (i.e., from two to eight) are primarily degenerate either due to equivalent carbons or due to equivalent oxygens and less likely due to the presence of both equivalent carbons and oxygens (71% due to either only C or only O and 22% due to both C and O). The degenerate reactions containing equivalent O (either standalone or in combination with other equivalent atoms such as C, N) are predominantly due to electronic orbital resonance of the oxygen atoms in the carboxyl groups [342]. Degeneracy due to only equivalent C and only N arise as a result of backbone symmetry of the reacting species (see also Figure 4). For example, in reaction TAUDO (see Figure C.1.3), reactant 2-oxoglutarate can be mapped to product succinate in four possible ways. This multiplicity arises from two equivalent carbon atom pairs 1,2 and 8,5 in 2-oxogluterate that can be mapped to either 17,16 or 12,13 positions, respectively present in the succinate product molecule.

Surprisingly, despite the presence of nearly 60% less number of oxygen atoms in the model than carbons, we find that equivalent oxygen atoms are by far the most frequently occurring (resulting in 44% of all degenerate reactions) whereas C atoms result only in 28% of total reaction degeneracy. The reaction HISTP (see Figure C.1.2) illustrates the reason for the above statistics, where all 24 alternative mappings are due to 4 oxygen atoms (3,4,5 and 6 in the reactant graph) although a greater number of carbon atoms are present in the graph. Often, multiple atoms (e.g., C, O, N or P) simultaneously contribute in the mapping degeneracy. Fairly ubiquitous are reactions with multiple mappings arising from both C and O atoms. For example, in the citrate hydro-lyase reaction of TCA cycle, both carbon and oxygen atoms in the symmetric citrate molecule are mapped in multiple ways to the product cis-aconitate molecule.

Phosphorous atoms accompanied by equivalent oxygen atoms (due to the presence of resonating phosphate groups) are involved in reactions with large numbers of mappings (i.e., more than 64). There exist ten reactions with number of mappings in the range of 513-1,024. These reactions contain four or more reacting molecules usually with multiple symmetric operators and are involved in cofactor and prosthetic group biosynthesis, murein recycling and nucleotide synthesis/salvage pathways. For example, in the asparigine synthetase reaction ASNS2, six molecules containing five reaction operators (two carboxyl groups and three phosphate groups) bring the reaction mapping degeneracy to 864 alternatives.

C.1.2.3 New reactions/metabolites in imPR90068

The introduced isotope mapping model imPR90068 contains mappings for reactions that were previously lumped or completely absent from isotope mapping models (even in imPS1485). These new additions include 68 reactions involved in the metabolism of 17 different amino acids (all but Asparagine, Glutamine and Glutamic acid), 65 reactions involved in central metabolism, 153 reactions in nucleotide biosynthesis and salvage pathways, 225 reactions in glycerophospholipid metabolism, 160 reactions in cofactor and prosthetic group biosynthesis and

181 reactions in alternate carbon metabolism. The inclusion of all biotransformations spanned by the genome-scale model implies that alternate metabolic routes can now fully be taken into account during flux elucidation using MFA. For example, in imPR90068, the xylose isomerase catalyzed reaction XYLI2 that reversibly isomerizes D-glucose to D-fructose combined with the fructose transport reaction FRUpts2pp which converts PEP (phosphoenolpyruvate) to pyruvate during the transport of D-fructose, creates a pathway from glucose to pyruvate alternate to glycolysis. Similarly, reactions such as the amylomaltase (AMALT1-4), maltodextrin glucosidase (MLTG1-5), a- and b-galactosidase (GALS3, LACZ, LACZpp) reactions of the alternate carbon metabolism pathway, which involve alternate routes for glucose metabolism, are included in the genome-scale model. Further, analysis under growth on 174 carbon sources is possible in the imPR90068 model as opposed to growth on glucose and few amino acids studied in imPS1485. In addition, all biomass components are mapped in imPR90068 model. As many as 45 biomass components absent from imPS1485 are now part of the model. These metabolites include cofactors (e.g., CoA), amino acids (e.g., His and Trp), riboflavin, murein, and inorganic ions (e.g., Fe+3). It is important to note that new reactions in imPR90068 are not necessarily far away from central metabolism. Even under aerobic glucose growth conditions, as many as 35 new reactions are added to central metabolism that are part of Citric Acid Cycle, Glycolysis/Gluconeogenesis, Oxidative Phosphorylation, Pentose Phosphate Pathway and Pyruvate Metabolism.

Notably, imPR90068 accounts for not only all reactions but also all metabolites present in iAF1260. 76 new metabolites are present in imPR90068 that were absent in imPS1485. These newly added metabolites link parts of metabolism previously treated before as separate. For example, (see Figure C.1.5) the added metabolite aicar (5-Amino-1-(5-Phospho-D-ribosyl)imidazole-4-carboxamide) directly participates in purine meabolism and the histidine pathway. It is also indirectly linked to thiamine metabolism (through metabolite air), glycine, serine and threonine metabolism (through glycine) and in alanine, aspartate and glutamate metabolism (through glutamate). Thus, the incorporation of a single additional metabolite in imPR90068 enables for the first time the ability to fully describe histidine and purine metabolism as well as account for interactions between many pathways.

C.1.2.3 Reduced and EMU based representation of imPR90068

Armed with a complete database of all atom mappings implied by the genome-scale model *i*AF1260, it is straightforward to select only the mappings which are relevant for a given isotope labeling experiment. The numbers of isotopomers present upon labeling various atoms present in the model are detailed in Table C.1.2. For example, by labeling only carbons we find that the 932 carbon-containing metabolites (with a total of 20,935 carbon atoms) yield 8.34 x 10^{93} ^{13}C isotopomers. We can tailor the set of considered isotopomers to the specifics of the system under consideration by removing all reactions/mappings that are suppressed under the experimental conditions. For example, under aerobic glucose minimal media conditions 752 blocked/suppressed reactions can be removed from the model leaving 793 metabolites containing 33,026 tractable carbon atoms and 3.02 x 10^{62} isotopomers.

An even more compact representation of the isotope mapping relations can be achieved using the EMU representation [323]. We have developed Python scripts that given the atom mapping matrices of imPR90068, the labeled substrate and measured fragments the EMU representation is automatically generated. The EMU representation of imPR90068 for aerobic labeled glucose minimal media conditions and using the 31 amino acid fragments listed in Table C.1.1 of [324] is provided as supplemental material. Table C.1.3 highlights the savings afforded by the EMU representation. While the 17,346 carbon isotopomers of imPS1485 are reduced to 1,215 EMU species and 3,912 mass isotopomers (Suthers et. al., 2009), the 10^{93} carbon isotopomers in

imPR90068 are reduced to 1,067,652 EMU species and 6,065,801 mass isotopomers. This is a tractable model size that can be handled by current solvers such as CPLEX 10 [343].

C.1.3 Summary

This work introduced the computational infrastructure for tracing all atoms present in every reaction in the *i*AF1260 metabolic reconstruction of *E. coli* from reactants to products to create a genome-scale mapping database. This automated procedure can be efficiently leveraged for genome-scale models of other organisms to create isotope mapping databases. Common reactions already present in *i*AF1260 can be directly culled from the imPR90068 reaction-mappings database thus significantly reducing the effort needed to construct other organism-specific mapping models. The potential to improve our understanding of flux allocation in different organisms is alluded by the gap in the size of genome scale vs. isotope mapping models. For example, there exists a 50-fold difference in the size of the genome-scale reconstruction of *Bacilus subtilis* that spans 1,020 reactions [68] and its current isotope mapping model [344] that accounts for only 25 reactions (all from central metabolism). It is expected that incorporating reactions into the mapping model already present in the genome-scale model could shed light onto metabolic pathway usage patterns with many practical implications, for example for an industrially relevant organism such as *B. subtillis*.

The ability to elucidate fluxes using the full complement of reactions and metabolites present in genome-scale level reconstructions comes at the expense of requiring additional labeling data. While lumped isotope models [318, 324, 325] typically require the analysis of spectra (i.e., NMR or GC/MS) for only about 20-50 fragments, using the totality of mapped isotopomers in imPR90068 will likely require significantly higher numbers of carefully chosen labeled fragments. This makes even more pertinent the use of methods such as OptMeas [326]; Suthers et al. 2009) to pinpoint minimal measurement sets and compact isotope representations such as EMU [323] for complete flux elucidation.

Finally, the use of molecular graph representations at a genome-scale level can be used to study the synthesis problem in metabolic networks [89]. An example application is in creating specific chemistry operations for the computational framework BNICE [89]. BNICE generates novel biochemical pathways and novel intermediate compounds given the bond-electron matrix (BEM) of the initial metabolites and a single or combination of reaction operators for each reaction in the pathway [89, 345]. The BEM specifies compound properties: the non-bonded valance electrons of all atoms in participating molecules and the connectivity, bond order (single, double bond etc.) between those atoms. On the other hand, the reaction operators used in BNICE are biotransformation rules that have been generalized based on EC reaction classification [346]. The molecular graph approach used to create the isotope mapping model specifies complete reaction rules for genome scale networks in the form of reaction mappings. Hence the data available in mapping files can be used to generate reaction operators required by BNICE for analysis of genome-scale networks. The metabolic network of *E.coli* (and eventually other organisms) can potentially be explored for hypothetical reaction steps which include novel intermediate/product metabolites with relative ease due to the availability of a genome-scale isotope mapping model [89].

C.1.4 Materials and Methods

The proposed procedure used to generate imPR90068, requires as input the stoichiometry of all reactions present in the metabolic network and data (e.g., MDL mol files) encoding the chemical structure of all metabolites involved in the network. The method described below can be applied to any genome-scale metabolic model and is amenable to the straightforward inclusion of additional reactions not present in the original organism models as well as user-supplied metabolite structures. During the automated procedure, a library of atom mappings and recurring

motifs is generated which can be leveraged for future isotope mapping efforts. The following four steps are performed on every reaction in the input network (see Figure C.1.1).

Step 1: Automated identification of metabolites with constant labeling and elucidation of recurring reaction motifs

The reaction stoichiometry, supplied as part of the input network, is parsed into reactants and products. Reaction stoichiometry is appropriately handled by accounting for multiple or partial occurrences of metabolites. Exchange reactions (i.e. a reaction in which the metabolite crosses the system boundary) are handled in a straightforward manner as labeling remains unaffected during transport. Similarly, reactions for which the same metabolite is present on both the reactant and product side are identified and the corresponding metabolite's labeling is flagged as identical for both compartments (Fig 1). A number of reaction motifs occurring in many biotransformations (such as atp + h_2o → adp + h + pi) are identified and their atom mappings are stored in a library (see Table C.1.1). Therefore, when parsing a particular reaction, metabolites that remain unaltered and metabolites identified as part of a reaction motif are temporarily removed before the molecular graph comparison step.

Step 2: Generation of reactant and product molecular graphs

Atom and bond information for all reactants and products in the reduced reaction is extracted from molecular geometry descriptors supplied as input data (i.e., MDL mol files). The chemical structure of each metabolite is represented as a graph where nodes depict atoms and edges refer to chemical bonds. The graphs of all reactants participating in a reaction are concatenated together (by combining atom and bond data of the individual metabolites) to yield a single *reactant graph*. Similarly, all the product molecule graphs are pooled to yield a single *product graph* (see Fig. 1). Note that reactant or product graphs are disjoint when multiple reactants or products are present, respectively. As a test, the total number of each atom type in the reactant graph is verified to be equal to that in the product graph.

Step 3: Construction of atom mappings between reactant and product graphs

We use the MCS method [333] to create the association graph AG between the reactant and product graphs. Subsequently, the branch and bound algorithm [341] is applied to detect the largest clique(s) in the AG. The largest clique corresponds to the largest subgraph (subset of nodes connected by the same set of bonds) shared between the reactant and the product graphs. In chemistry terms, this is the largest portion of the reacting molecules that remains invariant through the reaction step. The largest clique(s) encode the required mapping data for the current reaction (Fig. 1). The atom mappings of the metabolites with fixed labeling and those participating in reaction motifs (see Step 1) are generated and re-incorporated into the atom mapping database entry for the current reaction.

Step 4: Elucidation of consistent mappings

The MCS procedure often generates multiple atom mappings between reactant and product graphs (Fig 1). Alternate mappings are generated mainly due to the presence of many identical atoms within similar subgraphs between the reactant and product molecules. For example, the two oxygen atoms in a carboxyl group could, in principle, be routed in the same location in the product molecule. In addition, the presence of symmetric reactant metabolites (e.g., succinate) implies that positions equidistant from the middle are equivalent labeling choices in the product molecule. All the atom mappings obtained are verified to be correct by visually depicted the atom transition between the structures of reactant and product molecules (Fig. 1). If a particular atom transition is prohibited due to reaction chemistry, only the atom mappings permitted by the reaction heuristics are retained. One such example is a dephosphorylation reaction in which water molecule reacts with a phosphate-containing molecule thus displacing the phosphate group. Since

we know from reaction chemistry that the oxygen originating in the water molecule does not escape with the phosphate group, all such alternate mappings generated for dephosphorylation reactions can be automatically eliminated. Therefore, in some cases a post-processing step is needed to prune biologically irrelevant mappings. Using information of the reaction chemistry we retain only plausible mappings from the atom mapping file created in Step 3.

The end result of the atom mapping process is a library of atom mappings for every reaction in the input network. The procedure described above was used to create genome-scale atom mappings for the latest *E. coli* metabolic reconstruction *i*AF1260 [49]. Specifically, we constructed and used Python modules to extract atom and bond information for all 1,039 metabolites in iAF1260. This information was parsed from MDL mol files (whenever available) and from the KEGG [86] and the SDF PubChem databases. These data sets were used to create reactant and product graphs for all 2,077 reactions in iAF1260. The atom mappings were generated for each reaction separately using a cluster of Dell PowerEdge 1950 servers with dual 3.0 GHz Intel Xeon E5450 Quad-Core Processors and 32 GB of ECC RAM. Atom mappings were generated for every reaction in the network tracing all non-hydrogen elements including C, N, O, P, S and metal/non-metal ions. The obtained atom mappings were also manually curated as a final check.

The EMU representation [323] was implemented using Python modules. Briefly, given a set of mass isotopomer measurements and a set of source metabolites, this implementation calculates network fluxes through an EMU represention. The details of the procedure used to identify all EMU species and variables are outlined in (Suthers et al. 2009).

Reactant and product graph definitions1

The reactant and product graphs are defined by the following parameters.

G_1	Reactant graph
G_2	Product graph
$V(G_1)=\{u_i\}$	Vertices of G_1
$V(G_2)=\{v_i\}$	Vertices of G_2
$E(G_1)=\{(u_i, u_j)\}$	Edges of G_1
$E(G_2)=\{(v_i, v_j)\}$	Edges of G_2
$w(u_i)$	Atom type of vertex u_i
$w(u_i, u_j)$	Bond type between u_i and u_j

The maximum common subgraph: (MCS) approach

In the MCS approach, the association graph (AG) of G_1 and G_2 is defined by the set of vertices,

$V(AG)=\{(u_i, v_i)\}$, where $u_i \in V(G_1)$, $v_i \in V(G_2)$ and $w(u_i)=w(v_i)$

Two vertices (u_i, v_i) and (u_j, v_j) of the AG are connected whenever

$(u_i, u_j) \in E(G_1)$ and $(v_i, v_j) \in E(G_2)$

or

$(u_i, u_j) \notin E(G_1)$ and $(v_i, v_j) \notin E(G_2)$

This defines the edges of the AG.

Table C.1.1: List of frequently occurring reaction motifs

Reaction motif	# of occurrences in iAF1260	# of atoms mapped
atp + $h_2o \rightarrow$ adp + h + pi	162	32
atp + $h_2o \rightarrow$ amp + h + ppi	65	32
adp + $h_2o \rightarrow$ amp + h + pi	5	32
nad + h $\leftrightarrow$ nadh	110	44
nadp + h $\leftrightarrow$ nadph	82	48

Table C.1.2: Total number of most-prevalent atoms and their respective isotopomers

Atom type	Total # of atoms traced	Total # of isotopomers
Carbon	49,539	8.34×10^{93}
Oxygen	29,061	1.61×10^{60}
Phosphorous	3,280	1.00×10^{4}
Nitrogen	2,386	2.58×10^{7}
Sulfur	409	4.09×10^{3}
Others*	265	4.05×10^{3}
Total	90,068	1.37×10^{157}

* includes Ag, As, Ca, Cd, Cl, Co, Cu, halogens, Fe, Hg, K, Mg, Mn, Na, Ni, Se, W, Zn

Table C.1.3: Comparison of the sizes of imPS1485 and imPR90068 isotope mapping models of *E. coli*

		EMU model		EMU reduced model	
Isotope mapping model	^{13}C Isotopomers	EMU species	EMU mass isotopomers	EMU species	EMU mass isotopomers
Allowing for all uptakes with a transport mechanism					
imPR90068	8.3×10^{93}	1,067,652	6,065,801	621,311	2,786,978
imPS1485	17,346	1,215	3,912	762	2,438
Aerobic glucose minimal growth medium with all blocked reaction removed					
imPR90068	3.02×10^{62}	748,544	3,425,876	473,495	1,978,454
imPS1485	3,584	909	2,911	486	1,538

Table C.1.4: Distribution of metabolic reactions present in *im*PR90068 based on the number of alternate atom mappings of individual reactions. The break down of degenerate reactions with respect to equivalent carbons(C), oxygens(O), nitrogen(N) and phosphorous(P) are also shown.

Alternatives (Degeneracy)	Total # of reactions	# of reactions with equivalent C,O,N or P								
		C only	O only	N only	C,O	C,N	O,N	O,P	C,O,N	C,O,P
1	734									
2	232	138	105	4	30	3			1	0
3-4	117	17	48		41	2	2		4	1
5-8	179	41	66		58	1	2		7	3
9-16	71	2	31		33	1	1	1	1	0
17-32	121	9	68		31	1			4	7
33-64	35	1	14		14				3	2
65-128	29	0	16		9			2	1	1
129-256	19	0	5		6			8	0	0
257-512	126	1	107		4			3	3	1
513-1024	10	0	2		2			2	3	1

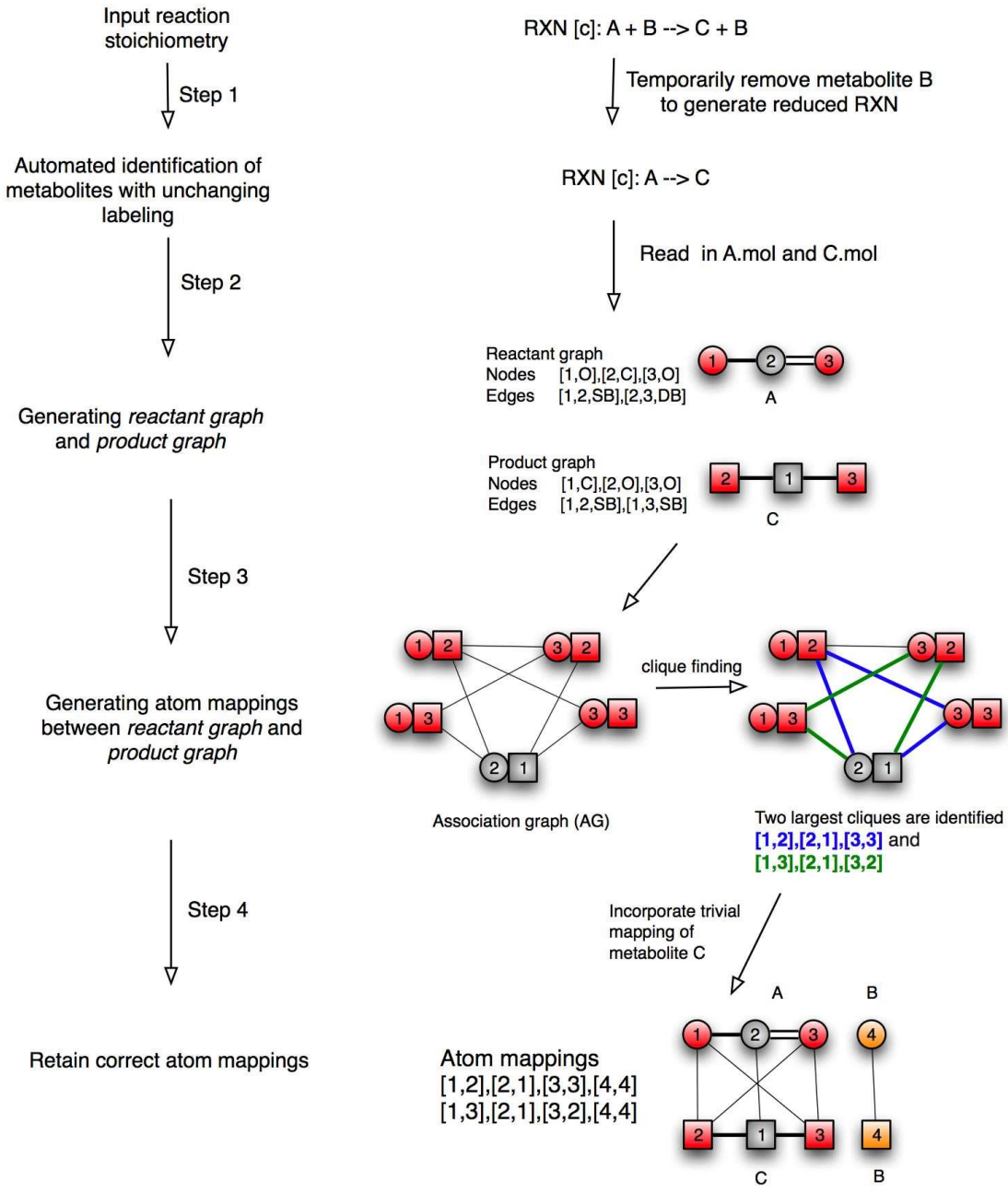

Figure C.1.1: Steps 1-4 are applied to a general reaction A + B --> C + B. The molecular structures of A, B and C are shown in Step 4. Grey circles and squares indicate carbon atoms (C), red denote oxygen (O) and orange map phosphorous (P) atoms. (Step 2) The atoms of reactant graph are shown as colored circles and that of product graph are shown as colored squares. (Step 3) The nodes of the AG are pairs of nodes from reactant and product graphs, and grey lines are the edges of the AG [see Appendix A for details]. The two cliques identified are the largest set of vertices that are completely connected to each other in the AG and are shown as thick green and blue lines respectively. (Step 4) The atom mappings are shown as lines (atom traces) between reactant and product molecular structures. From the visual representation we see that two alternate mappings exist due to symmetry of A and C molecules.

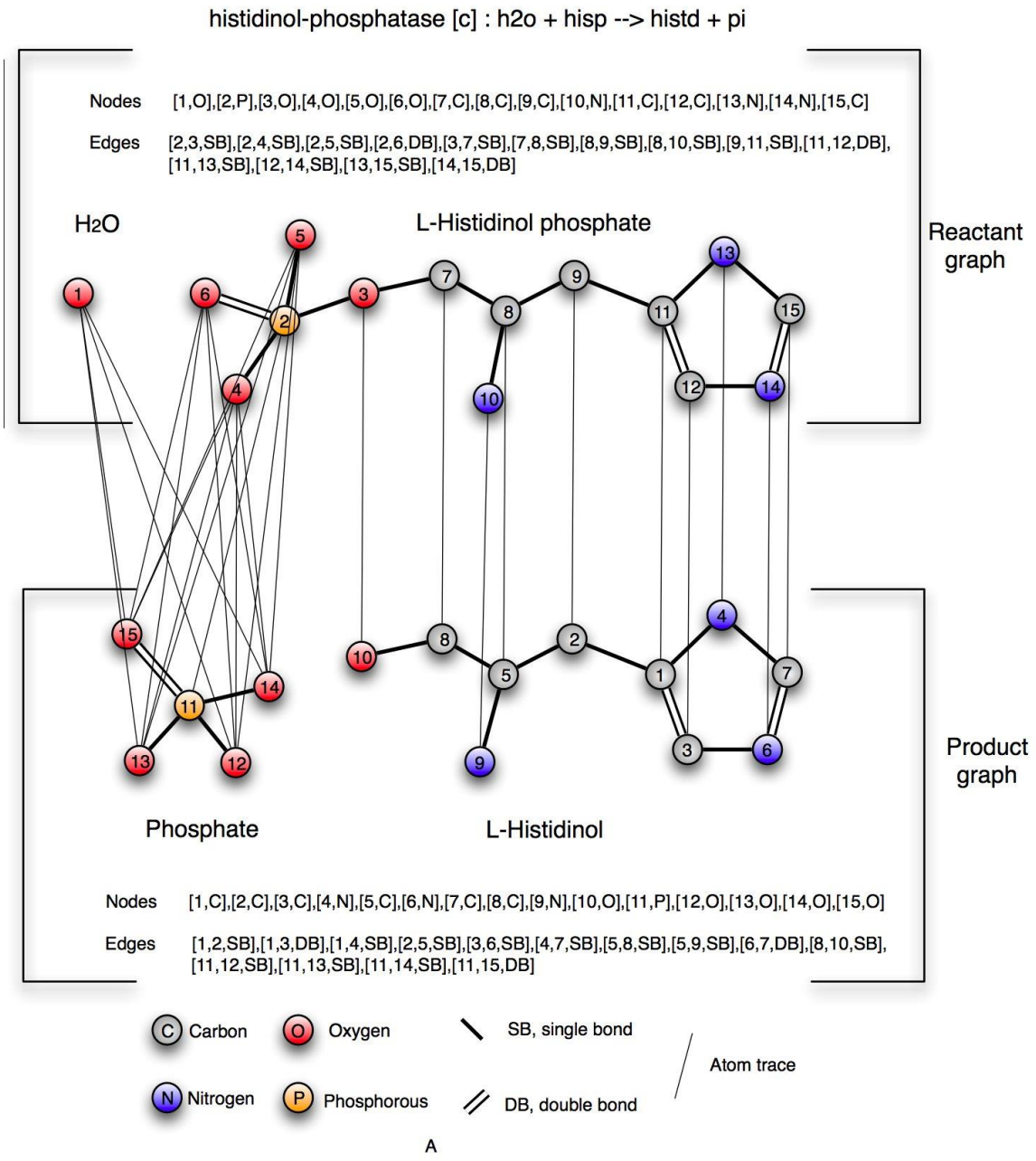

Figure C.1.2: The reactant and product graphs of reaction histidinol-phosphate. The mathematical form of the two graphs and the molecules represented by them are shown. The grey lines between the reactant and product graph trace atoms from reactants to products based on the 24 atom mappings generated for the reaction. Grey circles are carbon atoms; red circles are oxygen atoms; blue circles are nitrogen atoms; orange circles are phosphorous atoms.

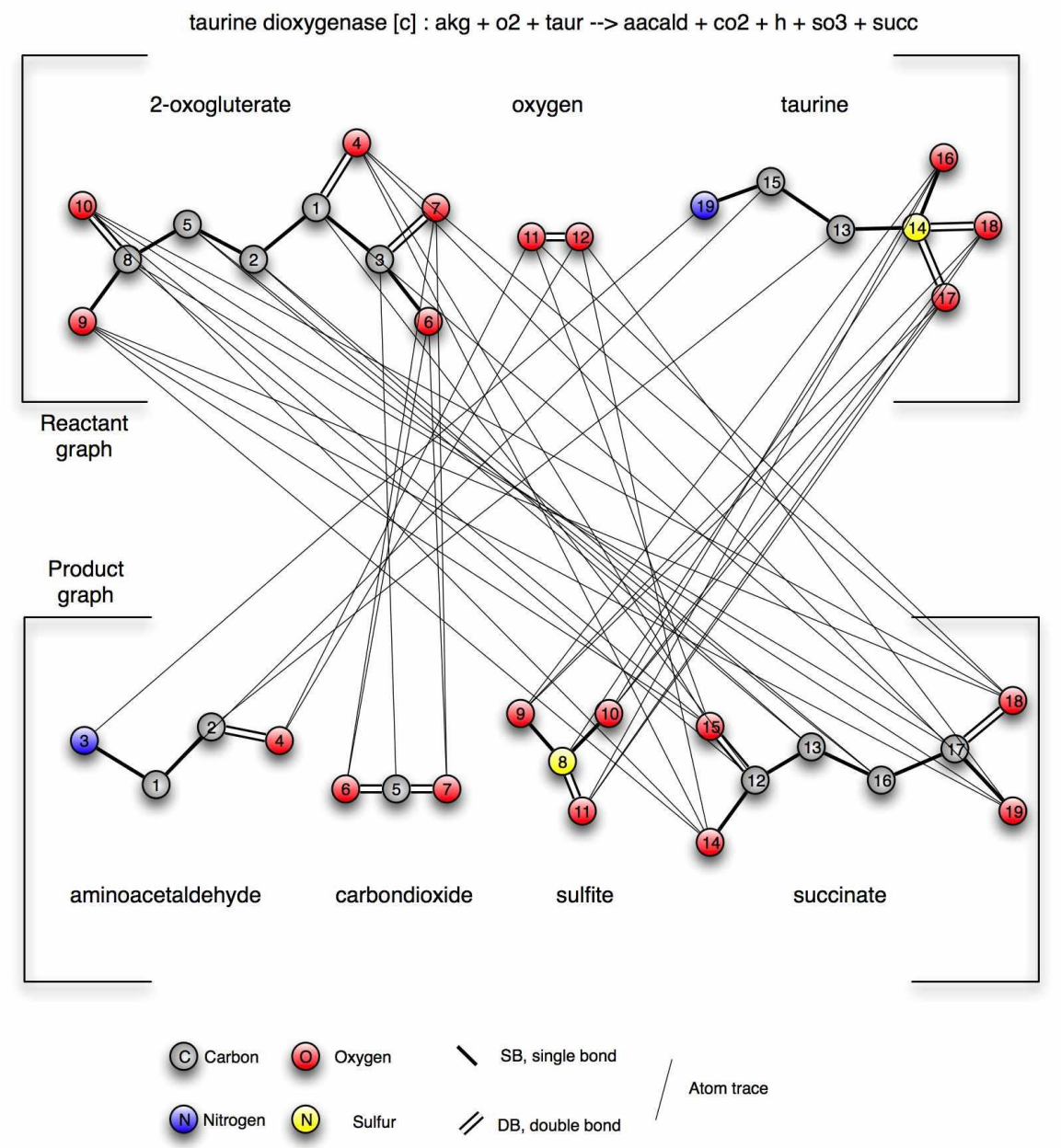

Figure C.1.3: Atom mappings of Taurine dioxygenase. The atom mappings between reactant graph and product graph are shown as a set of grey lines connecting reactant atoms to product atoms.

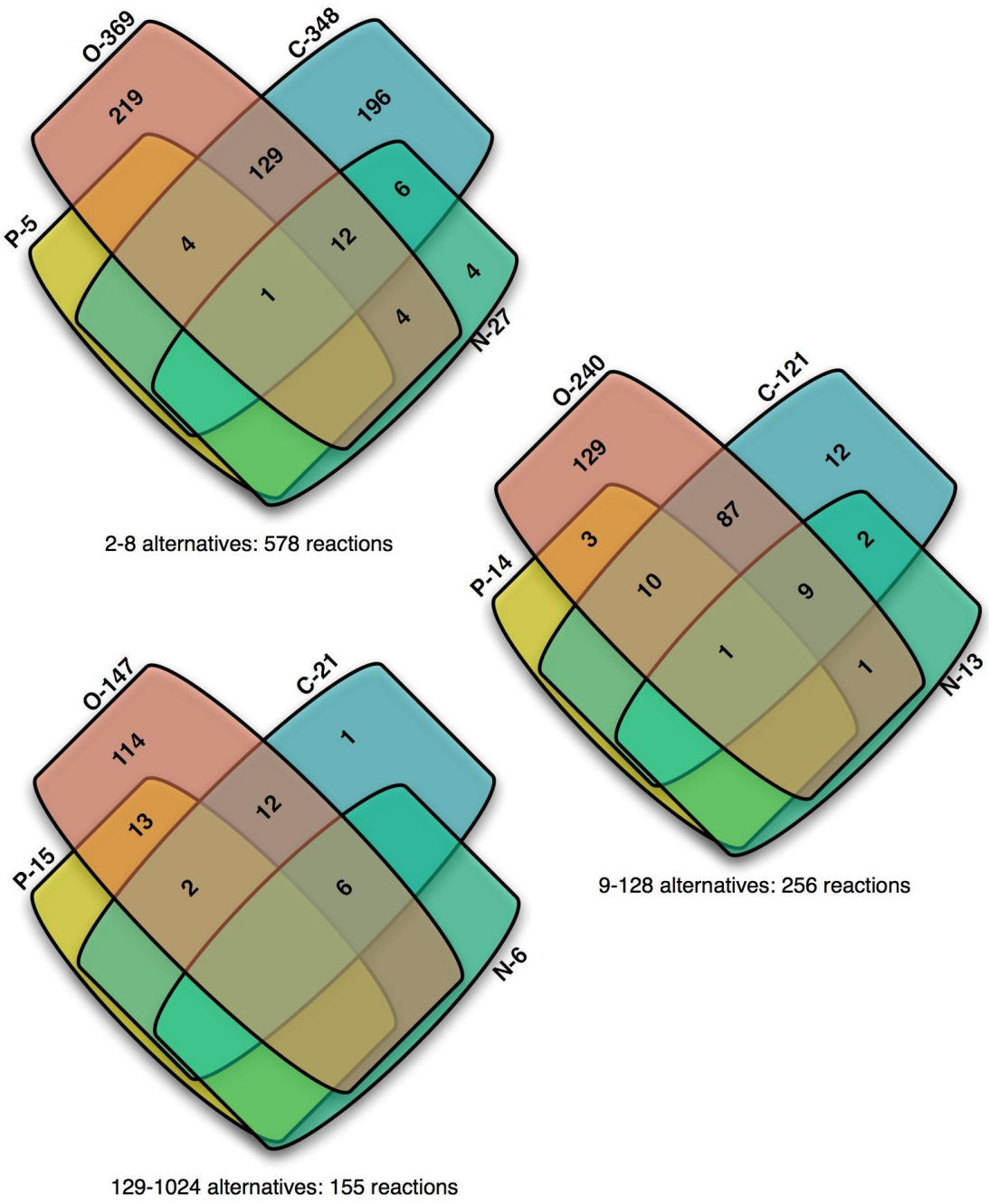

Figure C.1.4: The distribution of reaction mappings present in *im*PR90068 based on the type of equivalent atom(s) contributing to degeneracy in the mappings. Degeneracy arising due to equivalent C, O, N and P are shown respectively using blue, orange, green and yellow squares. The reactions are classified into three categories based on the number of alternative mappings present in individual reaction mappings: reactions containing 2-8, 9-128 or 129-1024 alternative mappings follow particular trends with respect to the reactant groups and atom types that result in degeneracy of mapping data.

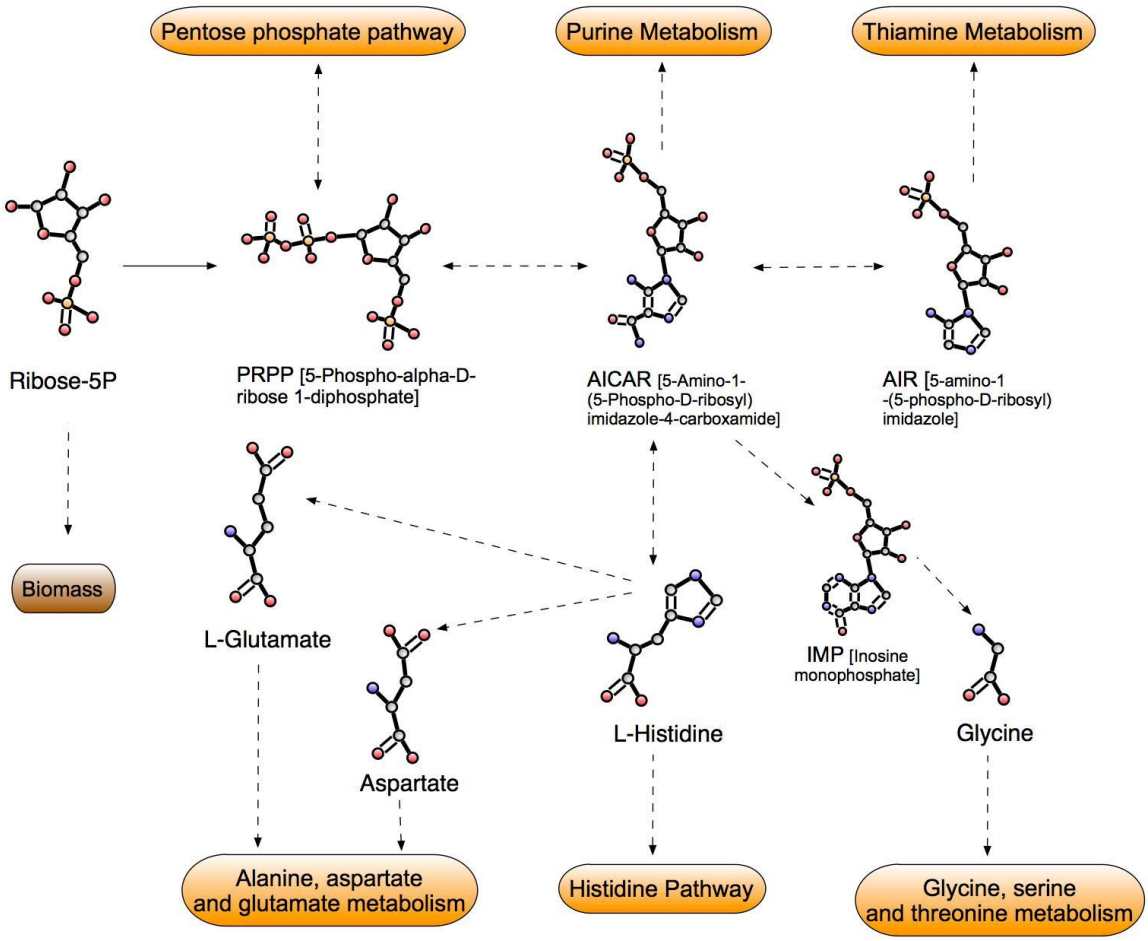

Figure C.1.5: An example of the expanded scope of the genome-scale isotope mapping model imPR90068. In imPS1485 Ribose-5P production was directly routed to biomass as a stand-in substitute for histidine. In imPR90068 R5P downstream conversion is linked to other amino acid synthesis pathways.

D. Specific Aim 3: Metabolic Flux Elucidation Algorithms Given GC-MS or NMR data

D.1 Identification of Optimal Measurement Sets for Complete Flux Elucidation in Isotopically Dynamic MFA Experiments

D.1.1 Introduction

Metabolic flux analysis (MFA) aims at the quantification of intracellular fluxes of metabolic networks. MFA methods based on ^{13}C labeling experiments have been successful in elucidating fluxes in many microorganisms under isotopic steady-state [347]. However, isotopic steady-state can be difficult to attain due to long duration of the experiment and high cost of the labeled substrates [348]. Moreover, isotopically stationary MFA (IS-MFA) cannot determine fluxes whose reactants and products reach steady-state isotopic distributions that are insensitive to the flux values. For instance, the intracellular fluxes of the one-carbon metabolism such as photoautotrophic metabolism fixing CO_2 are unidentifiable under IS-MFA, and this situation could arise in cases [349].

These limitations of IS-MFA are overcome in isotopically dynamic/non-stationary MFA (ID-MFA) experiments, which track the temporal changes in isotopic distribution of intracellular metabolites [350-353]. There have been several flux elucidation efforts utilizing non-stationary isotopic measurements [354-358]. In all these efforts, only a subset of internal fluxes and concentrations were determined as the available measurements were not sufficient for complete flux elucidation.

Recent progress in the ability to measure fragments from an expanded list of internal metabolites for microorganisms such as E. coli has enabled complete flux and concentration elucidation using relatively comprehensive model [359-361]. Recently, we developed an integer programming formulation OptMeas, which identifies a minimal number of measurements that uniquely determine the fluxes [362] using stationary isotopic data. Notably, the incidence structure analysis introduced in OptMeas does not require isotopic steady-state condition unlike other parameterization or linearization approaches [363].

In this work, we extend the OptMeas formulation to perform flux and concentration elucidation under isotopically transient conditions (ID-MFA). Additional variables are introduced in the optimization formulation to capture metabolite concentration information. The approximation gap, inherent with the incidence matrix analysis approach [362], is kept at a minimum through careful definition of the concentration variables. This gap is subsequently closed by relying on an iterative procedure. In addition, fast prescreening tests based on linear algebra are first carried out to detect infeasibility and sub-optimality. The screened solutions are further queried and refined until they fully determine the system.

The extended OptMeas formulation that can make use of isotopically non-stationary data is tested using a small network example producing minimal measurement sets for the complete determination of all fluxes and concentrations. OptMeas also correctly predicted that the measurement set can be further reduced if flux identifiability is the only requirement. OptMeas is subsequently applied to a medium-scale *E. coli* model (i.e., 75 reactions and 74 metabolites) to determine minimal measurement sets that resolve all identifiable fluxes.

D.1.2. Materials and methods

D.1.2.1 **Overview of Mathematical Model for ID-MFA**

For the systematic representation of the fluxes, metabolites, and isotopes present in ID-MFA models, we use the following sets throughout the section:

Sets:

$I = \{i\}$: metabolite pools (intermediate metabolites: $I^N \subset I$)

$J = \{j\}$: unidirectional fluxes

$K_i = \{k\}$: isotopomers of metabolite $i \in I$

$T = \{t_1,...,t_{|T|}\}$: sampling time points (initial point: $t_0 = 0$)

We define parameters S_{ij} for stoichiometry and $IMM^j_{i' \to i,k' \to k}$ for isotopomer mapping as in Chang et al. [362]. State variables of the ID-MFA models are:

Variables:

$v_j \geq 0 \quad j \in J$: flux values

$C_i \geq 0 \quad i \in I^N$: metabolite concentrations (pool sizes)

$I_{ik}(t) \in [0,1] \quad k \in K_i, i \in I$: isotopomer fractions (known for all substrates $I \setminus I^N$)

Using these variables, the metabolic and isotopic balances can be put forth as follows:

$$\sum_{j \in J} S_{ij} v_j = 0 \qquad i \in I^N \tag{1}$$

$$\sum_{k \in K_i} I_{ik} = 1 \qquad i \in I^N \tag{2}$$

$$C_i \frac{dI_{ik}}{dt} = \sum_{j|S_{ij}>0} \left(S_{ij} v_j \prod_{i' \in I} \sum_{k' \in K_{i'}} IMM^j_{i' \to i,k' \to k} I_{i'k'} \right) + \sum_{j|S_{ij}<0} S_{ij} v_j I_{ik} \qquad k \in K_i, i \in I^N \tag{3a}$$

Eq. (1) describes the metabolic steady-state where fluxes and concentrations remain constant, and Eq. (2) dictates that the sum of the isotopomer fractions must be equal to one. Eq. (3a) describes how the isotopomer fractions change over time [see 362 for a detailed description]. Given a set of measured fluxes and metabolites, the remaining unknowns v_j, C_i, and $I_{ik}(t)$ are determined by solving the system of differential and algebraic equations (DAE).

D.1.2.2 **Inverse Problems of ID-MFA**

We estimate unknown fluxes and concentrations from isotopic measurements by means of a least-squares parameter fitting (*inverse problem*). The isotopic non-stationarity renders the inverse problem to a dynamic optimization (DO) problem. This resulting problem has been solved before using iterative sequential procedures [354, 355] or evolutionary algorithms [357, 358]. These approaches require the repetitive simulation of Eq. (3a) which becomes very time-consuming for isotopomer models derived from comprehensive metabolic models.

In this work, we discretize variables $I_{ik}(t)$ in time to convert the DO problem into large-scale (algebraic) nonlinear programming (NLP) problems DynaCalc and DynaRange that minimize the sum of squared errors (SSE) (see supplementary text). Note that least-squares methods are used for overdetermined systems. We apply the incidence structure analysis [362] and find minimal sets of measurements that are required to fully determine the ID-MFA system.

D.1.2.3 **OptMeas Formulation for ID-MFA**

A key observation arising from Eq. (3a) is that fluxes always appear multiplicatively coupled with concentration variables. One can combine the two using a single variable set after dividing both sides of equation (3a) by C_i

$$\frac{dI_{ik}}{dt} = \sum_{j|S_{ij}>0} \left(w_{ij} \prod_{i' \in I} \sum_{k' \in K_{i'}} IMM^{j}_{i' \to i, k' \to k} I_{i'k'} \right) - w_{ii} I_{ik} \qquad k \in K_i, i \in I^N \tag{3}$$

where w_{ij} is the *space velocity* of influx j to metabolite i ($S_{ij} > 0$) defined as

$$w_{ij} = \frac{S_{ij} v_j}{C_i} \qquad i \in I^N, j \in J \tag{4}$$

and w_{ii} is the sum of all the outgoing space velocities of metabolite i

$$w_{ii} = -\sum_{j|S_{ij}<0} \frac{S_{ij} v_j}{C_i} \qquad i \in I^N \tag{5}$$

Note that the isotopic distribution of metabolite i does not depend on the magnitude of each efflux but only on their sum w_{ii} that is the reciprocal of the *residence time* or *turnover rate*. Similarly, if multiple fluxes lead to the production of a metabolite using the same reactants through identical atom mappings, the corresponding space velocities cannot be determined uniquely. Only their sum can be determined by solving Eq. (3) for the metabolite in question. We use Eqs. (1–5) to construct the incidence matrix of ID-MFA system (Figure D.1.1). Note that one isotopomer balance for each metabolite is dropped in Eq. (3) in order to eliminate the inherent redundancy due to Eq. (2).

We assign one output variable (column) to each equation (row) based on the incidence structure using binary variables x, y, and z for the rows, nonzero elements, and columns of the incidence matrix, respectively [362]:

$$\sum_{r \in RO} y_{rc} = z_c \qquad c \in CO$$
$$\sum_{c \in CO} y_{rc} = x_r \qquad r \in RO \tag{10}$$

Here, $y_{rc} = 1$ if and only if column c is assigned to row r. If column c is not assigned to any row ($z_c = 0$) then column c must be part of measurement set. Column variable z in the incidence matrix is denoted as $(z_j, z_i, z_{ij}, z_{ii}, z_{ik})$ and accounts for the state variables of v_j, C_i, w_{ij}, w_{ii}, and I_{ik}, respectively. Since w_{ij} and w_{ii} are not measured, we enforce $z_{ij} = 1$ and

$z_{ii} = 1$ at all times. For the sake of economy of presentation, concentration and IDV of metabolites are assumed to always be measurable. This restriction is lifted in D.1.2.5 to account for indirect measurement options such as lumped pool and mass isotopomer distribution vector (MDV).

Total measurement cost is taken to be a linear combination of individual measurement costs as tabulated below. Different cost structures are discussed in D.1.2.5.

Parameters:

q_j : flux measurement cost

q_{Ci}, q_{Ii} : metabolite measurement cost for concentration and IDV, respectively

OptMeas aims at minimizing the total measurement cost

$$\sum_{j \in J} q_j (1 - z_j) + \sum_{i \in I^N} \left(q_{Ci}(1 - z_i) + q_{Ii}(1 - u_i) \right) \tag{11}$$

where binary variable $u_i = 1$ if and only if the IDV of metabolite i is *not* measured, and satisfies

$$u_i \le z_{ik} \qquad k \in K_i \tag{12}$$

In D.1.2.5, we discuss two interesting variants of OptMeas for ID-MFA: 1) OptMeas that is used to estimate the number of intracellular fluxes that need to be known to determine the system, and 2) the OptMeas formulation that focuses on only flux identifiability.

D.1.2.4 **Solution Strategy**

The incidence structure analysis that OptMeas relies on may generate measurement sets that do not fully determine the nonlinear DAE system (1–5). Therefore, we use an iterative procedure adapted from Chang et al. [362] to recover the measurement sets for unique flux and concentration determination.

Step 0: Initialization. Preprocess the network so that S is of full row rank. Construct OptMeas formulation that is updated by introducing integer cuts in the course of the algorithm. Define set MS containing the list of optimal measurement sets, and initialize it to be empty.

Step 1: Solve OptMeas. Solve the current realization of OptMeas using an integer linear programming (ILP) solver, and obtain (J^*, I^{N1}, I^{N2}) as optimal measurement choices for fluxes, concentrations, and IDVs. Here, we used CPLEX 11 [364] accessed through Concert technology 2.5 [365].

Step 2: Remove linearly dependent flux measurements. Remove columns J^* from S. If the resulting matrix has full row rank, continue with Step 3. Otherwise, introduce the following integer cut into OptMeas and return to Step 1.

$$\sum_{j \in J^*} z_j \ge 1$$

Step 3: Prescreen for a unique flux/concentration elucidation. Test if unknown fluxes and concentrations can be uniquely determined as described in D.1.2.6. If so, continue with Step 4. Otherwise, cut off the current flux and concentration measurements and return to Step 1.

Step 4: Check for a unique flux/concentration elucidation. Test if the measurement set (J^*, I^{N1}, I^{N2}) uniquely determines all fluxes and concentrations in the network. This is accomplished by solving formulation TestUniq described in D.1.2.6. If the set determines all fluxes and concentrations, move to the next step. Otherwise, go to Step 7.

Step 5: Check for solution optimality. Test if (J^*, I^{N1}, I^{N2}) is optimal by solving TestOpt given in D.1.2.6. If TestOpt marks any measurement as unnecessary, then remove it from current set and move to the next step.

Step 6: Termination criterion. If the current measurement set has a higher relative cost than a predefined threshold, terminate and report the current MS as the final collection of all optimal measurement sets. Otherwise, include the current measurement set in MS.

Step 7: Search for alternative solutions. Remove (J^*, I^{N1}, I^{N2}) from the search space using the following integer cut and go back to Step 1.

$$\sum_{j\in J\setminus J^*} z_j + \sum_{i\in I^N\setminus I^{N1}} z_i + \sum_{i\in I^N\setminus I^{N2}} u_i - \left(\sum_{j\in J^*} z_j + \sum_{i\in I^{N1}} z_i + \sum_{i\in I^{N2}} u_i \right) < \left|J\setminus J^*\right| + \left|I^N\setminus I^{N1}\right| + \left|I\setminus I^{N2}\right| \quad (13)$$

D.1.2.5 **Extension of OptMeas**

Consideration of Lumped Pool Measurements

Metabolites X5P, R5P, and Ru5P are difficult to measure individually, but their lumped pool is not [358, 366]. For such measurements, we use set $L = \{l\}$ for lumped pools and I^l for constituent metabolites of pool l. Then, the pool size C_l and IDV I_{lk} of l are:

$$\sum_{i\in I^l} \gamma_{li} C_i = C_l \qquad l \in L$$

$$\sum_{i\in I^l} \sum_{k'\in K_i} \delta_{ik'lk} I_{ik'} = I_{lk} \qquad k \in K_l, l \in L$$

We denote the additional columns in the incidence matrix as z_l and z_{lk}. If the relative measurement cost of C_l and I_{lk} is q_{Cl} and q_{Il} respectively, the cost function is extended to:

$$\sum_{j\in J} q_j (1 - z_j) + \sum_{i\in I^N} \left(q_{Ci}(1 - z_i) + q_{Ii}(1 - u_i) \right) + \sum_{l\in L} \left(q_{Cl}(1 - z_l) + q_{Il}(1 - u_l) \right)$$

where binary variable $u_l = 1$ if and only if IDV of lumped pool l is *not* measured.

Consideration of Different Cost Structures for Metabolite Measurement

In general, the cost Q_i of measuring both concentration and IDV of metabolite i is smaller than $q_{Ci} + q_{Ii}$. We use binary variable U_i that is equal to 1 if and only if we do not make the simultaneous measurement of metabolite i. We substitute u_i in Eq. (12) by $u_i + U_i - 1$, and z_i in Eq. (10) by $z_i + U_i - 1$ in order to avoid duplicate measurements, and augment the cost function:

$$\sum_{j\in J} q_j\left(1-z_j\right)+\sum_{i\in I^N}\left(q_{Ci}\left(1-z_i\right)+q_{Ii}\left(1-u_i\right)+Q_i\left(1-U_i\right)\right)$$

Conservative OptMeas

OptMeas assigns an isotopomer variable to any equation that it participates in, which could cause nonlinear dependency. If we allow only the assignment of an isotopomer variable to its own balance equation, then OptMeas behaves conservative in predicting identifiability and the propagation of measurement errors upstream in the network is prevented. For this, we introduce

$$z_c \le \sum_{r\in R_c} y_{rc}$$

where column c corresponds to an isotopomer variable I_{ik} and row set R_c consists of Eq. (2) for metabolite i and Eq. (3) for isotopomer k.

Restriction to Flux Identifiability

In some situations when the flux identifiability is of primary interest, we can ignore concentration identifiability by introducing two additional binary variables

$y_i = 1$ iff concentration of metabolite i is required to resolve any flux

$\zeta_i = 1$ iff concentration measurement of i is not necessary for flux identifiability

Auxiliary variable ζ_i is used to linearize the concentration measurement term in the objective function $q_{Ci}y_i\left(1-z_i\right)$ to $q_{Ci}\left(1-\zeta_i\right)$. These variables are defined by the following constraints:

$$\left|R_i\right|\left|J\right|y_i \ge \sum_{r\in R_i}\sum_{j\in J} y_{rj} \ge y_i$$

$$z_i+\left(1-y_i\right)\ge \zeta_i$$

where R_i is the set of rows corresponding to $w_{i?}$.

D.1.2.6 **Sub-problems for the Proposed Procedure**

Prescreening Test for Unique Flux Elucidation

In order to devise a fast test for identifiability of a measurement set, we rely on linear algebra upon expanded stoichiometric matrix ES that accounts for the participation of fluxes and concentrations in Eqs. (1), (4), and (5). Matrix ES is similar to the first two columns and three rows of incidence matrix (Figure D.1.2). For flux column j, $ES_{rj}=S_{ij}$ if r corresponds to Eq. (1) for metabolite i, $ES_{rj}=S_{ij}$ if r corresponds to Eq. (4) for w_{ij} and $S_{ij}>0$, and $ES_{rj}=-S_{ij}$ if r corresponds to Eq. (5) for w_{ii} and $S_{ij}<0$. For concentration column i, $ES_{ri}=1$ if r corresponds to Eq. (4) for w_{ij}, and $ES_{ri}=v_i^{in}$ if r corresponds to Eq. (5) for w_{ii} where v_i^{in} is the number of incoming fluxes of metabolite i with distinct atom mapping. All the unassigned elements of ES are set to 0.

Then, the rank of ES indicates the maximum number of fluxes and concentrations that can be determined. Moreover, if we remove measured columns from ES, the rank of the resulting thinner matrix $\overline{ES}$ implies the maximum number of unmeasured fluxes and concentrations that

can be determined. Using this property of ES, we can derive the following integer cut that prohibits a measurement set that does not have full column-rank of $\overline{ES}$.

$$\sum_{j\in J}\left(2z_j^*-1\right)\left(z_j-z_j^*\right)+\sum_{i\in I^N}\left(2z_i^*-1\right)\left(z_i-z_i^*\right)<0.$$

Uniqueness and Optimality Tests

Optimization formulations TestUniq and TestOpt are similar to those for IS-MFA [362] except that they are DO problems with additional concentration variables:

$$\begin{aligned}(\text{TestUniq}) \quad \max \quad & \sum_{j\in J\setminus J^*}\left(v_j-\bar{v}_j\right)^2+\sum_{i\in I^N\setminus I^{N1}}\left(C_i-\bar{C}_i\right)^2 \\ \text{s.t.} \quad & \text{Eqs. } (1),(2),(3a), \\ & \text{Fix all the measured variables}\end{aligned}$$

$$\begin{aligned}(\text{TestOpt}(x)) \quad \max \quad & \|x-\bar{x}\| \\ \text{s.t.} \quad & \text{Eqs. } (1),(2),(3a), \\ & \text{Fix all the measured variables except } \ x\end{aligned}$$

As for DynaCalc, we apply the total discretization scheme to convert these sub-problems into large-scale NLP problems, which are then locally solved using multiple starting points.

D.1.3. Results and Discussion

D.1.3.1 Illustrative Network Example

We first consider the small network adapted from Nöh and Wiechert [367] as shown in Figure D.1.3a. The network is simulated using MATLAB 7.6 [368] using the flux, concentration and substrate labeling entries used in the original paper (see Figure D.1.2b and D.1.2c). We assume that external fluxes v_1, v_4, and v_5, the IDV of substrate A, and the concentration and MDV of all internal metabolites (B, C, D, and E) can be precisely measured. Five samples at time points 1, 3, 7, 15, and 31 of the simulated MDV time profiles are used for flux inference.

The measurement set used in the original paper [367] is composed of v_1 and the MDVs of B, C, D, E. These measurements are sufficient to infer all the intracellular fluxes (see Table D.1.1). However, because MDV measurements are conducted multiple times (five times for this example), measuring all four intracellular metabolites could be costly. Based on the cost coefficients shown in Table D.1.2, the measurement set of the original paper [367] has a relative cost of 56.

OptMeas successfully identified 48 optimal solutions with a relative cost of 27 (less than half of original cost). These solutions require measurement of two external fluxes chosen from v_1, v_4, or v_5. The required metabolite measurements include: 1) MDV of D plus MDV of either B or E, 2) any concentration plus MDV of C or MDVs of both B and E, 3) any combination of two concentrations plus MDV of D, or 4) any combination of three concentrations plus MDV of B or E. Flux and concentration identifiability of these suggested measurement sets were verified by solving TestUniq to global optimality using GAMS/BARON [369]. The DynaRange results for one measurement set $\left(v_1,v_4;C_{\text{B}},C_{\text{D}},C_{\text{E}};MDV_{\text{E}}\right)$ are shown in Table D.1.3.

If only flux and not concentration elucidation is sought after then all concentration measurements in the optimal sets can be eliminated. The optimal measurement sets for this case include any combination of two external fluxes plus the MDV of either B or E. This example demonstrated that OptMeas can generate measurement sets that allow for complete flux elucidation at substantial reduction in cost.

D.1.3.2 *Escherichia coli* Network

Analysis of isotopomer model

We next apply OptMeas to the metabolic network of the 1,3-propanediol (PDO) producing *E. coli* strain including 74 metabolites, 75 reactions, and 4,806 isotopomers [370]. An abridged version of this model was investigated before using ID-MFA [355] in order to elucidate net fluxes between carbon containing metabolites. However, many exchange rates of reversible reactions were left non-determined using the MDV measurements of proteinogenic amino acids and the intermediary metabolites (AKG, Cit, Mal, Pyr, Suc).

The model was earlier subjected to the stationary OptMeas analysis [362], which showed the potential redundancy of the MDV measurements used in the original paper and suggested novel measurement options for better flux elucidation. Here we perform ID-MFA for improved flux and concentration elucidation based on OptMeas. All fluxes are numbered and metabolites are named in agreement with the nomenclature scheme used in the original paper. We modified the original model to account for the symmetry of glycerol.

We summarize intracellular metabolite measurement alternatives that have been used for ID-MFA in Table D.1.4. The measurement of all other metabolites was prevented (see Figure D.1.3).

We first analyzed the expanded stoichiometric matrix (D.1.2.6) and detected that the exchange rate of the transhydrogenation reaction (reaction 64) and the branching ratio of two oxidative decarboxylations of malate (reactions 28 and 29) are both unidentifiable irrespective of substrate labeling or isotopic measurement. This agrees with earlier observations [362]. We also identified ten *conditionally unidentifiable* exchange rates of reversible reactions for the currently available measurement set (see Figure D.1.4). All of these *practically unidentifiable* fluxes [371, 372] are excluded from any further analysis.

Identification of measurements

First we reduce the problem dimensionality by identifying groups of fully coupled fluxes [373], shown in Table D.1.5, thus retaining only a single flux measurement for each group. OptMeas returned eight optimal measurement sets with some measurements shared amongst all of them. Common measurements include the rates of glucose uptake (v_{66}), glycerol uptake (v_{68}), oxygen uptake (v_{72}), PDO secretion (v_{69}), and CO_2 secretion (v_{71}), and concentrations of Ac, Arg, Cys, E4P, Leu, Pro, and TA-C3. Common isotopic measurements include IDV of PDO and MDV of Glu, Ile, Lys, Met, S7P, and Ser. Other than these common measurements, OptMeas required two MDV measurements of Phe or Tyr and Ala or Val.

Among these measurements, the oxygen uptake rate, IDV of PDO, and MDV of Glu, Met, and Phe are present in ID-MFA as they are for IS-MFA [362]. MDV of S7P and concentration of E4P together provide better elucidation of pentose phosphate pathway and determine the exchange rate of the transaldolase (v_{16}) and transketolase (v_{14}) reactions that were not determined before [355, 370]. Arg, Leu, and Pro do not produce any carbon-containing products as they only serve as biomass constituents in the model. Therefore, the most cost effective way of elucidating their concentration is by measuring them directly. Interestingly, both

Cys and TA-C3 cannot be elucidated through other measurements requiring instead their direct measurement which is prohibited in this example.

By investigating 80 near-optimal solutions derived by OptMeas, we observe that the concentration and MDV measurement of some metabolites are interchangeable. For example, if one solution requires the concentration of Tyr and MDV of Phe, then another solution requires the concentration of Phe and MDV of Tyr. The same is true for Ala and Val pair. These pairs receive carbons from the same metabolites through similar biosynthetic reactions (Phe and Tyr from E4P and PEP, and Ala and Val from Pyr). A similar observation is made for the triplet of Ile, Leu, and Pro, which receive carbons from TCA cycle metabolites (OAC, AcCoA, and AKG) respectively. The near-optimal solutions also suggest one extra MDV measurement of 6PG, F6P, FBP, G6P for a better elucidation of glucosephosphate isomerase (v_1) and fructose-biphosphate aldolase (v_3) in glycolytic pathway and transaldolase (v_{13}) in phosphate pathway.

Recall that the concentration measurement of Cys and TA-C3 are required but not allowed. Interestingly, OptMeas replaced these concentration measurements by citrate uptake (v_{67}), acetate secretion (v_{70}), biomass formation (v_{75}) rates and MDV of Asp when only the flux identifiability is required. This is an example where the measurement of metabolite concentrations directly leads to the elucidation of fluxes.

In summery, the identified measurement set is composed of eight external fluxes v_{66}, v_{67}, v_{68}, v_{69}, v_{70}, v_{71}, v_{72}, v_{75}, seven concentrations of Ac, Ala, Arg, E4P, Leu, Pro, Tyr, one IDV of PDO, and eleven MDVs of Asp, F6P, FBP, Glu, Ile, Lys, Met, Phe, S7P, Ser, Val. It is tested for the ability to fully identify fluxes in the metabolic model by solving DynaRange. We found that this measurement set still leaves the exchange rates of v_{25} and v_{27} undetermined. This is due to the symmetry of Fum and Suc and the cascade of reversible reactions in the TCA cycle that scramble their atom mapping. They can only be determined by adding the MDV measurements of Fum, Mal, and Suc in the list.

D.1.4 Summary

In this work, we extended the OptMeas formulation to account for flux and concentration inference in ID-MFA under isotopically non-stationary conditions. OptMeas exploits the multiplicative coupling of fluxes and concentrations in the isotope balance equations by recasting the isotope balance using space velocities as new variables. OptMeas correctly identified all optimal measurement sets for the small network example and predicted that some of the measurements are dispensable if only fluxes and not concentrations are needed. When applied to a medium-scale *E. coli* network [370], OptMeas found a set of fluxes that are unidentifiable under ID-MFA due to linear dependency, and suggested the measurement of intracellular metabolites that are distributed among the network to elucidate five more exchange rates than the previously reported ID-MFA results [355].

As demonstrated with *E. coli* network example, OptMeas can be improved by imposing the results of external analyses such as other identifiability analyses and flux/concentration coupling analyses [374] to refine its predictions. Moreover, OptMeas can be modified to focus on the variables that are specifically interested. For example, if certain fluxes and concentrations are not determined by an ID-MFA experiment, we can identify which measurements to make in the next experiment to pinpoint them.

The proposed solution procedure can greatly benefit from global optimization algorithms that exploit the structure of metabolic networks. We are currently exploring how to decompose an isotopomer network to generate sub-networks that are easy to solve and require the least

amount of effort in connecting their solutions. Both EMU and cumomer representations [355, 375] will be also considered for this purpose (Suthers et al. 2010).

We note that the incidence structure analysis used in OptMeas is also applicable to kinetic parameter estimation using metabolic non-stationary MFA experiments that have been explored recently [376-378]. We are also probing how to model the nonlinear structure of the mechanistic rate equations in order to help OptMeas produce the measurement sets that determine as many kinetic parameters as possible.

Table D.1.1 Estimated flux ranges for the illustrative toy network using the measurement set of Nöh and Wiechert [367]

Fluxes	true	Lobatto 5	Radau 5
v_1	10	10	10
v_2^f	13.7	[13.61, 13.80]	[13.65, 13.75]
v_2^b	6.7	[6.61, 6.79]	[6.64, 6.75]
v_3	3	[2.99, 3.00]	[2.99, 3.00]
v_4	7	[7.00, 7.01]	[7.00, 7.01]
SSE	-	1E-5[a]	1E-6[a]

The flux ranges are obtained by solving DynaRange formulation. They are tight enclosing the true value (with the relative error less than 0.2% for the unidirectional reactions and below 1% for the reversible reaction).

Table D.1.2 Measurement costs for the illustrative toy network

Type	Measurements	Relative cost
external fluxes	v_1, v_4, v_5	1
	v_6	2
concentrations	B, C, D, E	5
IDVs / MDVs	B, E	20 / 10
	D	30 / 15
	C	40 / 20

Table D.1.3 Identifiability results for the illustrative toy network

unknowns	samples used for inverse problem	
	<1>	<1, 3, 7, 15, 31>
v_2^f = 13.7	[13.047, 14.433]	[13.609, 14.646]
v_2^b = 6.7	[6.047, 7.433]	[6.609, 7.646]
C_C = 10	[3.58, 20[a]]	[9.802, 10.556]
SSE	0	1E-5

The range of exchange rate v_2^b is best captured by the first sample, which is most sensitive to this flux. On the other hand, the concentration of C is best estimated by using multiple samples. For discretization, 5th-order Lobatto nodes are used. DynaRange was solved using (number of included samples)*1E-5 as the cutoff.

[a]The upper bound of the concentration.

Table D.1.4. Metabolite measurement candidates from the literature

Pool	Concentration	MDV fragments	q_C	q_{MDV}
3PG	c	c	15	40
6PG	c, e	c	15	40

AcCoA	b, e	-	20	-
AKG	b, c, e	a, b, c	15	40
Cit	c, e	a, c	15	40
DHAP	b, e	-	20	-
E4P	b, e	-	20	-
F6P	b, c, e	b, c	15	40
FBP	b, e	b	15	40
Fum	c, e	c	25	40
G6P	b, c, e	b, c	15	40
GAP	b, e	-	20	-
ICit	e	-	20	-
Mal	b, c, e	a, c	15	40
OAC	e	-	20	-
PEP	b, c, e	b, c	15	40
Pyr	b, c, e	a, b	15	40
R5P	e	-	25	-
Ru5P	e	-	25	-
S7P		b	20	40
Suc	c, e	a	25	40
P5P (R5P+Ru5P+X5P)	b, c	b, c	15	-
SuccFum (Suc+SucCoA+Fum)	b	b	15	-
Ala	f	a, b, c, d, f	10	30
Arg	f	f	10	40
Asn	f	f	-	-
Asp	b, f	a, b, c, d, f	10	30
Gln	f	f	-	-
Glu	b, f	a, b, c, d, f	10	30
Gly	f	a, d, f	10	30
His	f	c, f	-	-
Ile	f	a, d, f	10	30
Leu	f	a, d, f	10	30
Lys		d	15	30
Met	f	a, d, f	10	30
Phe	f	a, c, d, f	10	40
Pro	f	d, f	10	30
Ser	f	a, c, d, f	10	30
Thr	f	a, d, f	10	30
Tyr	f	a, f	10	40
Val	f	a, d, f	10	30

Relative costs are scaled to the external flux measurements (one for liquid flux and ten for gas flux measurement). We assume that the concentration can be measured whenever MDV can be measured. References are (a) [370] or [355], (b) [354], (c) [357] or [358], (d) [324], (e) [379] or [380], (f) [381]. In addition, we assume that the IDV of Ac, CO_2, and PDO are measurable at the relative costs of 50, 30, and 100, respectively as they are metabolites transported into the extracellular medium. Concentration measurement cost for these three products are assumed to be equal to 25.

Table D.1.5. Flux coupling analysis of the medium-scale *E. coli* network

Fully coupled fluxes
v_{17}, v_{18}
v_{23}, v_{63}

v_{35}, v_{36}, v_{69}

v_{38}, v_{39}, v_{40}, v_{42}, v_{43}, v_{48}, v_{49}, v_{50}, v_{52}, v_{53}, v_{54}, v_{55}, v_{56}, v_{57}, v_{58}, v_{59}, v_{60}, v_{61}, v_{73}, v_{74}, v_{75}[a]

[a]This long list of coupled fluxes is due to the biomass equation.

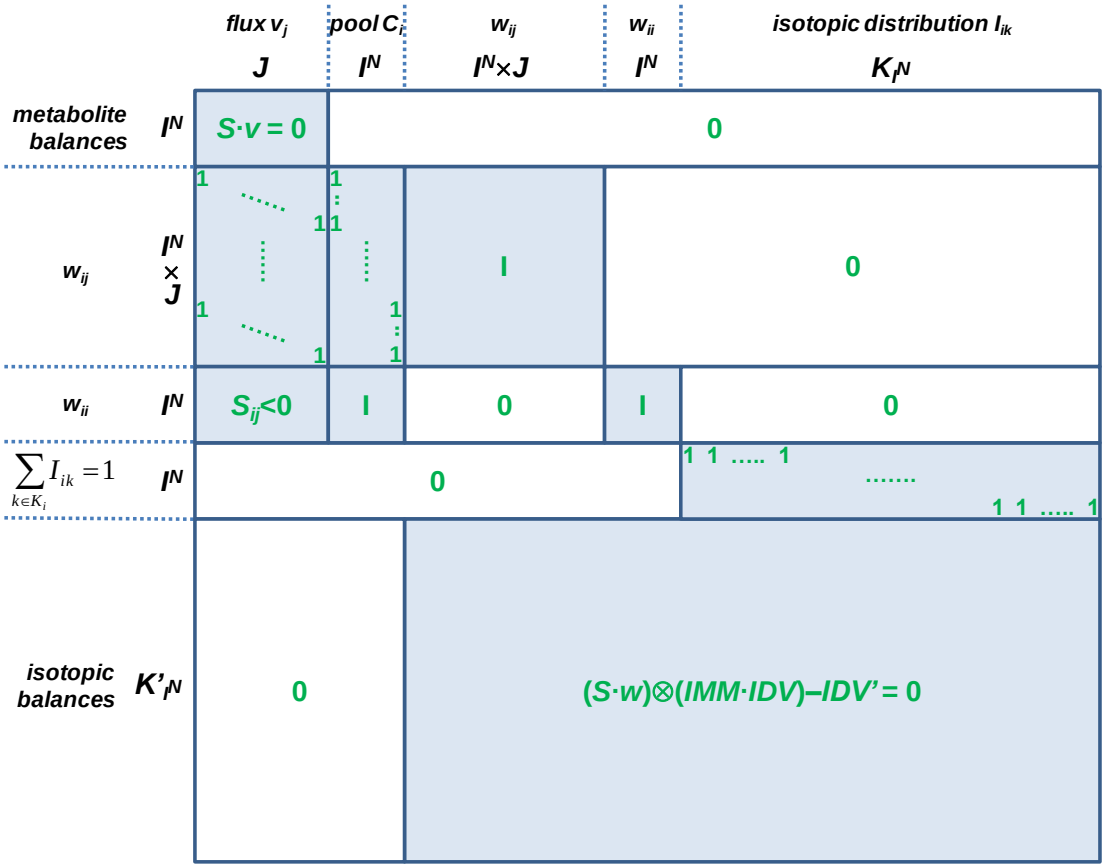

Figure D.1.1. Incidence matrix of the ID-MFA system. The incidence matrix of ID-MFA contains more rows and columns related to concentrations and space velocities than that of IS-MFA, which implies that more measurements are necessary for the identifiability of the underlying mathematical system. K'_i is the same as K_i except that the last isotopomer of metabolite i is excluded. Although denoted as $I^N \times J$ for notational convenience, w_{ij} is defined only for the (i, j) pairs that have $|S_{ij}| > 0$.

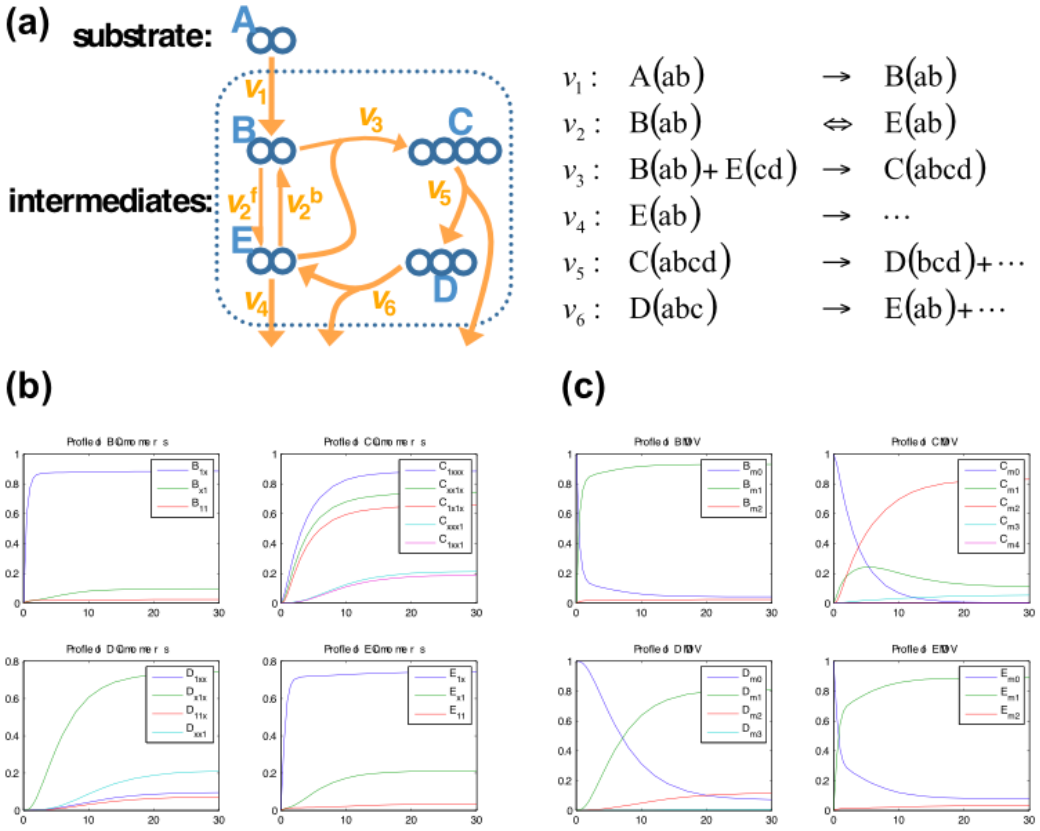

Figure D.1.2 Illustrative small network example. Panel (a) shows its network representation together with stoichiometry and atom mapping for each reaction. Note that metabolites leaving the system have been dropped for visualization purposes. Panel (b) shows the change in the isotopic distribution of each metabolite when the labeled substrate (a mixture of 2% unlabeled, 96% 1-labeled, and 2% fully labeled A) starts to feed. Cumomer fractions are shown for easy comparison with the original paper (Nöh and Wiechert 2006). Panel (c) is the MDV profile that is used to infer fluxes and concentrations. The simulation is conducted given the true flux distribution in Table D.1.2 and concentrations $C_B = 4$, $C_C = 10$, $C_D = 7$, and $C_E = 3$.

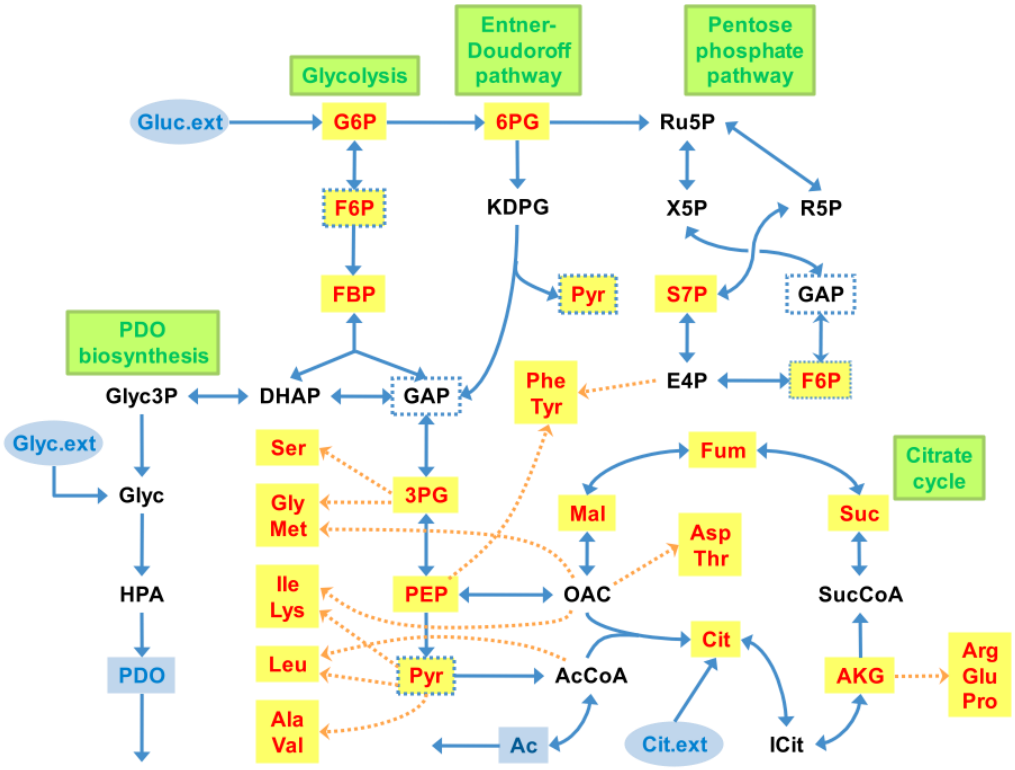

Figure D.1.3. Network representation of the medium-scale *E. coli* model. Substrates are circled, and the MDV and IDV measurements are shown in shaded squares in yellow and blue, respectively. Some metabolites (in dotted square) appear more than once to make the figure more readable.

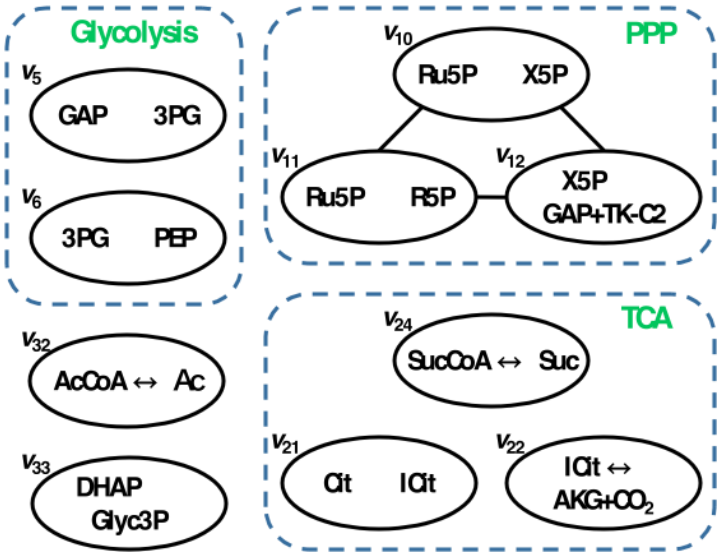

Figure D.1.4. Practically unidentifiable exchange rates of reversible reactions of medium-scale *E. coli* network using available measurements. Reactions in solid circle are found by solving DynaRange for the EMU representation of the network (Antoniewicz et al. 2007a) using GAMS/CONOPT3 NLP solver (Drud 1994) from multiple starting points. A pair of nodes that are connected by an edge correspond to alternative measurements recommended by conservative OptMeas for determining the system.

E. Specific Aim 4: Computational Strain Design

E.1 Identification of Non-native Production Routes and Engineering Interventions for the Microbial Synthesis of Long-chain Alcohols

The work in this section was published [99].

E.1.1. Introduction

Increasing demands for renewable energy and environmental concerns have stimulated an interest towards the production of second generation biofuels from renewable sources [109]. For the past few decades, bio-ethanol was considered as a substitute for transportation fuels. More recently, long-chain alcohols (C3-C5) have also emerged as biofuel alternatives because of their higher energy density and ease of storage [382]. Microorganisms from diverse environments naturally produce ethanol during fermentation. However, the natural synthesis of higher alcohols is not as commonplace with the exception of certain *Clostridia* strains [383, 384]. One possible production alternative for 1-butanol and 1-propanol is to use native pathways in *Clostridium acetobutylicum* [103, 105, 385-387]. An alternative approach is to integrate non-native pathways into standard microbial production hosts (i.e., *Escherichia coli* or yeast) by exploiting the conversion of key intermediary amino acids into long-chain alcohols [388, 389]. In this regard, numerous efforts have been made in the recent past to clone and express *Clostridia* genes (butyryl-CoA dehydrogenase, *bcd*) responsible for the production of 1-butanol in *E. coli* [390-392]. Homologs and isoenzymes of *bcd* from *Megasphaera elsdenii* [393, 394] and crotonoyl-CoA reductase (*ccr*) from *Streptomyces coelicolor* [395] have been tested. Recently, enzymes catalyzing the final steps of the Ehrlich pathway [396] in yeast were recruited in *E. coli* to convert 2-ketoacids into 1-butanol and isobutanol [15]. The global aim to converting biomass to energy has led to an increased interest in transferring non-native metabolic pathways and enzymes into industrial production hosts such as *E. coli* [104, 397] or *Saccharomyces cerevisiae* [398].

An important goal of this research requires extending the metabolic confines of microbial hosts by recruiting non-native biosynthetic pathways. So far, studies concerning the incorporation of heterologous pathways relied largely on human intuition and literature reports followed by experimentation [84, 85]. Nowadays, rapidly expanding compilations of biotransformations such as KEGG [86] and BRENDA [87] are increasingly being prospected to identify biosynthetic routes to long-chain alcohols. With a combined size that accounts for over 60,000 enzymatic reactions and 250,000 metabolites, these databases include reactant and product designation, stoichiometric coefficients, organism assignment, and occasional thermodynamic information for pathways [399]. Several optimization and graph-based methods have been employed to computationally assemble novel biochemical routes from these sources. Given a set of reactions (i.e., Universal database) the OptStrain [88] procedure uses a mixed-integer linear optimization representation to identify the minimal number of reactions to be added (i.e. knock-ins) into a genome-scale metabolic model to enable the production of the new molecule. However, the developed universal database, at the time, was limited to only approximately 4,000 reaction entries. The combinatorial nature of the problem poses a significant challenge to the OptStrain methodology as the number of reaction database entries increase from a few to tens of thousands. At the expense of not enforcing stoichiometric balances graph-based algorithms have inherently better-scaling properties for exhaustively identifying all min-path reaction entries that link a source with a target metabolite. Hatzimanikatis *et. al.* [89] introduced a graph-based heuristic approach to identify all possible biosynthetic routes from a given substrate to a target chemical by hypothesized enzymatic reaction rules. Recently, a new scoring algorithm [91] was introduced to evaluate and compare novel pathways generated using enzyme-reaction rules. The identified pathways may involve conversions for which no enzymatic activity has been isolated for before.

While this could shed light to truly novel production avenues, it may be more time-consuming to implement. In addition, several techniques such as PathMiner [92], PathComp [93], Pathway Tools [94, 95], MetaRoute [96], PathFinder [97] and UM-BBD Pathway Prediction System [98] are in use to search for bioconversion routes in reaction databases. Most of these methods, so far, have been employed to aid metabolic pathway reconstructions by matching putative enzymes with reference pathways while their contribution towards strain optimization has so far been limited.

In this work, we introduce a min-path graph procedure for in overcoming the complexity associated with exhaustively identifying all possible ways of linking a source with a target metabolite. The procedure is designed to remain tractable even when reaction database entries reach hundred of thousands. The first step, in this effort involved the incorporation of reaction and metabolite entries from both KEGG [86] and BRENDA [87] databases into a single repository. A customized min-path algorithm [100] is then employed to compute all possible pathways that enable the bio-production of a target alcohol molecule. We further scrutinize the identified pathways by first incorporating them into the genome-scale metabolic model of the production host microorganism and subsequently examining their maximum theoretical yields, number of enzymatic steps needed and cofactor availability. We demonstrate our integrated framework by exploring pathways from pyruvate (produced in *E. coli*) to 1-butanol. We then selectively add one or more of these pathways to the latest genome-scale metabolic model of *E. coli*, *i*AF1260 [49] and use our recent OptForce [196] procedure to predict metabolic interventions (i.e., up-/down-regulations and knockouts).

E.1.2. Methodology

The graph-based procedure discussed here is aimed at elucidating all possible biochemical routes from compounds found in the metabolic network of a desirable production host to a target molecule of interest. Alternatively, the procedure can also be used to track native routes that may increase productivity over known synthesis pathways by restricting the reaction entries to the ones present in the metabolic model of the production host. To provide the search procedure with known metabolic routes, we downloaded the most up-to-date version of the KEGG database [86] and extracted approximately 9,000 reactions and 16,000 metabolites. Unfortunately, the KEGG database does not contain complete production pathways of long-chain alcohols. We therefore, added a few hundred reaction entries from the BRENDA database [87] that are relevant to biofuels production to restore the metabolic connectivity to long-chain alcohols. It is important to note that we did not globally reconcile the entire KEGG database with BRENDA database (containing ~250,000 metabolites and 67,191 reactions). Instead, for all reactions in BRENDA associated with the synthesis of the target alcohol, we manually recorded identifiers for all the reactants, products and stoichiometric coefficients and integrated them with the KEGG entries into a single database.

Sorting out the naming inconsistencies for compounds was the most time consuming step. To accomplish this, we made use of available synonym data from PubChem [400] to arrive at unique metabolite identifiers. Reactions with generic (e.g. metabolites named as “alcohol”, “aldehyde” etc.) descriptions for reactant/product compounds, unknown stoichiometry and the ones that involve macromolecules (e.g. RNAP) were excluded. The integrated database used in this work spans 9,921 reactions and 17,013 metabolites from both BRENDA and KEGG.

We used the min-path procedure as depicted in Figure E.1.1 to trace all possible paths between a source and a target metabolite. We first computationally transformed the information contained within the stoichiometric coefficients (S_{ij}) that track participation of metabolites in reactions into a directed metabolite-to-metabolite graph (N_{ii}) where nodes represent metabolites. A directed arc with a weight of one exists between two nodes if one or more reactions in the

database allow the direct bioconversion from one metabolite to the other. If no such reaction exists then a very large cost value is assigned to signify that their direct interconversion is disallowed. Small molecules (e.g. water, carbon dioxide) and cofactors (e.g. NADP, ATP) are involved in a large number of reactions and thus can link reaction steps that do not share any additional metabolites. We therefore exclude all such associated directed arcs before employing the shortest path algorithm. We next compute all *k*-shortest "loopless" pathways [100] between a source and a target alcohol molecule. We start from the shortest path ($k = 1$) and exhaustively sample the combinatorial space of alternative pathways by subsequently eliminating arcs, one at a time, belonging to the shortest pathway. We recompute the shortest path until we record all "$k - 1$" shortest possible metabolic linkages to the target molecule.

We next evaluated the multiple identified pathways based on criteria such as maximum theoretical yield, number of reaction steps needed and co-factor requirements. Given a choice of a pathway to be added, we use our recent OptForce procedure [196] to identify additional strain manipulations (knockouts, up/down-regulations for fluxes) that guarantee a pre-specified yield for the alcohol molecule. The OptForce procedure uses metabolic flux measurements available for the wild-type strain and identifies which fluxes must depart from the original ranges to ensure the overproduction target for the desired alcohol molecule. Based on these necessary network changes, we combinatorially identify the minimal set of engineering interventions that result in a new flux distribution consistent with an overproducing strain of host microbe. All lexicographic searches needed to integrate database entries were performed using Python (version 2.4.3) and the algorithm for the identification of shortest paths was coded using C++ on a 2.6 GHz AMD Opteron Processor with 32 GB of ECC RAM.

E.1.3. Results

In this section, we demonstrate our min-path procedure by identifying all synthesis routes using KEGG and BRENDA database entries for producing 1-butanol from pyruvate. We first select promising pathways and subsequently integrate them with the genome-scale metabolic model of *E. coli*, *i*AF1260 [49]. Using OptForce [196] we next pinpoint metabolic engineering strategies for overproduction. Traditionally, two distinct synthesis routes have been employed in *E. coli* for the production of 1-butanol. The first pathway involves a fermentative transformation of pyruvate and acetyl-CoA to 1-butanol by the action enzymes from *C. acetobutylicum* [397]. The second pathway takes advantage of enzymes with broad-range substrate specificity to convert natural amino acids in *E. coli* into ketoacid precursors [15, 104] and eventually 1-butanol. In both pathways, pyruvate acts as an important precursor and a branching metabolite for butanol synthesis [401]. The fate of pyruvate at the end of glycolysis depends on the engineering strategies imparted to the production host. Therefore, here we selected pyruvate as a source metabolite in exploring pathways to 1-butanol (sink metabolite).

Figure E.1.2 illustrates all identified pathways from pyruvate to 1-butanol using the integrated reaction database. With the exception of the thiobutanoate pathway (present in the BRENDA database), all other pathways involved butanoyl-CoA and 1-butanal as shared intermediates that are converted to 1-butanol using secondary alcohol dehydrogenase (*adhE*) from *C. acetobutylicum*. The min-path procedure recapitulated both the fermentative and ketoacid pathways for 1-butanol synthesis (shown in dotted lines). In addition, the algorithm uncovered a number of possible transformations to butanoyl-CoA involving intermediate metabolites that are produced in *E. coli*. For example, pyruvate can be converted into acetyl-CoA using pyruvate dehydrogenase natively present in *E. coli*. However, the conversion from acetyl-CoA to butanoyl-CoA is not favored because 1-butanal produced along the pathway is used up as a co-reactant along other reactions in the same pathway. This severely reduces the flux of the 1-butanol to less than 10 mmol/g.DW.hr which is about ten times less than the yields from existing pathways [15, 382]. Similarly, pathways involving methylmalate and methylbutanoate as intermediates require

cofactors, which in turn, adversely reduce the yield of 1-butanol. Upon integrating these reactions in the metabolic model of *E. coli*, we estimated that the maximum theoretical yield of 1-butanol synthesis was only around 32 mmol/g.DW.hr.

The thiobutanoate pathway recruits a decarboxylase and a reductase enzyme and defines a novel synthesis route distinct from the two existing pathways. Instead of using dehydrogenases to convert butyraldehyde into 1-butanol, the new pathway proceeds with the transamination of methionine into 2-oxomethylthiobutanoate and eventually into 1-butanol. Notably, a native transaminase (E.C. 2.6.1.42) enzyme in *E. coli* is known to catalyze the conversion of L-methionine to L-glutamate with 2-ketoglutarate as a co-reactant [402]. The intermediate product, 2-methylthiobutanoate, is subsequently decarboxylated (E.C. 4.1.1.72) to 3-methylthiopropanal. This conversion is native in *Lactococcus lactis* [403]. Subsequently, 3-methylthiopropanal is reduced (E.C. 1.1.1.265) to 1-butanol by a reductase present in yeast [404]. It is important to note that the decarboxylase reaction removes a considerable amount carbon in the form of carbon dioxide, reducing the yield of 1-butanol by ~22% in comparison to the ketoacid pathway.

Next, we integrate these reactions in the *i*AF1260 metabolic model of *E. coli* and use OptForce [196] to identify metabolic interventions to meet an imposed overproduction target. The identified results are contrasted against the ones derived when the ketoacid pathway is integrated into the E. coli model. In both the case studies, the initial strain is first characterized by estimating the maximal range of flux variability using the intracellular flux measurements [405] available for the wild-type strain of *E. coli*, BW25113. The OptForce employs a bilevel optimization procedure to first identify the reaction fluxes that must increase or decrease (MUST sets) outside the wild-type flux ranges to meet the overproduction target. A minimal set of direct interventions (i.e. knock-up/down/outs) that guarantee a pre-specified yield for 1-butanol is next extracted from the MUST sets. All abbreviations for reactions and metabolites adhere to the *i*AF1260 metabolic model conventions.

E.1.3.1. Case 1: 1-butanol Synthesis using Thiobutanoate Pathway

Figure E.1.3 lists the identified MUST set of reactions considered one reaction at-a-time. The yield for 1-butanol was set at 95% of its theoretical maximum, while allowing the production of 5% biomass to support growth. The thiobutanoate pathway branches away from 2-ketoglutarate along the oxidative arm of the TCA cycle. In order to increase the pool of oxaloacetate available for the TCA cycle, the fluxes of reactions in the glycolytic pathway (PGI, PGM, PGK, PPC etc.) increase beyond their initial ranges. Many reactions in the pentose phosphate pathway (e.g. GND, TKT1/2, TALA etc.) were also classified in the $MUST^{U}$ sets. The increase in the fluxes for these reactions replenishes the glycolytic intermediary metabolites. Since, methionine is required as an important precursor for 1-butnaol pathway, reactions in methionine biosynthesis (e.g., CYSTL, METS, MTHFR2, CYSS) also members of the $MUST^{U}$ set. The fluxes of reactions leading to competing by-products, pyruvate kinase (PYK) and pyruvate formate lyase (PFL) decrease below their initial ranges. Since biomass production is reduced to 5% of its theoretical maximum, reactions in amino acid biosynthesis that are directly coupled to growth appear in the $MUST^{L}$ sets.

As expected, more complex flux changes are revealed in the network of $MUST^{UU}$, $MUST^{UL}$ and $MUST^{LL}$ sets shown in Figure E.1.4. These results underscore the importance of increasing the flux through the oxidative arm of the TCA cycle (FUM etc.) or at the same time negating the drain towards by-products such as acetate and ethanol. Additionally, in the $MUST^{UU}$ set, the flux of propanoyl CoA:succinyl CoA transferase (PPCSCT) or the flux of succinyl CoA synthetase (SUCOAS) must increase. Both of these fluxes are in close proximity to 2-ketoglutarate, which is an important branching metabolite in the TCA cycle for the thiobutanoate pathway. We carry out this hierarchical classification by considering three reactions at-a-time

(see Figure E.1.5). The increase in fluxes for IPPMI, IMPC and AIRC3 further boosts the synthesis of precursors for methionine through amino acid biosynthetic pathways.

It is to be noted that the MUST set of reactions represent the changes that must take place in the metabolic network for overproduction that can be directly or indirectly imparted by means of metabolic interventions. OptForce identifies the minimal set of reaction interventions (culled from the MUST sets) that forces the target yield for 1-butanol. Figure E.1.6a shows the FORCE set of reactions for overproducing 1-butanol in *E. coli* using the thiobutanoate pathway. Up regulating one of the two glycolytic fluxes, glucose-6-phosphate isomerase (PGI) or phosphoglycerate mutase (PGM), replenishes phosphoenol pyruvate available for the anaplerotic conversion to oxaloacetate. The up-regulation for phosphoenol pyruvate carboxylase (PPC) results in increasing the amount of oxaloacetate for the TCA cycle. Increase in fluxes of PPCSCT or SUCOAS ensure the availability of 2-ketoglutarate for transamination along the thiobutanoate pathway. In addition, the FORCE sets also include knockouts for pyruvate formate lyase (PFL) to reduce the drain towards by-products (acetate and ethanol) and methylenetetrahydrofolate dehydrogenase (MTHFD) to prevent the drain of L-methionine away from the thiobutanoate pathway. These coordinated set of interventions lead to a guaranteed yield for 1-butanol of 73 mmol / g.DW.hr.

E.1.3.2. Case 2: 1-butanol using Ketoacid Pathway

Figure E.1.6b contrasts the metabolic pathways and branching points for the ketoacid and thiobutanoate pathways on a metabolic map of *E. coli*, respectively. hile the thiobutanoate pathway branches out from a TCA cycle intermediate, pyruvate serves as an important precursor for 1-butanol produced via the ketoacid pathway. We integrated the reactions along this pathway to *i*AF1260 metabolic model of *E. coli* and applied our OptForce procedure to predict the MUST sets and subsequently, the FORCE sets. Figure E.1.6b shows the FORCE set of eight engineering interventions for 1-butanol synthesis in *E. coli* using the ketoacid pathway. Herein, OptForce suggested the up-regulation in the fluxes of reactions that convert key amino acids to 1-butanol precursors (i.e., serine deaminase (SERD) and methylglyoxal synthase (MGSA)). Presumably due to the proximity of the ketoacid pathway to the synthesis routes for natural fermentation products (acetate, ethanol, formate, lactate etc.), the down-regulations for pyruvate formate lyase (PFL) and lactate dehydrogenase (LDH) are needed to reduce carbon drain. Additionally, down-regulation of TCA cycle reactions, fumarate reductase (FRD3) and aconitase (ACONTa/b), also appear as essential network changes to ensure overproduction.

A notable difference between the two cases is the down-regulation of phosphogluconate dehydrogenase (GND) using the ketoacid pathway. While the flux of GND must increase for the thiobutanoate pathway (i.e., member of $MUST^{U}$ set), OptForce suggests that its flux must be reduced to facilitate 1-butanol synthesis when using the ketoacid pathway. In addition, while PGI and PGM were identified as up-regulations for the thiobutanoate pathway no glycolytic reactions were up- regulated in the FORCE set for the ketoacid route. Since the ketoacid pathway branches out from precursors synthesized at the end of glycolytic pathway, OptForce indicates that the depletion of carbon can be minimized through a number of down-regulations for competing pathways without the need of overexpressing glycolytic enzymes. However, in the thiobutanoate case, the anaplerotic phosphoenol pyruvate carboxylase (PPC) is required to replenish oxaloacetate and to sustain an increased flux through the TCA cycle.

E.1.4. Discussion and Summary

We have presented a graph-based min-path procedure that combines metabolic information from online databases (KEGG and BRENDA) to identify all possible biochemical synthesis routes to target biofuel candidates. The results for 1-butanol pathways reveal several new heterologous synthesis routes that can be computationally evaluated for overexpression and cloning

experiments. Our algorithm was able to identify existing pathways (ketoacid and fermentative pathways) used for 1-butanol production. Interestingly, we the results also suggested several native synthesis routes to precursors of 1-butanol in *E. coli*. For example, seven pathways from pyruvate to butanoyl-CoA involved intermediate metabolites produced by naturally occurring enzymes in *E. coli*. However, the yield of 1-butanol using these pathways was limited. In addition, the algorithm also uncovered a new alternative route to 1-butanol synthesis through the thiobutanoate pathway. Although, the decarboxylation of methylthiobutanoate reduced 1-butanol production, the computationally derived yield was comparable to the existing strains [15, 382, 397].

The results suggested by our OptForce procedure [196] revealed the differing nature of metabolic interventions required to overproduce 1-butanol using the thiobutanoate and ketoacid pathway. Recruiting the thiobutanoate pathway for 1-butanol overproduction required up-regulations for glycolytic fluxes (PGI, PGM). On the other hand, the ketoacid precursors were made available to 1-butanol synthesis by knocking down competing pathways (PFL, ACONTa/b etc.). The flux changes observed in the MUST sets for the two cases also showcased contrasting patterns. For example, for the thiobutanoate pathway, the fluxes of the pentose phosphate pathway increased so that alternative routes for glutamate and other amino acids are maintained to support growth. Although, none of the reactions from pentose phosphate pathway appeared in the FORCE sets, on the contrary, the OptForce procedure indicated that the fluxes of phosphogluconate dehydrogenase (GND) must be down-regulated while using ketoacid pathway to synthesize 1-butanol.

Several interventions that were identified in the FORCE sets have been used in existing strains to produce 1-butanol. For example, recent strategies to delete host competing pathways encoded by the genes *ldhA, frdBC, pta, pfl* and *adhE* [15, 382, 397] have resulted in a three-fold increase in the yield of 1-butanol. In addition, enhancing glycolytic fluxes by overexpressing NADH-regenerating enzymes were implemented in an *E. coli* strain [406] that yielded 580 mg/L of 1-butanol. In addition to the existing interventions, the OptForce procedure also uncovered new knockouts and up-regulations that coordinate an increased synthesis of 1-butanol. For example, the up-regulation of glycolytic fluxes and phoephoenolpyruvate carboxylase (PPC) increase the amount of oxaloacetate for the TCA cycle. However, in order to effectively utilize the transamination pathway, OptForce suggested up-regulations for PPCSCT and SUCOAS that are in close proximity to the branching thiobutanoate pathway.

The procedure detailed in this work allows for the enumeration of all possible metabolic routes to any target compound. Alternatively, the graph-based procedure can be used to identify alternative synthesis routes found entirely within the production host by selectively exploring pathways that are native. Currently, the procedure uses all the biotransformations found in the KEGG database [86, 407] and a selected set of reactions from the BRENDA [87] database. The min-path search procedure remains tractable for much larger compilations of reactions/metabolites. It is to be noted that the interventions proposed by OptForce pertain to the reactions. A complete mapping between the reactions and the genes is required for projecting the results at the gene-level.

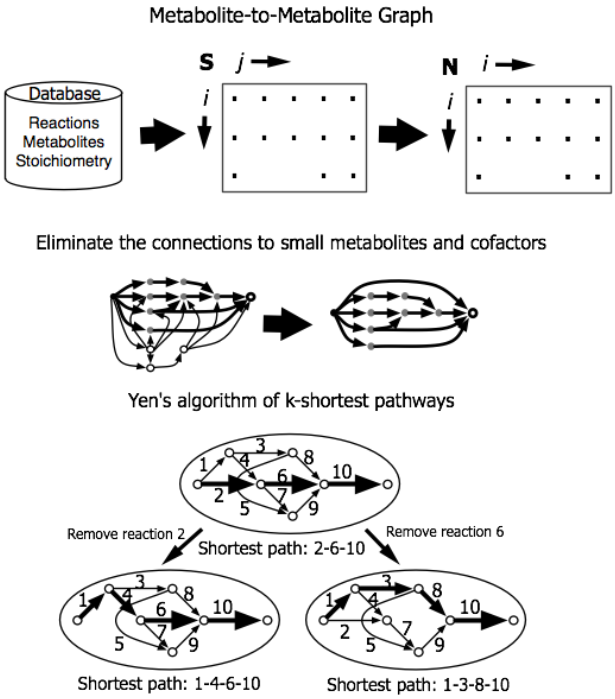

Figure E.1.1: Graph-based procedure to min novel pathways from reaction databases using Yen's shortest path algorithm.

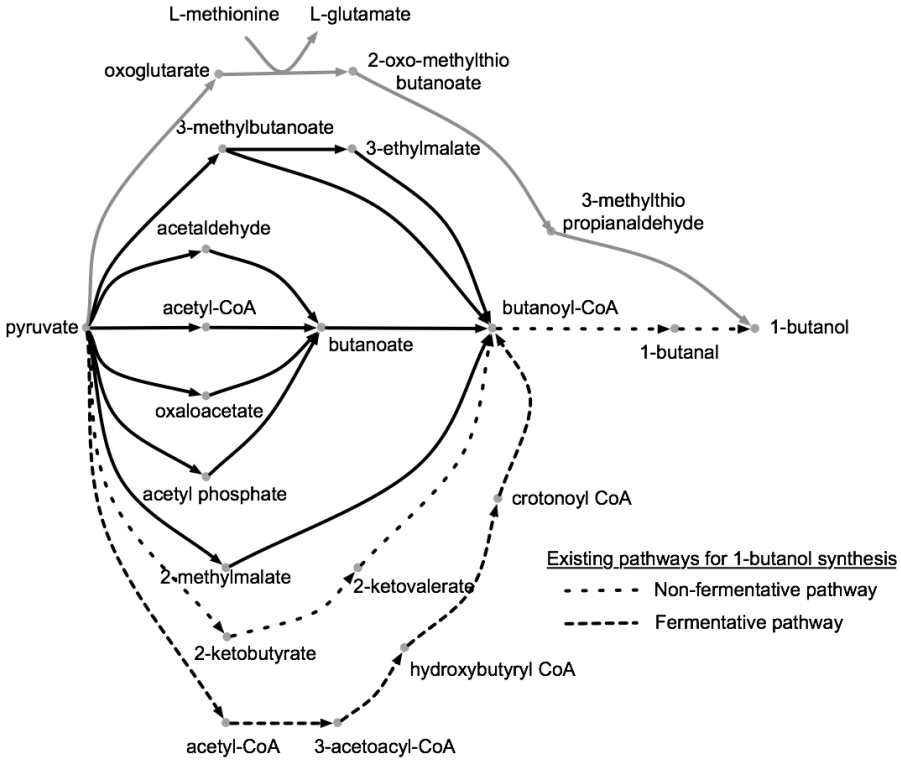

Figure E.1.2: Pathways identified from pyruvate to 1-butanol using the graph-based procedure. Widely spaced dotted arrows represent the ketoacid pathway and the closely spaced arrows represent the fermentative pathways for 1-butanol synthesis. The thiobutanoate pathway is shown in grey.

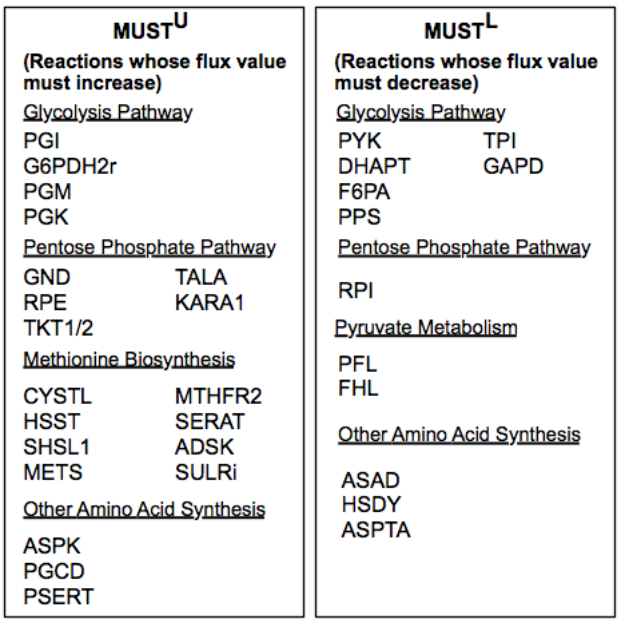

Figure E.1.3: MUST^{U} and MUST^{L} set of reactions for 1-butanol synthesis in *E. coli* using the thiobutanoate pathway.

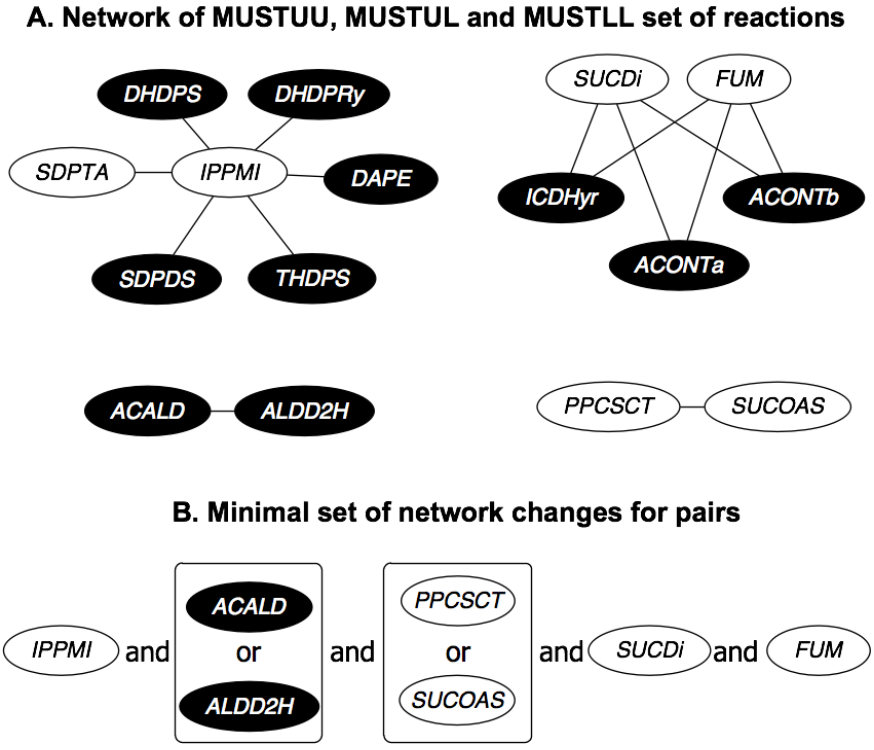

Figure E.1.4: MUST^{UU}, MUST^{UL} and MUST^{LL} set of reactions for 1-butanol synthesis using the thiobutanoate pathway. Black ovals represent reaction flux down-regulations while white ovals denote up-regulations.

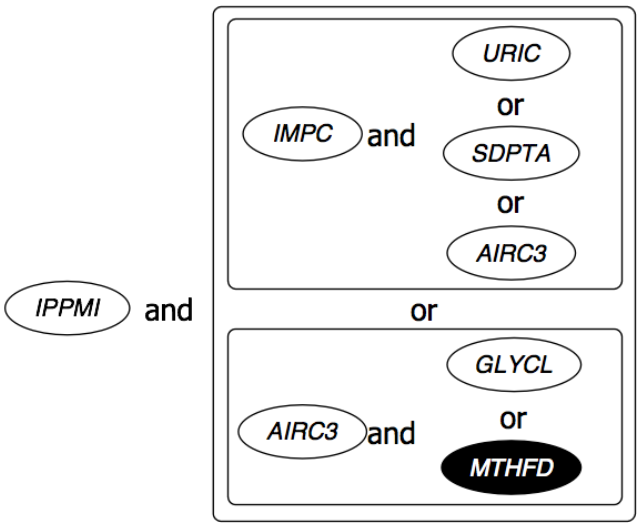

Figure E.1.5: Minimal set of network changes for triples (i.e. MUST^{UUU}, MUST^{UUL}, MUST^{ULL} etc). Reactions whose fluxes must increase are shown in white ovals while reactions whose fluxes decrease are shown in black ovals.

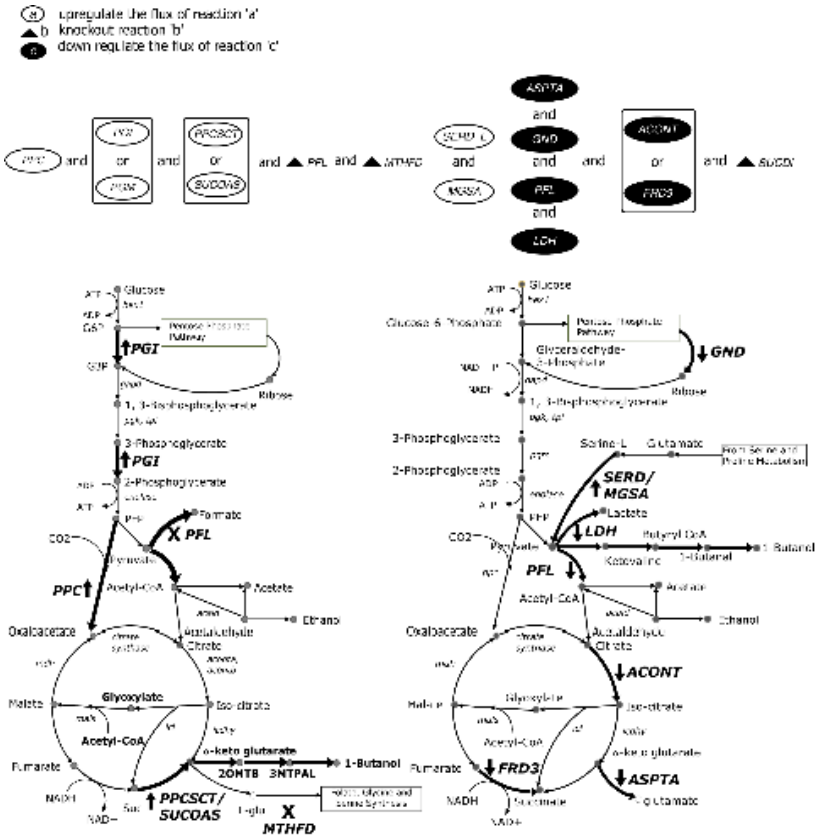

Figure E.1.6: The FORCE set of reactions for 1-butanol synthesis using the thiobutanoate (left) and the ketoacid (right) pathways. All the reaction interventions are shown in bold. Up-regulations are denoted with (↑) symbol, down-regulations are denoted with (↓) symbol and the knockouts are shown with (×) symbol.

E.1 **Ground and transition state computations for enzymatic reactivity and specificity**

E.1.1. Introduction

A challenge in enzyme design is to improve substrate specificity, active site access, and binding while maintaining or even improving transition state stabilization. Because of enzymes' immense potential to provide solutions to challenges in biomass treatment, biosensing and environmental pollutants treatments, the goal of this project is to develop a new computational workflow utilizing highly accurate quantum mechanical methods. This will be done as purely (exhaustive) experimental library screening approaches cannot predictably lead to optimized designs within a reasonable amount of time/cost.

We are addressing this challenge by putting forth and demonstrating an enzyme design workflow relying on computations at multiple states. We are fine-tuning and benchmarking the new computational workflow using the *E. coli* dihydrofolate reductase catalyzing the reduction of dihydrofolate to tetrahydrofolate as the test system given the abundance of available mutant activity data and modeled transition states. Concurrently we are deploying the developed IPRO computational base to re-engineer cytochrome $P450_{BM-3}$ monooxygenase, which is functionally expressed at high levels in *E. coli* and has become a prime target for hydroxylase engineering of small alkanes towards alcohols. $P450_{BM-3}$ is being engineered to hydroxylate ethane, a non-natural substrate. The reaction mechanism for $P450_{BM-3}$ is well established, experimental design attempts exist for comparison, and the system is computationally tractable. From a practical viewpoint, the selective oxidation of light alkanes can produce liquid fuels or value-added chemicals from remote natural gas sources or less valuable refinery by-products. By studying these systems, our goal is to develop and demonstrate a general computational workflow that can create enzymatic activity for a non-natural substrate.

E.2.2. Methodology

We explored the application of molecular mechanics (MM) and quantum mechanically (QM)-parameterized MM calculations to test our computational methodology against existing experimental data prior to moving forward with computational design. Arnold and coworkers used directed evolution to identify a mutant of $P450_{BM-3}$, 535-h, which was capable of hydroxylating ethane to ethanol. This mutant involved 14 amino acid substitutions relative to the wild-type, with 3 mutations occurring in the active site region (Positions 78, 82, 328). Our preliminary binding calculations explored whether the 535-h mutant performance can be explained by improvements in enzyme-ethane binding and enhanced transition state stabilization. A computational saturation mutagenesis procedure written in Python using CHARMM was used to sequentially mutate each one of the 14 positions identified by the Arnold lab in mutant 535-h to every possible amino acid. Interaction energy changes upon mutation were calculated using the generalized born implicit solvent model (GBSW).[408]

E.2.3. Results

In Figure E.2.1, we plot the interaction energy improvement ($-\Delta\Delta G_{calculated}$) compared to the wild-type enzyme for every position and single mutation choice. A positive value in Figure E.2.1 indicates stronger binding of ethane to the mutant as compared to the wild-type enzyme. Looking at the 53-5h mutations (one at a time) arrived at through the directed evolution procedure we find that they are sometimes but not always the most energetically beneficial. In particular, for residues 78 and 328 that are in contact with the substrate (but not for position 82) the identified mutations are near at the interaction energy optimum. These results confirm that energy interactions at the ground state provide only part of the answer to the enzymatic activity level improvement puzzle. We next explored whether transition state energy interactions may provide any missing pieces. We applied DFT calculations to obtain the transition state structure and charge distribution obtained from the DFT-determined transition state to reparameterize the MM force-field for evaluation of transition state stability. This QM-derived MM force-field was then used to estimate the impact of mutations on transition state stability.

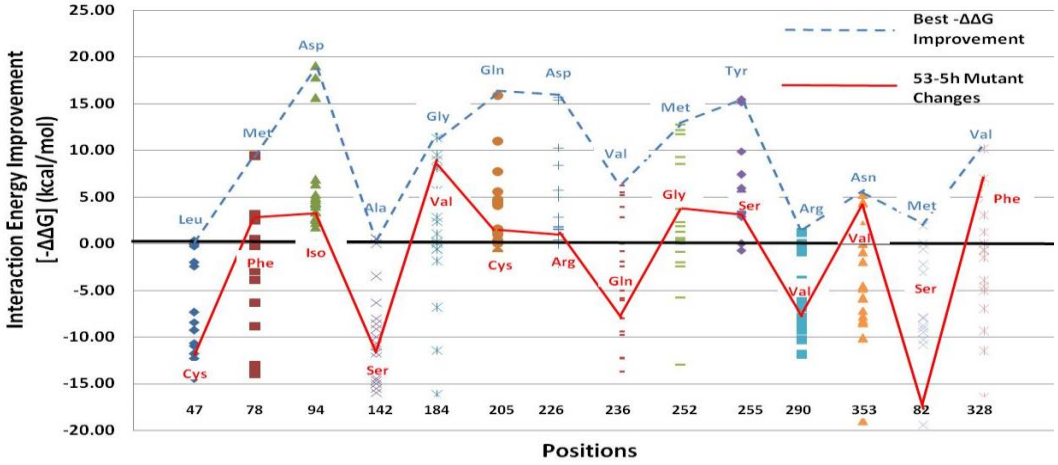

Figure E.2.1: Interaction energy improvement ($-\Delta\Delta G$), compared to the wild-type $P450_{BM-3}$, upon single amino acid mutations at the 14 positions changed in mutant 535-h for the binding of the ground state (ethane) structure. The x-axis value represents the mutated position in the enzyme. The blue (top) amino-acid abbreviations represent the computationally determined optimal mutation at that position, whereas in cases the experimental and computationally optimal mutant differ, red values (bottom) indicate the experimental mutation.

Figure E.2.2 illustrates the results of the computational saturation mutagenesis applied to the transition state, where the interaction energy was calculated exactly the same way as in the ground state calculations. Interaction energy improvements at the TS are significantly higher on average than the corresponding ones at the ground state due to the difference in charge distribution between the ground and transition states. We find that single point mutations (i.e. K94I, A290V, F205C) that had little or no effect on the energy interactions at the ground state provide significant stabilization at the transition state. Conversely, mutations that seem to not make a difference at the transition state (i.e. A184V, A328F) are important for energy stabilization at the ground state. These results demonstrate the complementary nature of GS and TS calculations for explaining and improving enzymatic activity levels.

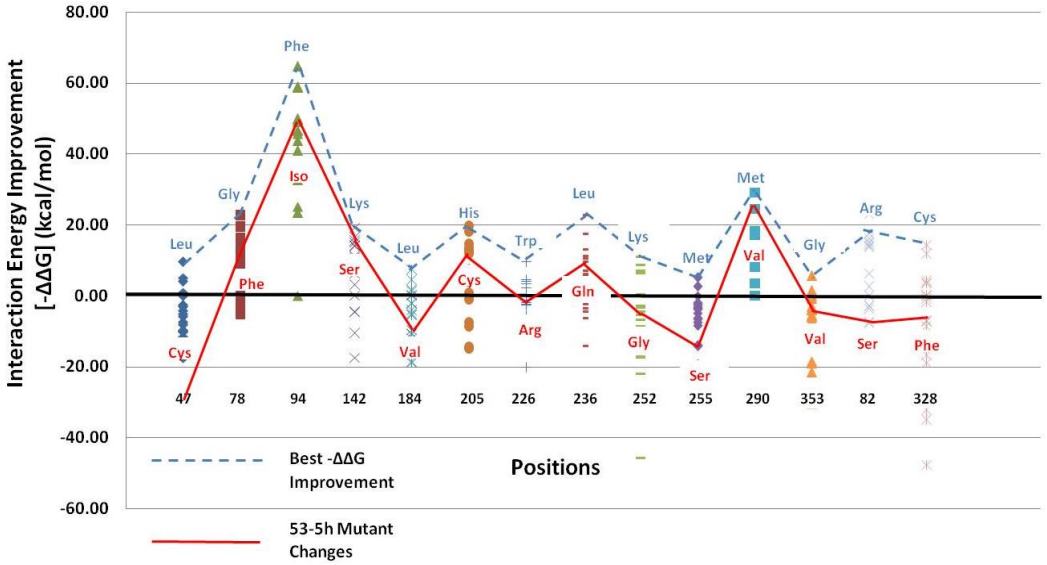

Figure E.2.2: Improvement in interaction energy (-ΔΔG), compared to the wild-type $P450_{BM-3}$, upon single amino acid mutations at the 14 positions changed in mutant 535-h for the transition state structure. Mutations were found that significantly improve the interaction energy between the protein and the transition state structure that were not found to improve the binding of the reactant state (ethane).

Building upon what we've learned from the mutagenic analysis above, we next proceeded to design the P450. We first had to select the positions to be designed. Therefore we proposed a new approach to systematically select design positions. In our approach we employed sequence, structural, and energetic factors. Shannon entropy analysis selected positions with intermediate sequence variability. Next, the distances of the entropically identified positions to the ethane were calculated, and only those within 8Å were selected as part of the final group of design positions. Lastly, we developed and performed a computational alanine scanning mutagenesis mutating every sequence position of P450BM3 to alanine, and identifying which of the positions affected the interaction energy with the ethane most drastically. Design positions that changed the interaction energy by more than 1 standard deviation were considered in the final pool of design positions. Based on the sequence, structure, and energetic factors, as well as knowledge of the active site residues, we refined our final # of design positions to 16 positions.

We next used the IPRO framework running in parallel and with solvation and optimized the interaction energy between the P450 and the ground and the transition states calculated previously. IPRO generated 8 ground state and 6 transition state solutions that optimized the interaction energy between the P450 and the substrates presented in Table E.2.1.

Table E.2.1: IPRO generated designs optimizing the interaction energy between the ground and transition states.

Ground State Designs	***Transition State Designs***
260G	75D, 78K, 82G
88G, 260G	75D, 78K, 82G, 260G
88G, 260G, 327G, 328G	75D, 78K, 82G, 260G, 327G, 328G
88G, 200K, 260G, 327G, 328G	75D, 78K, 82G, 177G, 182K, 260G, 327G, 328G
88G, 177K, 182G, 200K, 260G, 327G, 328G	75D, 78K, 82G, 177G, 182K, 200K, 260G, 327G, 328G
47K, 88G, 177K, 182G, 200K, 260G, 327G, 328G	47H, 75D, 78K, 82G, 177G, 182K, 200K, 260G, 327G, 328G
47K, 88G, 177K, 182G, 200E, 260G, 327G, 328G	
47K, 94R, 88G, 177K, 182G, 200E, 260G, 327G, 328G	

At this stage in the design process, we cannot describe any specific designs in detail without experimental results. Instead, we will highlight some of the general trends found. We are seeing that IPRO predicted more positive and more hydrophobic residues at the ground state. The change in charge can be explained by the partial negative charge on the oxygen portion of the iron-oxo species. For the transition state, the residues predicted were net smaller than the wild-type. Mutations to glycine can be rationalized by the backbone needing more flexibility to conform around the smaller ethane substrate compared to the large fatty acids P450 naturally hydroxylates.

We next employed IPRO using the design positions found by Arnold and coworkers with directed evolution. The goal of this was to compare whether the experimentally-found positions would improve interaction energy and the number of stabilizing residue contacts within 3 angstroms to the ethane relative to the design position selection procedure outlined above.

Wild-type P450 had 11 contacts within 3 angstroms of the ethane. Our best ground state design improved the number of contacts to 17, whereas the experimentally determined positions improved the number of contacts to 18. IPRO using the experimentally derived positions improved the interaction energy by 25.6% relative to the best design predicted by our systematically determined design positions. At the transition state, we observed just the opposite. The IPRO designs using our design positions

improved the number of contacts to 22 from 11, whereas the designs predicted by Arnold and coworkers' design positions improved the number of contacts to 16. Our best design improved the interaction energy by 58.1% relative to the best design predicted with the Meinhold et al. design positions at the transition state. The design positions found experimentally improved the interaction energy the best at the ground state, whereas the systematically selected design positions improved the interaction energy the best at the transition state.

Finally, with several designs found to improve the ground and transition state interactions, we confirmed that the ethane was still capable of entering the binding pocket of the top designs. Figure E.2.3 shows the binding pockets of the best ground and transition state designs using our systematically selected design positions, relative to the wild-type binding pocket. Clearly the substrate can still access the binding pocket to bind/unbind.

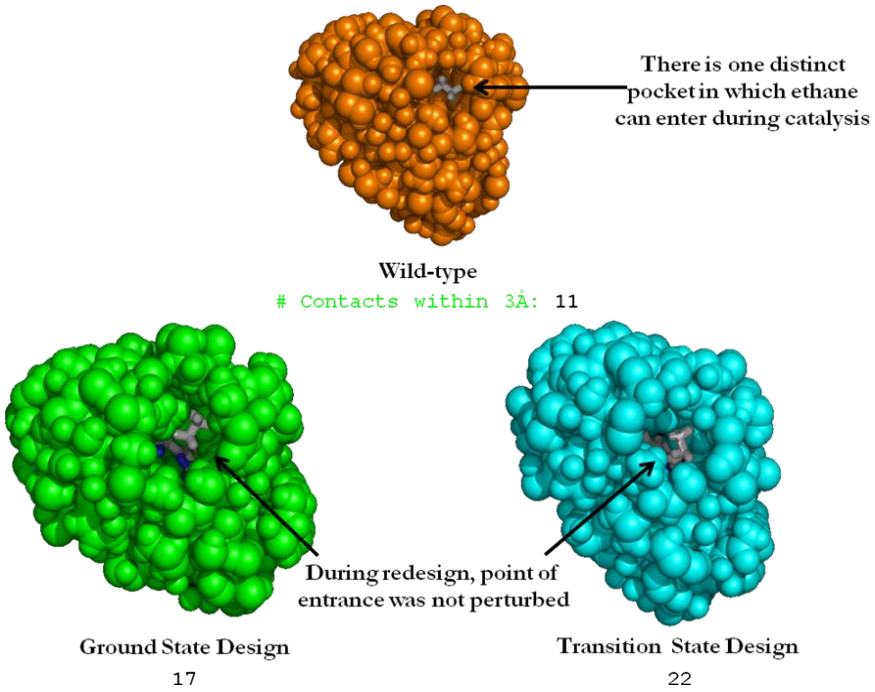

Figure E.2.3: Visual depiction of best ground and transition state binding pockets relative to the wild-type binding pocket. The best designs improved the number of contacts while still allowing the substrate to bind/unbind.

With the shortage of experimental data for this system, the next steps would be to construct the designs predicted both by both sets of design positions for experimental quantitative comparison. These limited number of sequence designs are offered for further experimental study

We are currently carrying out the calculations described above on dihydrofolate reductase (DHFR) to reduce dihydrofolate to tetrahydrofolate. This system is being explored since there is an abundance of experimental mutagenesis data to perform benchmarking on both positive and negative designs. These calculations will lead to experimental constructs of top designs as DHFR is much easier experimental system to

work with. Computational designs leading to improved DHFR activity will validate our preliminary hypothesis that ground and transition state interaction energies are complementary so we can proceed to experimentally construct the best P450 designs.

List of Recent Publications from Research Supported by this Grant

Zomorrodi, A.R. and C.D. Maranas (2012), "OptCom: A multi-level optimization framework for the metabolic modeling and analysis of microbial communities," PLoS Computational Biology, 8(2):e1002363. PMID: 22319433

Kumar, A., P.F. Suthers and C.D. Maranas (2012), "MetRxn: A knowledgebase of metabolites and reactions spanning metabolic models and databases," BMC Bioinformatics, 13(6):e1002363. PMID: 22233419

Saha, R., P.F. Suthers and C.D. Maranas (2011), "Zea mays iRS1563: A Comprehensive Genome-Scale Metabolic Reconstruction of Maize Metabolism.," PLoS ONE, 6(7): e21784. PMID: 21755001

Satish Kumar, V., J.G. Ferry and C.D. Maranas (2011), "Metabolic reconstruction of the archaeon methanogen Methanosarcina acetivorans.," BMC Systems Biology, 5(1): 28. PMID: 21324125

Ravikirthi, P., P.F. Suthers and C.D. Maranas (2011), " Construction of an E. coli Genome-scale Atom Mapping Model for MFA Calculations,," Biotechnology & Bioengineering, 108(6): 1372-82. PMID: 21328316

Xu, P., S. Ranganathan, Z.L. Fowler, C.D. Maranas and M.A.G. Koffas (2011), "Genome-scale metabolic network modeling results in minimal interventions that cooperatively force carbon flux towards malonyl-CoA," Metabolic Engineering, Vol. 13, Issue 5, 578-587, DOI:10.1016/j.ymben.2011.06.008.

Zomorrodi, A.R. and C.D. Maranas (2010), "Improving the iMM904 S. cerevisiae metabolic model using essentiality and synthetic lethality data.," BMC Systems Biology, 4(1): 178. PMID: 21190580

Ranganathan, S. and C.D. Maranas (2010), "Microbial 1-butanol production: Identification of non-native production routes and in silico engineering interventions," Biotechnology Journal, 5: 716-25. PMID: 20665644

Ranganathan, S., P.F. Suthers and C.D. Maranas (2010), "OptForce: An Optimization Procedure for Identifying All Genetic Manipulations Leading to Targeted Overproductions," PLoS Comput. Biol., Vol. 6, No. 4, e100074

Suthers, P.F., Y.J. Chang and C.D. Maranas (2010), "Improved computational performance of MFA using elementary metabolite units and flux coupling," Metabolic Engineering, Vol. 12, 123-128.

Refeences

1. Nakamura, C.E. and G.M. Whited, *Metabolic engineering for the microbial production of 1,3-propanediol.* Curr Opin Biotechnol, 2003. **14**(5): p. 454-9.
2. Misawa, N., S. Yamano, and H. Ikenaga, *Production of beta-carotene in Zymomonas mobilis and Agrobacterium tumefaciens by introduction of the biosynthesis genes from Erwinia uredovora.* Appl Environ Microbiol, 1991. **57**(6): p. 1847-9.
3. Scott, E., F. Peter, and J. Sanders, *Biomass in the manufacture of industrial products-the use of proteins and amino acids.* Appl Microbiol Biotechnol, 2007. **75**(4): p. 751-62.
4. Das, A., S.H. Yoon, S.H. Lee, J.Y. Kim, D.K. Oh, and S.W. Kim, *An update on microbial carotenoid production: application of recent metabolic engineering tools.* Appl Microbiol Biotechnol, 2007. **77**(3): p. 505-12.
5. Nishizaki, T., K. Tsuge, M. Itaya, N. Doi, and H. Yanagawa, *Metabolic engineering of carotenoid biosynthesis in Escherichia coli by ordered gene assembly in Bacillus subtilis.* Appl Environ Microbiol, 2007. **73**(4): p. 1355-61.
6. Sauer, M., D. Porro, D. Mattanovich, and P. Branduardi, *Microbial production of organic acids: expanding the markets.* Trends Biotechnol, 2008. **26**(2): p. 100-108.
7. Thoen, J. and R. Busch, *Industrial Chemicals from Biomass - Industrial Concepts*, in *Biorefineries-Industrial Processes and Products*, B. Kamm, P.R. Gruber, and M. Kamm, Editors. 2006, WILEY-VCH Verlag GmbH & Co. KGaA. p. 347-365.
8. Liu, H., R. Ramnarayanan, and B.E. Logan, *Production of electricity during wastewater treatment using a single chamber microbial fuel cell.* Environmental Sceince and Technology, 2004. **In press**.
9. Bond, D.R., D.E. Holmes, L.M. Tender, and D.R. Lovley, *Electrode-reducing microorganisms that harvest energy from marine sediments.* Science, 2002. **295**(5554): p. 483-5.
10. Bond, D.R. and D.R. Lovley, *Electricity production by Geobacter sulfurreducens attached to electrodes.* Appl Environ Microbiol, 2003. **69**(3): p. 1548-55.
11. Ter Heijne, A., H.V. Hamelers, and C.J. Buisman, *Microbial fuel cell operation with continuous biological ferrous iron oxidation of the catholyte.* Environ Sci Technol, 2007. **41**(11): p. 4130-4.
12. Ishii, S., T. Shimoyama, Y. Hotta, and K. Watanabe, *Characterization of a filamentous biofilm community established in a cellulose-fed microbial fuel cell.* BMC Microbiol, 2008. **8**(1): p. 6.
13. Jarboe, L.R., T.B. Grabar, L.P. Yomano, K.T. Shanmugan, and L.O. Ingram, *Development of ethanologenic bacteria.* Adv Biochem Eng Biotechnol, 2007. **108**: p. 237-61.
14. Atsumi, S., A.F. Cann, M.R. Connor, C.R. Shen, K.M. Smith, M.P. Brynildsen, K.J. Chou, T. Hanai, and J.C. Liao, *Metabolic engineering of Escherichia coli for 1-butanol production.* Metab Eng, 2007.
15. Atsumi, S., T. Hanai, and J.C. Liao, *Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels.* Nature, 2008. **451**(7174): p. 86-9.
16. Ragauskas, A.J., C.K. Williams, B.H. Davison, G. Britovsek, J. Cairney, C.A. Eckert, W.J. Frederick, Jr., J.P. Hallett, D.J. Leak, C.L. Liotta, J.R. Mielenz, R. Murphy, R. Templer, and T. Tschaplinski, *The path forward for biofuels and biomaterials.* Science, 2006. **311**(5760): p. 484-9.
17. 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**(1): p. 6.

18. Liolios, K., N. Tavernarakis, P. Hugenholtz, and N.C. Kyrpides, *The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide.* Nucleic Acids Res, 2006. **34**(Database issue): p. D332-4.
19. Stolyar, S., S. Van Dien, K.L. Hillesland, N. Pinel, T.J. Lie, J.A. Leigh, and D.A. Stahl, *Metabolic modeling of a mutualistic microbial community.* Mol Syst Biol, 2007. **3**: p. 92.
20. Reed, J.L., I. Famili, I. Thiele, and B.O. Palsson, *Towards multidimensional genome annotation.* Nat Rev Genet, 2006. **7**(2): p. 130-41.
21. 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, 2003. **84**(6): p. 647-57.
22. Oliveira, A.P., J. Nielsen, and J. Forster, *Modeling Lactococcus lactis using a genome-scale flux model.* BMC Microbiol, 2005. **5**: p. 39.
23. Alper, H., Y.S. Jin, J.F. Moxley, and G. Stephanopoulos, *Identifying gene targets for the metabolic engineering of lycopene biosynthesis in Escherichia coli.* Metab Eng, 2005. **7**(3): p. 155-64.
24. Pharkya, P. and C.D. Maranas, *An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems.* Metab Eng, 2006. **8**(1): p. 1-13.
25. Almaas, E., B. Kovacs, T. Vicsek, Z.N. Oltvai, and A.L. Barabasi, *Global organization of metabolic fluxes in the bacterium Escherichia coli.* Nature, 2004. **427**(6977): p. 839-43.
26. Burgard, A.P., E.V. Nikolaev, C.H. Schilling, and C.D. Maranas, *Flux coupling analysis of genome-scale metabolic network reconstructions.* Genome Res, 2004. **14**(2): p. 301-12.
27. Motter, A.E., N. Gulbahce, E. Almaas, and A.L. Barabasi, *Predicting synthetic rescues in metabolic networks.* Mol Syst Biol, 2008. **4**: p. 168.
28. Jin, Y.S. and T.W. Jeffries, *Stoichiometric network constraints on xylose metabolism by recombinant Saccharomyces cerevisiae.* Metab Eng, 2004. **6**(3): p. 229-38.
29. Lee, D.Y., L.T. Fan, S. Park, S.Y. Lee, S. Shafie, B. Bertok, and F. Friedler, *Complementary identification of multiple flux distributions and multiple metabolic pathways.* Metab Eng, 2005. **7**(3): p. 182-200.
30. Jamshidi, N. and B.O. Palsson, *Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets.* BMC Syst Biol, 2007. **1**: p. 26.
31. Reed, J.L., T.R. Patel, K.H. Chen, A.R. Joyce, M.K. Applebee, C.D. Herring, O.T. Bui, E.M. Knight, S.S. Fong, and B.O. Palsson, *Systems approach to refining genome annotation.* Proc Natl Acad Sci U S A, 2006. **103**(46): p. 17480-4.
32. Scheer, M., A. Grote, A. Chang, I. Schomburg, C. Munaretto, M. Rother, C. Sohngen, M. Stelzer, J. Thiele, and D. Schomburg, *BRENDA, the enzyme information system in 2011.* Nucleic Acids Res, 2011. **39**(Database issue): p. D670-6.
33. Kanehisa, M., S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa, *KEGG for representation and analysis of molecular networks involving diseases and drugs.* Nucleic Acids Res, 2009. **38**(Database issue): p. D355-60.
34. Caspi, R., T. Altman, J.M. Dale, K. Dreher, C.A. Fulcher, F. Gilham, P. Kaipa, A.S. Karthikeyan, A. Kothari, M. Krummenacker, M. Latendresse, L.A. Mueller, S. Paley, L. Popescu, A. Pujar, A.G. Shearer, P. Zhang, and P.D. Karp, *The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases.* Nucleic Acids Res, 2009. **38**(Database issue): p. D473-9.
35. Lang, M., M. Stelzer, and D. Schomburg, *BKM-react, an integrated biochemical reaction database.* BMC Biochem, 2011. **12**: p. 42.

36. Gao, J., L.B. Ellis, and L.P. Wackett, *The University of Minnesota Biocatalysis/Biodegradation Database: improving public access.* Nucleic Acids Res, 2010. **38**(Database issue): p. D488-91.
37. Feist, A.M., M.J. Herrgard, I. Thiele, J.L. Reed, and B.O. Palsson, *Reconstruction of biochemical networks in microorganisms.* Nat Rev Microbiol, 2009. **7**(2): p. 129-43.
38. Gevorgyan, A., M.G. Poolman, and D.A. Fell, *Detection of stoichiometric inconsistencies in biomolecular models.* Bioinformatics, 2008. **24**(19): p. 2245-51.
39. Notebaart, R.A., F.H. van Enckevort, C. Francke, R.J. Siezen, and B. Teusink, *Accelerating the reconstruction of genome-scale metabolic networks.* BMC Bioinformatics, 2006. **7**: p. 296.
40. Ott, M.A. and G. Vriend, *Correcting ligands, metabolites, and pathways.* BMC Bioinformatics, 2006. **7**: p. 517.
41. Fleischmann, A., M. Darsow, K. Degtyarenko, W. Fleischmann, S. Boyce, K.B. Axelsen, A. Bairoch, D. Schomburg, K.F. Tipton, and R. Apweiler, *IntEnz, the integrated relational enzyme database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D434-7.
42. Bairoch, A., *The ENZYME database in 2000.* Nucleic Acids Res, 2000. **28**(1): p. 304-5.
43. Vastrik, I., P. D'Eustachio, E. Schmidt, G. Gopinath, D. Croft, B. de Bono, M. Gillespie, B. Jassal, S. Lewis, L. Matthews, G. Wu, E. Birney, and L. Stein, *Reactome: a knowledge base of biologic pathways and processes.* Genome Biol, 2007. **8**(3): p. R39.
44. Matthews, L., G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein, and P. D'Eustachio, *Reactome knowledgebase of human biological pathways and processes.* Nucleic Acids Res, 2009. **37**(Database issue): p. D619-22.
45. Stobbe, M.D., S.M. Houten, G.A. Jansen, A.H. van Kampen, and P.D. Moerland, *Critical assessment of human metabolic pathway databases: a stepping stone for future integration.* BMC Syst Biol, 2011. **5**: p. 165.
46. Bornstein, B.J., S.M. Keating, A. Jouraku, and M. Hucka, *LibSBML: an API library for SBML.* Bioinformatics, 2008. **24**(6): p. 880-1.
47. Hucka, M., A. Finney, H.M. Sauro, H. Bolouri, J.C. Doyle, H. Kitano, A.P. Arkin, B.J. Bornstein, D. Bray, A. Cornish-Bowden, A.A. Cuellar, S. Dronov, E.D. Gilles, M. Ginkel, V. Gor, Goryanin, II, W.J. Hedley, T.C. Hodgman, J.H. Hofmeyr, P.J. Hunter, N.S. Juty, J.L. Kasberger, A. Kremling, U. Kummer, N. Le Novere, L.M. Loew, D. Lucio, P. Mendes, E. Minch, E.D. Mjolsness, Y. Nakayama, M.R. Nelson, P.F. Nielsen, T. Sakurada, J.C. Schaff, B.E. Shapiro, T.S. Shimizu, H.D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang, *The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models.* Bioinformatics, 2003. **19**(4): p. 524-31.
48. Stromback, L. and P. Lambrix, *Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX.* Bioinformatics, 2005. **21**(24): p. 4401-7.
49. Feist, A.M., C.S. Henry, J.L. Reed, M. Krummenacker, A.R. Joyce, P.D. Karp, L.J. Broadbelt, V. Hatzimanikatis, and B.O. Palsson, *A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information.* Mol Syst Biol, 2007. **3**: p. 121.
50. Radrich, K., Y. Tsuruoka, P. Dobson, A. Gevorgyan, N. Swainston, G. Baart, and J.M. Schwartz, *Integration of metabolic databases for the reconstruction of genome-scale metabolic networks.* BMC Syst Biol, 2010. **4**: p. 114.
51. Pitkanen, E., A. Akerlund, A. Rantanen, P. Jouhten, and E. Ukkonen, *ReMatch: a web-based tool to construct, store and share stoichiometric metabolic models with carbon maps for metabolic flux analysis.* J Integr Bioinform, 2008. **5**(2).

52. Quek, L.E. and L.K. Nielsen, *On the reconstruction of the Mus musculus genome-scale metabolic network model.* Genome Inform, 2008. **21**: p. 89-100.
53. 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, 2008. **4**(2): p. 148-59.
54. Henry, C.S., M. DeJongh, A.A. Best, P.M. Frybarger, B. Linsay, and R.L. Stevens, *High-throughput generation, optimization and analysis of genome-scale metabolic models.* Nat Biotechnol, 2010. **28**(9): p. 977-82.
55. Weininger, D., *SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules.* Journal of Chemical Information and Computer Sciences, 1988. **28**(1): p. 31-36.
56. Weininger, D., A. Weininger, and J.L. Weininger, *SMILES. 2. Algorithm for generation of unique SMILES notation.* Journal of Chemical Information and Computer Sciences, 1989. **29**(2): p. 97-101.
57. *Daylight Theory Manual.* [http://www.daylight.com/dayhtml/doc/theory/].
58. Weininger, D., *SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules.* Journal of chemical information and computer sciences, 1988. **28**(1): p. 31.
59. Weininger, D., A. Weininger, and J. Weininger, *SMILES. 2. Algorithm for generation of unique SMILES notation.* J Chem Inf Comput Sci, 1989. **29**: p. 97 - 101.
60. Varma, A. and B.O. Palsson, *Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110.* Appl Environ Microbiol, 1994. **60**(10): p. 3724-31.
61. Kjeldsen, K.R. and J. Nielsen, *In silico genome-scale reconstruction and validation of the Corynebacterium glutamicum metabolic network.* Biotechnol Bioeng, 2009. **102**(2): p. 583-97.
62. Durot, M., F. Le Fevre, 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, 2008. **2**: p. 85.
63. Henry, C.S., J.F. Zinner, M.P. Cohoon, and R.L. Stevens, *iBsu1103: a new genome-scale metabolic model of Bacillus subtilis based on SEED annotations.* Genome Biol, 2009. **10**(6): p. R69.
64. David, H., I.S. Ozcelik, G. Hofmann, and J. Nielsen, *Analysis of Aspergillus nidulans metabolism at the genome-scale.* BMC Genomics, 2008. **9**: p. 163.
65. 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, 2006. **281**(52): p. 40041-8.
66. 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, 2008. **190**(8): p. 2790-803.
67. Becker, S.A. and B.O. Palsson, *Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation.* BMC Microbiol, 2005. **5**: p. 8.
68. Oh, Y.K., B.O. 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, 2007. **282**(39): p. 28791-9.
69. Nogales, J., B.O. Palsson, and I. Thiele, *A genome-scale metabolic reconstruction of Pseudomonas putida KT2440: iJN746 as a cell factory.* BMC Syst Biol, 2008. **2**: p. 79.

70. Resendis-Antonio, O., J.L. Reed, S. Encarnacion, J. Collado-Vides, and B.O. Palsson, *Metabolic reconstruction and modeling of nitrogen fixation in Rhizobium etli.* PLoS Comput Biol, 2007. **3**(10): p. 1887-95.
71. Duarte, N.C., M.J. Herrgard, and B.O. Palsson, *Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model.* Genome Res, 2004. **14**(7): p. 1298-309.
72. Herrgard, M.J., N. Swainston, P. Dobson, W.B. Dunn, K.Y. Arga, M. Arvas, N. Bluthgen, S. Borger, R. Costenoble, M. Heinemann, M. Hucka, N. Le Novere, P. Li, W. Liebermeister, M.L. Mo, A.P. Oliveira, D. Petranovic, S. Pettifer, E. Simeonidis, K. Smallbone, I. Spasic, D. Weichart, R. Brent, D.S. Broomhead, H.V. Westerhoff, B. Kirdar, M. Penttila, E. Klipp, B.O. Palsson, U. Sauer, S.G. Oliver, P. Mendes, J. Nielsen, and D.B. Kell, *A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology.* Nat Biotechnol, 2008. **26**(10): p. 1155-60.
73. Reed, J.L., T.D. Vo, C.H. Schilling, and B.O. Palsson, *An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR).* Genome Biol, 2003. **4**(9): p. R54.
74. Feist, A.M., J.C. Scholten, B.O. Palsson, F.J. Brockman, and T. Ideker, *Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri.* Mol Syst Biol, 2006. **2**: p. 2006 0004.
75. 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.
76. Sheikh, K., J. Forster, and L.K. Nielsen, *Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of Mus musculus.* Biotechnol Prog, 2005. **21**(1): p. 112-21.
77. Kuepfer, L., U. Sauer, and L.M. Blank, *Metabolic functions of duplicate genes in Saccharomyces cerevisiae.* Genome Res, 2005. **15**(10): p. 1421-30.
78. Forster, J., I. Famili, P. Fu, B.O. Palsson, and J. Nielsen, *Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network.* Genome Res, 2003. **13**(2): p. 244-53.
79. 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, 2007. **97**(4): p. 657-71.
80. Borodina, I., P. Krabben, and J. Nielsen, *Genome-scale analysis of Streptomyces coelicolor A3(2) metabolism.* Genome Res, 2005. **15**(6): p. 820-9.
81. Schilling, C.H., M.W. Covert, I. Famili, G.M. Church, J.S. Edwards, and B.O. Palsson, *Genome-scale metabolic model of Helicobacter pylori 26695.* J Bacteriol, 2002. **184**(16): p. 4582-93.
82. Lee, J., H. Yun, A.M. Feist, B.O. Palsson, and S.Y. Lee, *Genome-scale reconstruction and in silico analysis of the Clostridium acetobutylicum ATCC 824 metabolic network.* Appl Microbiol Biotechnol, 2008. **80**(5): p. 849-62.
83. Roberts, S.B., C.M. Gowen, J.P. Brooks, and S.S. Fong, *Genome-scale metabolic analysis of Clostridium thermocellum for bioethanol production.* BMC Syst Biol, 2010. **4**: p. 31.
84. Bode, H.B. and R. Muller, *The impact of bacterial genomics on natural product research.* Angew Chem Int Ed Engl, 2005. **44**(42): p. 6828-46.
85. Wenzel, S.C. and R. Muller, *Recent developments towards the heterologous expression of complex bacterial natural product biosynthetic pathways.* Curr Opin Biotechnol, 2005. **16**(6): p. 594-606.
86. Kanehisa, M., M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi, *KEGG for linking genomes to life and the environment.* Nucleic Acids Res, 2008. **36**(Database issue): p. D480-4.

87. Chang, A., M. Scheer, A. Grote, I. Schomburg, and D. Schomburg, *BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009.* Nucleic Acids Res, 2009. **37**(Database issue): p. D588-92.
88. Pharkya, P., A.P. Burgard, and C.D. Maranas, *OptStrain: a computational framework for redesign of microbial production systems.* Genome Res, 2004. **14**(11): p. 2367-76.
89. Hatzimanikatis, V., C. Li, J.A. Ionita, C.S. Henry, M.D. Jankowski, and L.J. Broadbelt, *Exploring the diversity of complex metabolic networks.* Bioinformatics, 2005. **21**(8): p. 1603-9.
90. Henry, C.S., L.J. Broadbelt, and V. Hatzimanikatis, *Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate.* Biotechnol Bioeng, 2010. **106**(3): p. 462-73.
91. Cho, A., H. Yun, J.H. Park, S.Y. Lee, and S. Park, *Prediction of novel synthetic pathways for the production of desired chemicals.* BMC Systems Biology, 2010. **4**(35).
92. McShan, D.C., S. Rao, and I. Shah, *PathMiner: predicting metabolic pathways by heuristic search.* Bioinformatics, 2003. **19**(13): p. 1692-8.
93. Kanehisa, M., S. Goto, M. Hattori, K.F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa, *From genomics to chemical genomics: new developments in KEGG.* Nucleic Acids Res, 2006. **34**(Database issue): p. D354-7.
94. Karp, P.D., S. Paley, and P. Romero, *The Pathway Tools software.* Bioinformatics, 2002. **18 Suppl 1**: p. S225-32.
95. Karp, P.D., S.M. Paley, M. Krummenacker, M. Latendresse, J.M. Dale, T.J. Lee, P. Kaipa, F. Gilham, A. Spaulding, L. Popescu, T. Altman, I. Paulsen, I.M. Keseler, and R. Caspi, *Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology.* Brief Bioinform, 2009.
96. Blum, T. and O. Kohlbacher, *MetaRoute: fast search for relevant metabolic routes for interactive network navigation and visualization.* Bioinformatics, 2008. **24**(18): p. 2108-9.
97. Goesmann, A., M. Haubrock, F. Meyer, J. Kalinowski, and R. Giegerich, *PathFinder: reconstruction and dynamic visualization of metabolic pathways.* Bioinformatics, 2002. **18**(1): p. 124-9.
98. Ellis, L.B., D. Roe, and L.P. Wackett, *The University of Minnesota Biocatalysis/Biodegradation Database: the first decade.* Nucleic Acids Res, 2006. **34**(Database issue): p. D517-21.
99. 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.
100. Yen, J.Y., *Finding K Shortest Loopless Paths in a Network.* Management Science Series a-Theory, 1971. **17**(11): p. 712-716.
101. Atsumi, S., A.F. Cann, M.R. Connor, C.R. Shen, K.M. Smith, M.P. Brynildsen, K.J.Y. Chou, T. Hanai, and J.C. Liao, *Metabolic engineering of Escherichia coli for 1-butanol production.* Metabolic Engineering, 2008. **10**(6): p. 305-311.
102. Formanek, J., R. Mackie, and H.P. Blaschek, *Enhanced Butanol Production by Clostridium beijerinckii BA101 Grown in Semidefined P2 Medium Containing 6 Percent Maltodextrin or Glucose.* Appl Environ Microbiol, 1997. **63**(6): p. 2306-10.
103. Lee, J.Y., Y.S. Jang, J. Lee, E.T. Papoutsakis, and S.Y. Lee, *Metabolic engineering of Clostridium acetobutylicum M5 for highly selective butanol production.* Biotechnol J, 2009. **4**(10): p. 1432-40.
104. Shen, C.R. and J.C. Liao, *Metabolic engineering of Escherichia coli for 1-butanol and 1-propanol production via the keto-acid pathways.* Metab Eng, 2008. **10**(6): p. 312-20.
105. Sillers, R., A. Chow, B. Tracy, and E.T. Papoutsakis, *Metabolic engineering of the non-sporulating, non-solventogenic Clostridium acetobutylicum strain M5 to produce butanol*

without acetone demonstrate the robustness of the acid-formation pathways and the importance of the electron balance. Metab Eng, 2008. **10**(6): p. 321-32.
106. Zomorrodi, A.R. and C.D. Maranas, *OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities.* PLoS Comput Biol, 2012. **8**(2): p. e1002363.
107. Follows, M.J., S. Dutkiewicz, S. Grant, and S.W. Chisholm, *Emergent biogeography of microbial communities in a model ocean.* Science, 2007. **315**(5820): p. 1843-6.
108. Warnecke, F., P. Luginbuhl, N. Ivanova, M. Ghassemian, T.H. Richardson, J.T. Stege, M. Cayouette, A.C. McHardy, G. Djordjevic, N. Aboushadi, R. Sorek, S.G. Tringe, M. Podar, H.G. Martin, V. Kunin, D. Dalevi, J. Madejska, E. Kirton, D. Platt, E. Szeto, A. Salamov, K. Barry, N. Mikhailova, N.C. Kyrpides, E.G. Matson, E.A. Ottesen, X. Zhang, M. Hernandez, C. Murillo, L.G. Acosta, I. Rigoutsos, G. Tamayo, B.D. Green, C. Chang, E.M. Rubin, E.J. Mathur, D.E. Robertson, P. Hugenholtz, and J.R. Leadbetter, *Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite.* Nature, 2007. **450**(7169): p. 560-5.
109. Stephanopoulos, G., *Challenges in engineering microbes for biofuels production.* Science, 2007. **315**(5813): p. 801-4.
110. Vinas, M., J. Sabate, C. Guasp, J. Lalucat, and A.M. Solanas, *Culture-dependent and -independent approaches establish the complexity of a PAH-degrading microbial consortium.* Can J Microbiol, 2005. **51**(11): p. 897-909.
111. Peng, R.H., A.S. Xiong, Y. Xue, X.Y. Fu, F. Gao, W. Zhao, Y.S. Tian, and Q.H. Yao, *Microbial biodegradation of polyaromatic hydrocarbons.* FEMS Microbiol Rev, 2008. **32**(6): p. 927-55.
112. Katsuyama, C., S. Nakaoka, Y. Takeuchi, K. Tago, M. Hayatsu, and K. Kato, *Complementary cooperation between two syntrophic bacteria in pesticide degradation.* J Theor Biol, 2009. **256**(4): p. 644-54.
113. Wagner, M. and A. Loy, *Bacterial community composition and function in sewage treatment systems.* Curr Opin Biotechnol, 2002. **13**(3): p. 218-27.
114. Daims, H., M.W. Taylor, and M. Wagner, *Wastewater treatment: a model system for microbial ecology.* Trends Biotechnol, 2006. **24**(11): p. 483-9.
115. Rittmann, B.E., M. Hausner, F. Loffler, N.G. Love, G. Muyzer, S. Okabe, D.B. Oerther, J. Peccia, L. Raskin, and M. Wagner, *A vista for microbial ecology and environmental biotechnology.* Environ Sci Technol, 2006. **40**(4): p. 1096-103.
116. Sabra, W., D. Dietz, D. Tjahjasari, and A.P. Zeng, *Biosystems analysis and engineering of microbial consortia for industrial biotechnology.* Eng Life Sci, 2010. **10**(5): p. 407-421.
117. Hansen, S.K., P.B. Rainey, J.A. Haagensen, and S. Molin, *Evolution of species interactions in a biofilm community.* Nature, 2007. **445**(7127): p. 533-6.
118. Losos, J.B., M. Leal, R.E. Glor, K. De Queiroz, P.E. Hertz, L. Rodriguez Schettino, A.C. Lara, T.R. Jackman, and A. Larson, *Niche lability in the evolution of a Caribbean lizard community.* Nature, 2003. **424**(6948): p. 542-5.
119. Kerr, B., M.A. Riley, M.W. Feldman, and B.J. Bohannan, *Local dispersal promotes biodiversity in a real-life game of rock-paper-scissors.* Nature, 2002. **418**(6894): p. 171-4.
120. Tilman, D., *Niche tradeoffs, neutrality, and community structure: a stochastic theory of resource competition, invasion, and community assembly.* Proc Natl Acad Sci U S A, 2004. **101**(30): p. 10854-61.
121. Xavier, J.B., *Social interaction in synthetic and natural microbial communities.* Mol Syst Biol, 2011. **7**: p. 483.
122. Fuhrman, J.A., *Microbial community structure and its functional implications.* Nature, 2009. **459**(7244): p. 193-9.

123. DeLong, E.F., *The microbial ocean from genomes to biomes.* Nature, 2009. **459**(7244): p. 200-6.
124. Lozupone, C.A., M. Hamady, B.L. Cantarel, P.M. Coutinho, B. Henrissat, J.I. Gordon, and R. Knight, *The convergence of carbohydrate active gene repertoires in human gut microbes.* Proc Natl Acad Sci U S A, 2008. **105**(39): p. 15076-81.
125. Mo, M.L., N. Jamshidi, and B.O. Palsson, *A genome-scale, constraint-based approach to systems biology of human metabolism.* Mol Biosyst, 2007. **3**(9): p. 598-603.
126. Dobson, P.D., K. Smallbone, D. Jameson, E. Simeonidis, K. Lanthaler, P. Pir, C. Lu, N. Swainston, W.B. Dunn, P. Fisher, D. Hull, M. Brown, O. Oshota, N.J. Stanford, D.B. Kell, R.D. King, S.G. Oliver, R.D. Stevens, and P. Mendes, *Further developments towards a genome-scale metabolic model of yeast.* BMC Syst Biol, 2010. **4**: p. 145.
127. Bizukojc, M., D. Dietz, J. Sun, and A.P. Zeng, *Metabolic modelling of syntrophic-like growth of a 1,3-propanediol producer, Clostridium butyricum, and a methanogenic archeon, Methanosarcina mazei, under anaerobic conditions.* Bioprocess Biosyst Eng, 2010. **33**(4): p. 507-23.
128. Lewis, N.E., G. Schramm, A. Bordbar, J. Schellenberger, M.P. Andersen, J.K. Cheng, N. Patel, A. Yee, R.A. Lewis, R. Eils, R. Konig, and B.O. Palsson, *Large-scale in silico modeling of metabolic interactions between cell types in the human brain.* Nat Biotechnol, 2010. **28**(12): p. 1279-85.
129. Bordbar, A., A.M. Feist, R. Usaite-Black, J. Woodcock, B.O. Palsson, and I. Famili, *A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology.* BMC Syst Biol, 2011. **5**: p. 180.
130. Tzamali, T., P. Poirazi, I. Tollis, and M. Reczko, *Computational identification of bacterial communities.* Int J Biol Life Sci, 2009. **1**(4): p. 185-191.
131. Tzamali, E., P. Poirazi, I.G. Tollis, and M. Reczko, *A computational exploration of bacterial metabolic diversity identifying metabolic interactions and growth-efficient strain communities.* BMC Syst Biol, 2011. **5**(1): p. 167.
132. Wintermute, E.H. and P.A. Silver, *Emergent cooperation in microbial metabolism.* Mol Syst Biol, 2010. **6**: p. 407.
133. Segre, D., D. Vitkup, and G.M. Church, *Analysis of optimality in natural and perturbed metabolic networks.* Proc Natl Acad Sci U S A, 2002. **99**(23): p. 15112-7.
134. Klitgord, N. and D. Segre, *Environments that induce synthetic microbial ecosystems.* PLoS Comput Biol, 2010. **6**(11): p. e1001002.
135. Zhuang, K., M. Izallalen, P. Mouser, H. Richter, C. Risso, R. Mahadevan, and D.R. Lovley, *Genome-scale dynamic modeling of the competition between Rhodoferax and Geobacter in anoxic subsurface environments.* ISME J, 2010.
136. Mahadevan, R., J.S. Edwards, and F.J. Doyle, 3rd, *Dynamic flux balance analysis of diauxic growth in Escherichia coli.* Biophys J, 2002. **83**(3): p. 1331-40.
137. Salimi, F., K. Zhuang, and R. Mahadevan, *Genome-scale metabolic modeling of a clostridial co-culture for consolidated bioprocessing.* Biotechnol J, 2010. **5**(7): p. 726-38.
138. Borenstein, E. and M.W. Feldman, *Topological signatures of species interactions in metabolic networks.* J Comput Biol, 2009. **16**(2): p. 191-200.
139. Chuang, J.S., O. Rivoire, and S. Leibler, *Simpson's paradox in a synthetic microbial system.* Science, 2009. **323**(5911): p. 272-5.
140. Chuang, J.S., O. Rivoire, and S. Leibler, *Cooperation and Hamilton's rule in a simple synthetic microbial system.* Mol Syst Biol, 2010. **6**: p. 398.
141. Freilich, S., A. Kreimer, E. Borenstein, N. Yosef, R. Sharan, U. Gophna, and E. Ruppin, *Metabolic-network-driven analysis of bacterial ecological strategies.* Genome Biol, 2009. **10**(6): p. R61.

142. Freilich, S., A. Kreimer, I. Meilijson, U. Gophna, R. Sharan, and E. Ruppin, *The large-scale organization of the bacterial network of ecological co-occurrence interactions.* Nucleic Acids Res, 2010. **38**(12): p. 3857-68.
143. Frey, E., *Evolutionary game theory: Theoretical concepts and applications to microbial communities.* Physica A, 2010. **389**(20): p. 4265-4298.
144. Gore, J., H. Youk, and A. van Oudenaarden, *Snowdrift game dynamics and facultative cheating in yeast.* Nature, 2009. **459**(7244): p. 253-6.
145. Lehmann, L. and L. Keller, *The evolution of cooperation and altruism--a general framework and a classification of models.* J Evol Biol, 2006. **19**(5): p. 1365-76.
146. Nadell, C.D., K.R. Foster, and J.B. Xavier, *Emergence of spatial structure in cell groups and the evolution of cooperation.* PLoS Comput Biol, 2010. **6**(3): p. e1000716.
147. Schuster, S., J.U. Kreft, N. Brenner, F. Wessely, G. Theissen, E. Ruppin, and A. Schroeter, *Cooperation and cheating in microbial exoenzyme production--theoretical analysis for biotechnological applications.* Biotechnol J, 2010. **5**(7): p. 751-8.
148. Shou, W., S. Ram, and J.M. Vilar, *Synthetic cooperation in engineered yeast populations.* Proc Natl Acad Sci U S A, 2007. **104**(6): p. 1877-82.
149. Vallino, J.J., *Modeling microbial consortiums as distributed metabolic networks.* Biol Bull, 2003. **204**(2): p. 174-9.
150. Taffs, R., J.E. Aston, K. Brileya, Z. Jay, C.G. Klatt, S. McGlynn, N. Mallette, S. Montross, R. Gerlach, W.P. Inskeep, D.M. Ward, and R.P. Carlson, *In silico approaches to study mass and energy flows in microbial consortia: a syntrophic case study.* BMC Syst Biol, 2009. **3**: p. 114.
151. Miller, L.D., J.J. Mosher, A. Venkateswaran, Z.K. Yang, A.V. Palumbo, T.J. Phelps, M. Podar, C.W. Schadt, and M. Keller, *Establishment and metabolic analysis of a model microbial community for understanding trophic and electron accepting interactions of subsurface anaerobic environments.* BMC Microbiol, 2010. **10**: p. 149.
152. Kumar, V.S. and C.D. Maranas, *GrowMatch: an automated method for reconciling in silico/in vivo growth predictions.* PLoS Comput Biol, 2009. **5**(3): p. e1000308.
153. 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.
154. Sahinidis, N.V., *BARON: A general purpose global optimization software package.* J Global Optim, 1996. **8**(2): p. 201-205.
155. Pfeiffer, T., S. Schuster, and S. Bonhoeffer, *Cooperation and competition in the evolution of ATP-producing pathways.* Science, 2001. **292**(5516): p. 504-7.
156. Stams, A.J., *Metabolic interactions between anaerobic bacteria in methanogenic environments.* Antonie Van Leeuwenhoek, 1994. **66**(1-3): p. 271-94.
157. Schink, B., *Synergistic interactions in the microbial world.* Antonie Van Leeuwenhoek, 2002. **81**(1-4): p. 257-61.
158. Schink, B., *Syntrophic associations in methanogenic degradation.* Prog Mol Subcell Biol, 2006. **41**: p. 1-19.
159. Dolfing, J., B. Jiang, A.M. Henstra, A.J. Stams, and C.M. Plugge, *Syntrophic growth on formate: a new microbial niche in anoxic environments.* Appl Environ Microbiol, 2008. **74**(19): p. 6126-31.
160. Schink, B., *Energetics of syntrophic cooperation in methanogenic degradation.* Microbiol Mol Biol Rev, 1997. **61**(2): p. 262-80.
161. Garczarek, F., M. Dong, D. Typke, H.E. Witkowska, T.C. Hazen, E. Nogales, M.D. Biggin, and R.M. Glaeser, *Octomeric pyruvate-ferredoxin oxidoreductase from Desulfovibrio vulgaris.* J Struct Biol, 2007. **159**(1): p. 9-18.
162. Ward, D.M., M.J. Ferris, S.C. Nold, and M.M. Bateson, *A natural view of microbial biodiversity within hot spring cyanobacterial mat communities.* Microbiol Mol Biol Rev, 1998. **62**(4): p. 1353-70.

163. van der Meer, M.T., S. Schouten, M.M. Bateson, U. Nubel, A. Wieland, M. Kuhl, J.W. de Leeuw, J.S. Sinninghe Damste, and D.M. Ward, *Diel variations in carbon metabolism by green nonsulfur-like bacteria in alkaline siliceous hot spring microbial mats from Yellowstone National Park.* Appl Environ Microbiol, 2005. **71**(7): p. 3978-86.
164. Steunou, A.S., S.I. Jensen, E. Brecht, E.D. Becraft, M.M. Bateson, O. Kilian, D. Bhaya, D.M. Ward, J.W. Peters, A.R. Grossman, and M. Kuhl, *Regulation of nif gene expression and the energetics of N2 fixation over the diel cycle in a hot spring microbial mat.* ISME J, 2008. **2**(4): p. 364-78.
165. Nold, S.C. and D.M. Ward, *Photosynthate partitioning and fermentation in hot spring microbial mat communities.* Appl Environ Microbiol, 1996. **62**(12): p. 4598-607.
166. Anderson, K.L., T.A. Tayne, and D.M. Ward, *Formation and fate of fermentation products in hot spring cyanobacterial mats.* Appl Environ Microbiol, 1987. **53**(10): p. 2343-52.
167. Sandbeck, K.A. and D.M. Ward, *Fate of immediate methane precursors in low-sulfate, hot-spring algal-bacterial mats.* Appl Environ Microbiol, 1981. **41**(3): p. 775-82.
168. Frund, C. and Y. Cohen, *Diurnal Cycles of Sulfate Reduction under Oxic Conditions in Cyanobacterial Mats.* Appl Environ Microbiol, 1992. **58**(1): p. 70-7.
169. Konopka, A., *Accumulation and utilization of polysaccharide by hot spring phototrophs during a light-dark transition.* FEMS Microbiol Ecol, 1992. **102**(1): p. 27-32.
170. Bateson, M.M. and D.M. Ward, *Photoexcretion and fate of glycolate in a hot spring cyanobacterial mat.* Appl Environ Microbiol, 1988. **54**(7): p. 1738-43.
171. Mahadevan, R., D.R. Bond, J.E. Butler, A. Esteve-Nunez, M.V. Coppi, B.O. Palsson, C.H. Schilling, and D.R. Lovley, *Characterization of metabolism in the Fe(III)-reducing organism Geobacter sulfurreducens by constraint-based modeling.* Appl Environ Microbiol, 2006. **72**(2): p. 1558-68.
172. Yang, T.H., M.V. Coppi, D.R. Lovley, and J. Sun, *Metabolic response of Geobacter sulfurreducens towards electron donor/acceptor variation.* Microb Cell Fact, 2010. **9**: p. 90.
173. Engel, P., R. Kramer, and G. Unden, *Transport of C4-dicarboxylates by anaerobically grown Escherichia coli. Energetics and mechanism of exchange, uptake and efflux.* Eur J Biochem, 1994. **222**(2): p. 605-14.
174. Winfrey, M.R. and J.G. Zeikus, *Anaerobic metabolism of immediate methane precursors in Lake Mendota.* Appl Environ Microbiol, 1979. **37**(2): p. 244-53.
175. Lovley, D.R., D.F. Dwyer, and M.J. Klug, *Kinetic analysis of competition between sulfate reducers and methanogens for hydrogen in sediments.* Appl Environ Microbiol, 1982. **43**(6): p. 1373-9.
176. Lupton, F. and J. Zeikus, *Physiological basis for sulfate-dependent hydrogen competition between sulfidogens and methanogens.* Curr Microbiol, 1984. **11**(1): p. 7-11.
177. Robinson, J. and J. Tiedj, *Competition between sulfate-reducing and methanogenic bacteria for H2 under resting and growing conditions.* Arch. Microbiol., 1984. **137**(1): p. 26-32.
178. Lovley, D.R. and J.G. Ferry, *Production and Consumption of H(2) during Growth of Methanosarcina spp. on Acetate.* Appl Environ Microbiol, 1985. **49**(1): p. 247-9.
179. O'Brien, J.M., R.H. Wolkin, T.T. Moench, J.B. Morgan, and J.G. Zeikus, *Association of hydrogen metabolism with unitrophic or mixotrophic growth of Methanosarcina barkeri on carbon monoxide.* J Bacteriol, 1984. **158**(1): p. 373-5.
180. Phelps, T.J., R. Conrad, and J.G. Zeikus, *Sulfate-Dependent Interspecies H(2) Transfer between Methanosarcina barkeri and Desulfovibrio vulgaris during Coculture Metabolism of Acetate or Methanol.* Appl Environ Microbiol, 1985. **50**(3): p. 589-94.
181. Knorr, A.L., R. Jain, and R. Srivastava, *Bayesian-based selection of metabolic objective functions.* Bioinformatics, 2007. **23**(3): p. 351-7.

182. Ow, D.S., D.Y. Lee, M.G. Yap, and S.K. Oh, *Identification of cellular objective for elucidating the physiological state of plasmid-bearing Escherichia coli using genome-scale in silico analysis.* Biotechnol Prog, 2009. **25**(1): p. 61-7.
183. Pramanik, J. and J.D. Keasling, *Stoichiometric model of Escherichia coli metabolism: incorporation of growth-rate dependent biomass composition and mechanistic energy requirements.* Biotechnol Bioeng, 1997. **56**(4): p. 398-421.
184. Savinell, J.M. and B.O. Palsson, *Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism.* J Theor Biol, 1992. **154**(4): p. 421-54.
185. Schuetz, R., L. Kuepfer, and U. Sauer, *Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli.* Mol Syst Biol, 2007. **3**: p. 119.
186. Feist, A.M. and B.O. Palsson, *The biomass objective function.* Curr Opin Microbiol, 2010. **13**(3): p. 344-349.
187. Burgard, A.P. and C.D. Maranas, *Optimization-based framework for inferring and testing hypothesized metabolic objective functions.* Biotechnol Bioeng, 2003. **82**(6): p. 670-7.
188. Gianchandani, E.P., M.A. Oberhardt, A.P. Burgard, C.D. Maranas, and J.A. Papin, *Predicting biological system objectives de novo from internal state measurements.* BMC Bioinformatics, 2008. **9**: p. 43.
189. Dias, J.M., A. Oehmen, L.S. Serafim, P.C. Lemos, M.A. Reis, and R. Oliveira, *Metabolic modelling of polyhydroxyalkanoate copolymers production by mixed microbial cultures.* BMC Syst Biol, 2008. **2**: p. 59.
190. Mcinerney, M.J., D.A. Amos, K.S. Kealy, and J.A. Palmer, *Synthesis and Function of Polyhydroxyalkanoates in Anaerobic Syntrophic Bacteria.* FEMS Microbiol Rev, 1992. **103**(2-4): p. 195-205.
191. Venkateswaran, S. and A.L. Demain, *The Clostridium Thermocellum-Clostridium Thermosaccharolyticum Ethanol-Production Process - Nutritional Studies and Scale-Down.* Chem Eng Commun, 1986. **45**(1-6): p. 53-60.
192. Ng, T.K., A. Benbassat, and J.G. Zeikus, *Ethanol-Production by Thermophilic Bacteria - Fermentation of Cellulosic Substrates by Cocultures of Clostridium-Thermocellum and Clostridium-Thermohydrosulfuricum.* Appl Environ Microb, 1981. **41**(6): p. 1337-1343.
193. Wiegel, J. and L.G. Ljungdahl, *Thermoanaerobacter-Ethanolicus Gen-Nov, Spec-Nov, a New, Extreme Thermophilic, Anaerobic Bacterium.* Arch Microbiol, 1981. **128**(4): p. 343-348.
194. Lamed, R. and J.G. Zeikus, *Ethanol-Production by Thermophilic Bacteria - Relationship between Fermentation Product Yields of and Catabolic Enzyme-Activities in Clostridium-Thermocellum and Thermoanaerobium-Brockii.* J Bacteriol, 1980. **144**(2): p. 569-578.
195. Demain, A.L., M. Newcomb, and J.H. Wu, *Cellulase, clostridia, and ethanol.* Microbiol Mol Biol Rev, 2005. **69**(1): p. 124-54.
196. Ranganathan, S., P.F. Suthers, and C.D. Maranas, *OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions.* PLoS Comput Biol, 2010. **6**(4): p. e1000744.
197. Saha, R., P.F. Suthers, and C.D. Maranas, *Zea mays iRS1563: a comprehensive genome-scale metabolic reconstruction of maize metabolism.* PLoS One, 2011. **6**(7): p. e21784.
198. Bennetzen, J.L., S. Hake, and SpringerLink (Online service), *Handbook of Maize Genetics and Genomics*. 2009, Springer New York: New York, NY.
199. Sanchez, O.J. and C.A. Cardona, *Trends in biotechnological production of fuel ethanol from different feedstocks.* Bioresource Technology, 2008. **99**(13): p. 5270-5295.
200. Farrell, A.E., R.J. Plevin, B.T. Turner, A.D. Jones, M. O'Hare, and D.M. Kammen, *Ethanol can contribute to energy and environmental goals.* Science, 2006. **311**(5760): p. 506-508.
201. Stewart, C.N., Jr., *Biofuels and biocontainment.* Nat Biotechnol, 2007. **25**(3): p. 283-4.

202. Mechin, V., O. Argillier, F. Rocher, Y. Hebert, I. Mila, B. Pollet, Y. Barriere, and C. Lapierre, *In search of a maize ideotype for cell wall enzymatic degradability using histological and biochemical lignin characterization.* J Agric Food Chem, 2005. **53**(15): p. 5872-81.
203. Dennis, C. and C. Surridge, *A. thaliana genome.* Nature, 2000. **408**(6814): p. 791-791.
204. Yu, J., S.N. Hu, J. Wang, G.K.S. Wong, S.G. Li, and a.l. et, *A draft sequence of the rice genome (Oryza sativa L. ssp indica).* Science, 2002. **296**(5565): p. 79-92.
205. Goff, S.A., D. Ricke, T.H. Lan, G. Presting, R.L. Wang, and a.l. et, *A draft sequence of the rice genome (Oryza sativa L. ssp japonica).* Science, 2002. **296**(5565): p. 92-100.
206. Paterson, A.H., J.E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, and a.l. et, *The Sorghum bicolor genome and the diversification of grasses.* Nature, 2009. **457**(7229): p. 551-556.
207. Schnable, P.S., D. Ware, R.S. Fulton, J.C. Stein, F. Wei, and a.l. et, *The B73 maize genome: complexity, diversity, and dynamics.* Science, 2009. **326**(5956): p. 1112-5.
208. Xavier Argout, J.S., Jean Marc Aury, Gaetan Droc, Jerome Gouzy, et al, *Deciphering the genome structure and paleohistory of Theobroma cacao.* Nature Proceedings, 2010.
209. Dal'Molin, C.G.D., L.E. Quek, R.W. Palfreyman, S.M. Brumbley, and L.K. Nielsen, *AraGEM, a Genome-Scale Reconstruction of the Primary Metabolic Network in Arabidopsis.* Plant Physiology, 2010. **152**(2): p. 579-589.
210. Sweetlove, L.J., R.L. Last, and A.R. Fernie, *Predictive metabolic engineering: A goal for systems biology.* Plant Physiology, 2003. **132**(2): p. 420-425.
211. Gutierrez, R.A., D.E. Shasha, and G.M. Coruzzi, *Systems biology for the virtual plant.* Plant Physiology, 2005. **138**(2): p. 550-554.
212. Feist, A.M., M.J. Herrgard, I. Thiele, J.L. Reed, and B.O. Palsson, *Reconstruction of biochemical networks in microorganisms.* Nature Reviews Microbiology, 2009. **7**(2): p. 129-143.
213. Park, J.M., T.Y. Kim, and S.Y. Lee, *Constraints-based genome-scale metabolic simulation for systems metabolic engineering.* Biotechnology Advances, 2009. **27**(6): p. 979-988.
214. 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, 2009. **4**(12): p. 1653-70.
215. Poolman, M.G., L. Miguet, L.J. Sweetlove, and D.A. Fell, *A Genome-Scale Metabolic Model of Arabidopsis and Some of Its Properties.* Plant Physiology, 2009. **151**(3): p. 1570-1581.
216. Grafahrend-Belau, E., F. Schreiber, D. Koschutzki, and B.H. Junker, *Flux Balance Analysis of Barley Seeds: A Computational Approach to Study Systemic Properties of Central Metabolism.* Plant Physiology, 2009. **149**(1): p. 585-598.
217. Dal'Molin, C.G.D., L.E. Quek, R.W. Palfreyman, S.M. Brumbley, and L.K. Nielsen, *C4GEM, a Genome-Scale Metabolic Model to Study C-4 Plant Metabolism.* Plant Physiology, 2010. **154**(4): p. 1871-1885.
218. Pilalis, E., A. Chatziioannou, B. Thomasset, and F. Kolisis, *An in silico compartmentalized metabolic model of Brassica napus enables the systemic study of regulatory aspects of plant central metabolism.* Biotechnol Bioeng.
219. Bennett, M.D., I.J. Leitch, H.J. Price, and J.S. Johnston, *Comparisons with Caenorhabditis (approximately 100 Mb) and Drosophila (approximately 175 Mb) using flow cytometry show genome size in Arabidopsis to be approximately 157 Mb and thus approximately 25% larger than the Arabidopsis genome initiative estimate of approximately 125 Mb.* Ann Bot, 2003. **91**(5): p. 547-57.
220. Satish Kumar, V., M.S. Dasika, and C.D. Maranas, *Optimization based automated curation of metabolic reconstructions.* BMC bioinformatics, 2007. **8**: p. 212.

221. Liang, C., L. Mao, D. Ware, and L. Stein, *Evidence-based gene predictions in plant genomes.* Genome Res, 2009. **19**(10): p. 1912-23.
222. Salamov, A.A. and V.V. Solovyev, *Ab initio gene finding in Drosophila genomic DNA.* Genome Res, 2000. **10**(4): p. 516-22.
223. 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, 2009. **5**(2): p. e1000285.
224. Penningd, F.W., A.H. Brunstin, and H.H. Vanlaar, *Products, Requirements and Efficiency of Biosynthesis - Quantitative Approach.* Journal of Theoretical Biology, 1974. **45**(2): p. 339-377.
225. Spector, W.S., *Handbook of biological data*. 1956, Philadelphia,: Saunders. xxxvi, 584 p.
226. Muller, F., Dijkhuis, DJ, Heida, YS, *On the relationship between chemical composition and digestibility in vivo of roughages.* Agricultural Research Report 1970. **736**: p. 1-27.
227. Wedig, C., Jaster, EH, Moore, KJ, *Hemicellulose monosaccharide composition and in vitro disappearance of orchard grass and alfalfa hay.* Journal of Agricultaral and Food Chemistry, 1987. **35**(2): p. 23-27.
228. Sun, Q., B. Zybailov, W. Majeran, G. Friso, P.D.B. Olinares, and K.J. van Wijk, *PPDB, the Plant Proteomics Database at Cornell.* Nucleic Acids Research, 2009. **37**: p. D969-D974.
229. Heazlewood, J.L., R.E. Verboom, J. Tonti-Filippini, I. Small, and A.H. Millar, *SUBA: the Arabidopsis Subcellular Database.* Nucleic Acids Res, 2007. **35**(Database issue): p. D213-8.
230. Volk, R.J. and W.A. Jackson, *Photorespiratory Phenomena in Maize - Oxygen-Uptake, Isotope Discrimination, and Carbon-Dioxide Efflux.* Plant Physiology, 1972. **49**(2): p. 218-&.
231. Dai, Z.Y., M.S.B. Ku, and G.E. Edwards, *C-4 Photosynthesis - the Effects of Leaf Development on the Co2-Concentrating Mechanism and Photorespiration in Maize.* Plant Physiology, 1995. **107**(3): p. 815-825.
232. Jolivettournier, P. and R. Gerster, *Incorporation of Oxygen into Glycolate, Glycine, and Serine during Photorespiration in Maize Leaves.* Plant Physiology, 1984. **74**(1): p. 108-111.
233. Kumar, V.S., M.S. Dasika, and C.D. Maranas, *Optimization based automated curation of metabolic reconstructions.* BMC Bioinformatics, 2007. **8**: p. 212.
234. Wei, Y., M. Lin, D.J. Oliver, and P.S. Schnable, *The roles of aldehyde dehydrogenases (ALDHs) in the PDH bypass of Arabidopsis.* BMC Biochem, 2009. **10**: p. 7.
235. Ouzounis, C.A. and P.D. Karp, *Global properties of the metabolic map of Escherichia coli.* Genome Res, 2000. **10**(4): p. 568-76.
236. Hong, S.J. and C.G. Lee, *Evaluation of central metabolism based on a genomic database of Synechocystis PCC6803.* Biotechnology and Bioprocess Engineering, 2007. **12**(2): p. 165-173.
237. Wise RR, H.J., *Synthesis, export and partitioning of end products of photosynthesis.* Structure and Function of Plastids, ed. H.J. Wise RR. Vol. 23. 2007, Dordrecht, The Netherlands: Springer.
238. Dennis, D.T. and J.A. Miernyk, *Compartmentation of Non-Photosynthetic Carbohydrate-Metabolism.* Annual Review of Plant Physiology and Plant Molecular Biology, 1982. **33**: p. 27-50.
239. Taiz, L. and E. Zeiger, *Plant Physilogy*. Third ed. 2002, Sunderland, Massachusetts: Sinauer Associates, Inc., Publishers.
240. Allen, J.F., *Photosynthesis of ATP - Electrons, proton pumps, rotors, and poise.* Cell, 2002. **110**(3): p. 273-276.

241. Hervas, M., J.A. Navarro, and M.A. De La Rosa, *Electron transfer between membrane complexes and soluble proteins in photosynthesis.* Accounts of Chemical Research, 2003. **36**(10): p. 798-805.
242. Gregory, R., *Biochemistry of Photosynthesis*. 1989, Chichester, NY, USA: John Wiley & Sons.
243. Tsaftaris, A.S., A.M. Bosabalidis, and J.G. Scandalios, *Cell-Type-Specific Gene-Expression and Acatalasemic Peroxisomes in a Null Cat2 Catalase Mutant of Maize.* Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences, 1983. **80**(14): p. 4455-4459.
244. Hisano, H., R. Nandakumar, and Z.Y. Wang, *Genetic modification of lignin biosynthesis for improved biofuel production.* In Vitro Cellular & Developmental Biology-Plant, 2009. **45**(3): p. 306-313.
245. Winkel-Shirley, B., *Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology.* Plant Physiology, 2001. **126**(2): p. 485-493.
246. Styles, E.D. and O. Ceska, *Genetic-Control of 3-Hydroxy-Flavonoids and 3-Deoxy-Flavonoids in Zea-Mays.* Phytochemistry, 1975. **14**(2): p. 413-415.
247. Winkel-Shirley, B., *Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology.* Plant Physiol, 2001. **126**(2): p. 485-93.
248. Weidemann, C., R. Tenhaken, U. Hohl, and W. Barz, *Medicarpin and Maackiain 3-O-Glucoside-6'-O-Malonate Conjugates Are Constitutive Compounds in Chickpea (Cicer-Arietinum L) Cell-Cultures.* Plant Cell Reports, 1991. **10**(6-7): p. 371-374.
249. Vanholme, R., K. Morreel, J. Ralph, and W. Boerjan, *Lignin engineering.* Current Opinion in Plant Biology, 2008. **11**(3): p. 278-285.
250. Sattler, S.E., D.L. Funnell-Harris, and J.F. Pedersen, *Brown midrib mutations and their importance to the utilization of maize, sorghum, and pearl millet lignocellulosic tissues.* Plant Science, 2010. **178**(3): p. 229-238.
251. Marita, J.M., W. Vermerris, J. Ralph, and R.D. Hatfield, *Variations in the cell wall composition of maize brown midrib mutants.* Journal of Agricultural and Food Chemistry, 2003. **51**(5): p. 1313-1321.
252. Kuc, J. and O.E. Nelson, *Abnormal Lignins Produced by Brown-Midrib Mutants of Maize .I. Brown-Midrib-1 Mutant.* Archives of Biochemistry and Biophysics, 1964. **105**(1): p. 103-&.
253. Guillaumie, S., M. Pichon, J.P. Martinant, M. Bosio, D. Goffner, and Y. Barriere, *Differential expression of phenylpropanoid and related genes in brown-midrib bm1, bm2, bm3, and bm4 young near-isogenic maize plants.* Planta, 2007. **226**(1): p. 235-50.
254. Sticklen, M.B., *Expediting the biofuels agenda via genetic manipulations of cellulosic bioenergy crops.* Biofuels Bioproducts & Biorefining-Biofpr, 2009. **3**(4): p. 448-455.
255. Sticklen, M.B., *Plant genetic engineering for biofuel production: towards affordable cellulosic ethanol.* Nat Rev Genet, 2008. **9**(6): p. 433-43.
256. Li, X., J.K. Weng, and C. Chapple, *Improvement of biomass through lignin modification.* Plant Journal, 2008. **54**(4): p. 569-581.
257. Vega-Sanchez, M.E. and P.C. Ronald, *Genetic and biotechnological approaches for biofuel crop improvement.* Current Opinion in Biotechnology, 2010. **21**(2): p. 218-224.
258. Grabber, J.H., P.F. Schatz, H. Kim, F.C. Lu, and J. Ralph, *Identifying new lignin bioengineering targets: 1. Monolignol-substitute impacts on lignin formation and cell wall fermentability.* Bmc Plant Biology, 2010. **10**: p. -.
259. Abramson, M., O. Shoseyov, and Z. Shani, *Plant cell wall reconstruction toward improved lignocellulosic production and processability.* Plant Science, 2010. **178**(2): p. 61-72.

260. Torney, F., L. Moeller, A. Scarpa, and K. Wang, *Genetic engineering approaches to improve bioethanol production from maize.* Current Opinion in Biotechnology, 2007. **18**(3): p. 193-199.
261. Smidansky, E.D., J.M. Martin, L.C. Hannah, A.M. Fischer, and M.J. Giroux, *Seed yield and plant biomass increases in rice are conferred by deregulation of endosperm ADP-glucose pyrophosphorylase.* Planta, 2003. **216**(4): p. 656-664.
262. Kim, J. and J.L. Reed, *OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains.* BMC Syst Biol, 2010. **4**: p. 53.
263. Thiele, I. and B.O. Palsson, *A protocol for generating a high-quality genome-scale metabolic reconstruction.* Nature Protocols, 2010. **5**(1): p. 93-121.
264. Feist, A.M. and B.O. Palsson, *The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli.* Nature Biotechnology, 2008. **26**(6): p. 659-667.
265. Duarte, N.C., S.A. Becker, N. Jamshidi, I. Thiele, M.L. Mo, T.D. Vo, R. Srivas, and B.O. Palsson, *Global reconstruction of the human metabolic network based on genomic and bibliomic data.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(6): p. 1777-1782.
266. Shlomi, T., M.N. Cabili, M.J. Herrgard, B.O. Palsson, and E. Ruppin, *Network-based prediction of human tissue-specific metabolism.* Nature Biotechnology, 2008. **26**(9): p. 1003-1010.
267. Jerby, L., T. Shlomi, and E. Ruppin, *Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism.* Molecular Systems Biology, 2010. **6**: p. -.
268. Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, *Basic Local Alignment Search Tool.* Journal of Molecular Biology, 1990. **215**(3): p. 403-410.
269. Varma, A. and B.O. Palsson, *Metabolic Flux Balancing - Basic Concepts, Scientific and Practical Use.* Bio-Technology, 1994. **12**(10): p. 994-998.
270. Satish Kumar, V., J.G. Ferry, and C.D. Maranas, *Metabolic reconstruction of the archaeon methanogen Methanosarcina Acetivorans.* BMC Syst Biol, 2011. **5**: p. 28.
271. Park, J.M., T.Y. Kim, and S.Y. Lee, *Constraints-based genome-scale metabolic simulation for systems metabolic engineering.* Biotechnol Adv, 2009. **27**(6): p. 979-88.
272. Feist, A.M. and B.O. Palsson, *The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli.* Nat Biotechnol, 2008. **26**(6): p. 659-67.
273. Pharkya, P., A.P. Burgard, and C.D. Maranas, *Exploring the overproduction of amino acids using the bilevel optimization framework OptKnock.* Biotechnol Bioeng, 2003. **84**(7): p. 887-99.
274. Hatzimanikatis, V., M. Emmerling, U. Sauer, and J.E. Bailey, *Application of mathematical tools for metabolic design of microbial ethanol production.* Biotechnol Bioeng, 1998. **58**(2-3): p. 154-61.
275. Lee, D.S., H. Burd, J. Liu, E. Almaas, O. Wiest, A.L. Barabasi, Z.N. Oltvai, and V. Kapatral, *Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple Staphylococcus aureus genomes identify novel antimicrobial drug targets.* J Bacteriol, 2009. **191**(12): p. 4015-24.
276. Pinney, J.W., B. Papp, C. Hyland, L. Wambua, D.R. Westhead, and G.A. McConkey, *Metabolic reconstruction and analysis for parasite genomes.* Trends Parasitol, 2007. **23**(11): p. 548-54.
277. Shlomi, T., M.N. Cabili, and E. Ruppin, *Predicting metabolic biomarkers of human inborn errors of metabolism.* Mol Syst Biol, 2009. **5**: p. 263.
278. Maeder, D.L., I. Anderson, T.S. Brettin, D.C. Bruce, P. Gilna, C.S. Han, A. Lapidus, W.W. Metcalf, E. Saunders, R. Tapia, and K.R. Sowers, *The Methanosarcina barkeri genome: comparative analysis with Methanosarcina acetivorans and Methanosarcina*

mazei reveals extensive rearrangement within methanosarcinal genomes. J Bacteriol, 2006. **188**(22): p. 7922-31.
279. Deppenmeier, U., A. Johann, T. Hartsch, R. Merkl, R.A. Schmitz, R. Martinez-Arias, A. Henne, A. Wiezer, S. Baumer, C. Jacobi, H. Bruggemann, T. Lienard, A. Christmann, M. Bomeke, S. Steckel, A. Bhattacharyya, A. Lykidis, R. Overbeek, H.P. Klenk, R.P. Gunsalus, H.J. Fritz, and G. Gottschalk, *The genome of Methanosarcina mazei: evidence for lateral gene transfer between bacteria and archaea.* J Mol Microbiol Biotechnol, 2002. **4**(4): p. 453-61.
280. Galagan, J.E., *The Genome of M. acetivorans Reveals Extensive Metabolic and Physiological Diversity.* Genome Research, 2002. **12**(4): p. 532-542.
281. Rother, M., E. Oelgeschläger, and W. W. Metcalf, *Genetic and proteomic analyses of CO utilization by Methanosarcina acetivorans.* Archives of Microbiology, 2007.
282. SabrinaTachdjian, K. and S. Conners, *Functional Genomics of Stress Response in Extremophilic Archaea.* Archaea: New Models for Prokaryotic Biology, 2008.
283. Li, L., Q. Li, L. Rohlin, U. Kim, K. Salmon, T. Rejtar, R.P. Gunsalus, B.L. Karger, and J.G. Ferry, *Quantitative Proteomic and Microarray Analysis of the Archaeon MethanosarcinaacetivoransGrown with Acetate versus Methanol.* Journal of Proteome Research, 2007. **6**(2): p. 759-771.
284. Li, Q., L. Li, T. Rejtar, B. Karger, and J. Ferry, *Proteome of Methanosarcinaacetivorans Part II: Comparison of Protein Levels in Acetate- ….* Journal of Proteome Research, 2005.
285. Li, Q., L. Li, T. Rejtar, B. Karger, and J. Ferry, *Proteome of Methanosarcinaacetivorans Part I: An Expanded View of the Biology of the Cell.* Journal of Proteome Research, 2005.
286. Schlesinger, W.H., *Biogeochemistry : an analysis of global change*. 2nd ed. 1997, San Diego, Calif.: Academic Press. xiii, 588 p., [2] p. of plates.
287. Ferry, J.G., *Methanogenesis : ecology, physiology, biochemistry & genetics*. Chapman & Hall microbiology series. 1993, New York: Chapman & Hall. x, 536 p.
288. McInerney, M.J., J.R. Sieber, and R.P. Gunsalus, *Syntrophy in anaerobic global carbon cycles.* Curr Opin Biotechnol, 2009. **20**(6): p. 623-32.
289. Bloom, A.A., P.I. Palmer, A. Fraser, D.S. Reay, and C. Frankenberg, *Large-scale controls of methanogenesis inferred from methane and gravity spaceborne data.* Science. **327**(5963): p. 322-5.
290. Cheng, S., D. Xing, D.F. Call, and B.E. Logan, *Direct biological conversion of electrical current into methane by electromethanogenesis.* Environ Sci Technol, 2009. **43**(10): p. 3953-8.
291. Ferry, J. and C. House, *The stepwise evolution of early life driven by energy conservation.* Molecular biology and evolution, 2006.
292. Battistuzzi, F.U., A. Feijao, and S.B. Hedges, *A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land.* BMC Evol Biol, 2004. **4**: p. 44.
293. Ferry JG, K.K., ed. *Methanogenesis*. Archaea: Molecular Cell Biology, ed. R Cabicchioli. 2007, ASM Press. 7.
294. Sowers, K., S. Baron, and J. Ferry, *Methanosarcina acetivorans sp. nov., an acetotrophic methane-producing bacterium ….* Applied and Environmental Microbiology, 1984.
295. Ferry, J.G. and D.J. Lessner, *Methanogenesis in Marine Sediments.* Annals of the New York Academy of Sciences, 2007. **1125**(1): p. 147-157.
296. Lessner, D., L. Li, Q. Li, T. Rejtar, and V. Andreev, *… of CO2 to methane in CO-grown Methanosarcina acetivorans revealed by proteomics.* Proceedings of the National Academy of Sciences, 2006.
297. *http://cmr.jcvi.org/cgi-bin/CMR/CmrHomePage.cgi*.

298. Ding, Y., S. Zhang, J. Tomb, and J. Ferry, *... system that are differentially expressed in methanol-and acetate-grown Methanosarcina ....* FEMS Microbiology Letters, 2002.
299. Sowers, K. *http://carb.umbi.umd.edu/g2f/*.
300. Overbeek, R., T. Begley, R.M. Butler, J.V. Choudhuri, H.Y. Chuang, M. Cohoon, V. de Crecy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E.D. Frank, S. Gerdes, E.M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A.C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G.D. Pusch, D.A. Rodionov, C. Ruckert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, and V. Vonstein, *The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.* Nucleic Acids Res, 2005. **33**(17): p. 5691-702.
301. Bose, A., M.A. Pritchett, and W.W. Metcalf, *Genetic Analysis of the Methanol- and Methylamine-Specific Methyltransferase 2 Genes of Methanosarcina acetivorans C2A.* Journal of Bacteriology, 2008. **190**(11): p. 4017-4026.
302. Joyce, A.R., J.L. Reed, A. White, R. Edwards, A. Osterman, T. Baba, H. Mori, S.A. Lesely, B.O. Palsson, and S. Agarwalla, *Experimental and computational assessment of conditionally essential genes in Escherichia coli.* J Bacteriol, 2006. **188**(23): p. 8259-71.
303. Rother, M., P. Boccazzi, A. Bose, and M. Pritchett s, *... methyl-coenzyme M reductase is essential in Methanosarcina acetivorans C2A and allows ....* Journal of Bacteriology, 2005.
304. Thauer, R., A. Kaster, H. Seedorf, and W. Buckel, *Methanogenic archaea: ecologically relevant differences in energy conservation.* Nature Reviews Microbiology, 2008.
305. Thiele, I. and B.O. Palsson, *A protocol for generating a high-quality genome-scale metabolic reconstruction.* Nat Protoc. **5**(1): p. 93-121.
306. Pernthaler, A., A.E. Dekas, C.T. Brown, S.K. Goffredi, T. Embaye, and V.J. Orphan, *Diverse syntrophic partnerships from deep-sea methane vents revealed by direct cell capture and metagenomics.* Proc Natl Acad Sci U S A, 2008. **105**(19): p. 7052-7.
307. Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
308. Bairoch, A., R. Apweiler, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, and L.S. Yeh, *The Universal Protein Resource (UniProt).* Nucleic Acids Res, 2005. **33**(Database issue): p. D154-9.
309. Ravikirthi, P., P.F. Suthers, and C.D. Maranas, *Construction of an E. Coli genome-scale atom mapping model for MFA calculations.* Biotechnol Bioeng, 2011. **108**(6): p. 1372-82.
310. Vallino, J.J. and G. Stephanopoulos, *Metabolic flux distributions in Corynebacterium glutamicum during growth and lysine overproduction.* Biotechnol Bioeng, 1993. **41**(6): p. 633-46.
311. Nielsen, J., *It is all about metabolic fluxes.* J Bacteriol, 2003. **185**(24): p. 7031-5.
312. Bailey, J.E., *Toward a science of metabolic engineering.* Science, 1991. **252**(5013): p. 1668-75.
313. Stephanopoulos, G. and J.J. Vallino, *Network rigidity and metabolic engineering in metabolite overproduction.* Science, 1991. **252**(5013): p. 1675-81.
314. Adelbert, B., R. Christoph, E. Dietmar, A. Duilio, F. Georg, and E. Wolfgang, *Elucidation of novel biosynthetic pathways and metabolite flux patterns by retrobiosynthetic NMR analysis.* FEMS Microbiology Reviews, 1998. **22**(5): p. 567-598.
315. Kelleher, J.K., *Flux estimation using isotopic tracers: common ground for metabolic physiology and metabolic engineering.* Metab Eng, 2001. **3**(2): p. 100-10.

316. Wittmann, C. and E. Heinzle, *Genealogy profiling through strain improvement by using metabolic network analysis: metabolic flux genealogy of several generations of lysine-producing corynebacteria.* Appl Environ Microbiol, 2002. **68**(12): p. 5843-59.
317. Schmidt, K., J. Nielsen, and J. Villadsen, *Quantitative analysis of metabolic fluxes in Escherichia coli, using two-dimensional NMR spectroscopy and complete isotopomer models.* J Biotechnol, 1999. **71**(1-3): p. 175-89.
318. Kim, H.U., T.Y. Kim, and S.Y. Lee, *Metabolic flux analysis and metabolic engineering of microorganisms.* Mol Biosyst, 2008. **4**(2): p. 113-20.
319. Zupke, C. and G. Stephanopoulos, *Modeling of Isotope Distributions and Intracellular Fluxes in Metabolic Networks Using Atom Mapping Matrixes.* Biotechnol. Prog., 1994. **10**(5): p. 489-498.
320. Schmidt, K., M. Carlsen, J. Nielsen, and J. Villadsen, *Modeling isotopomer distributions in biochemical networks using isotopomer mapping matrices.* Biotechnol Bioeng, 1997. **55**(6): p. 831-40.
321. Wiechert, W., M. Mollney, N. Isermann, M. Wurzel, and A.A. de Graaf, *Bidirectional reaction steps in metabolic networks: III. Explicit solution and analysis of isotopomer labeling systems.* Biotechnol Bioeng, 1999. **66**(2): p. 69-85.
322. van Winden, W.A., C. Wittmann, E. Heinzle, and J.J. Heijnen, *Correcting mass isotopomer distributions for naturally occurring isotopes.* Biotechnol Bioeng, 2002. **80**(4): p. 477-9.
323. Antoniewicz, M.R., J.K. Kelleher, and G. Stephanopoulos, *Elementary metabolite units (EMU): a novel framework for modeling isotopic distributions.* Metab Eng, 2007. **9**(1): p. 68-86.
324. Suthers, P.F., A.P. Burgard, M.S. Dasika, F. Nowroozi, S. van Dien, J.D. Keasling, and C.D. Maranas, *Metabolic flux elucidation for large-scale models using ^{13}C labeled isotopes.* Metab Eng, 2007. **9**(5-6): p. 387-405.
325. Antoniewicz, M.R., D.F. Kraynie, L.A. Laffend, J. Gonzalez-Lergier, J.K. Kelleher, and G. Stephanopoulos, *Metabolic flux analysis in a nonstationary system: fed-batch fermentation of a high yielding strain of E. coli producing 1,3-propanediol.* Metab Eng, 2007. **9**(3): p. 277-92.
326. Chang, Y., P.F. Suthers, and C.D. Maranas, *Identification of optimal measurement sets for complete flux elucidation in metabolic flux analysis experiments.* Biotechnol Bioeng, 2008. **100**(6): p. 1039-49.
327. Raymond, J.W. and P. Willett, *Maximum common subgraph isomorphism algorithms for the matching of chemical structures.* J Comput Aided Mol Des, 2002. **16**(7): p. 521-33.
328. Willett, P., *Searching for pharmacophoric patterns in databases of three-dimensional chemical structures.* J Mol Recognit, 1995. **8**(5): p. 290-303.
329. Gillet, V.J., D.J. Wild, P. Willett, and J. Bradshaw, *Similarity and Dissimilarity Methods for Processing Chemical Structure Databases.* The Computer Journal, 1998. **41**(8): p. 547-558.
330. Raymond, J.W., E.J. Gardiner, and P. Willett, *Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithm.* J Chem Inf Comput Sci, 2002. **42**(2): p. 305-16.
331. Raymond, J.W., E.J. Gardiner, and P. Willett, *RASCAL: Calculation of graph similarity using maximum common edge subgraphs.* Computer Journal, 2002. **45**(6): p. 631-644.
332. Hattori, M., Y. Okuno, S. Goto, and M. Kanehisa, *Heuristics for chemical compound matching.* Genome Inform, 2003. **14**: p. 144-53.
333. Hattori, M., Y. Okuno, S. Goto, and M. Kanehisa, *Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways.* J Am Chem Soc, 2003. **125**(39): p. 11853-65.

334. Mu, F., R.F. Williams, C.J. Unkefer, P.J. Unkefer, J.R. Faeder, and W.S. Hlavacek, *Carbon-fate maps for metabolic reactions.* Bioinformatics, 2007. **23**(23): p. 3193-3199.
335. Goto, S., T. Nishioka, and M. Kanehisa, *LIGAND: chemical database for enzyme reactions.* Bioinformatics, 1998. **14**(7): p. 591-9.
336. Goto, S., Y. Okuno, M. Hattori, T. Nishioka, and M. Kanehisa, *LIGAND: database of chemical compounds and reactions in biological pathways.* Nucleic Acids Res, 2002. **30**(1): p. 402-4.
337. Wipke, W.T. and T.M. Dyott, *Stereochemically unique naming algorithm.* J. Am. Chem. Soc., 1974. **96**(15): p. 4834-4842.
338. Arita, M., *In silico atomic tracing by substrate-product relationships in Escherichia coli intermediary metabolism.* Genome Res, 2003. **13**(11): p. 2455-66.
339. Flower, D.R., *On the Properties of Bit String-Based Measures of Chemical Similarity.* J. Chem. Inf. Comput. Sci., 1998. **38**(3): p. 379-386.
340. Arita, M., *The metabolic world of Escherichia coli is not small.* Proc Natl Acad Sci U S A, 2004. **101**(6): p. 1543-7.
341. Coen, B. and K. Joep, *Algorithm 457: finding all cliques of an undirected graph.* Commun. ACM, 1973. **16**(9): p. 575-577.
342. Linus, P. and J. Sherman, *The Nature of the Chemical Bond. VI. The Calculation from Thermochemical Data of the Energy of Resonance of Molecules Among Several Electronic Structures.* The Journal of Chemical Physics, 1933. **1**(8): p. 606-617.
343. ILOG, *ILOG CPLEX 10.1 User's Manual.* Mountain View, CA: ILOG Inc., 2006.
344. Dauner, M., J.E. Bailey, and U. Sauer, *Metabolic flux analysis with a comprehensive isotopomer model in Bacillus subtilis.* Biotechnol Bioeng, 2001. **76**(2): p. 144-56.
345. Gonzalez-Lergier, J., L.J. Broadbelt, and V. Hatzimanikatis, *Theoretical considerations and computational analysis of the complexity in polyketide synthesis pathways.* J Am Chem Soc, 2005. **127**(27): p. 9930-8.
346. Tipton, K. and S. Boyce, *History of the enzyme nomenclature system.* Bioinformatics, 2000. **16**(1): p. 34-40.
347. Iwatani, S., Y. Yamada, and Y. Usuda, *Metabolic flux analysis in biotechnology processes.* Biotechnol Lett, 2008. **30**(5): p. 791-799.
348. Wiechert, W., O. Schweissgut, H. Takanaga, and W.B. Frommer, *Fluxomics: mass spectrometry versus quantitative imaging.* Curr Opin Plant Biol, 2007. **10**(3): p. 323-330.
349. Shastri, A.A. and J.A. Morgan, *A transient isotopic labeling methodology for ^{13}C metabolic flux analysis of photoautotrophic microorganisms.* Phytochemistry, 2007. **68**(16-18): p. 2302-2312.
350. de Koning, W. and K. van Dam, *A method for the determination of changes of glycolytic metabolites in yeast on a subsecond time scale using extraction at neutral pH.* Anal Biochem, 1992. **204**(1): p. 118-123.
351. van Winden, W.A., J.C. van Dam, C. Ras, R.J. Kleijn, J.L. Vinke, W.M. van Gulik, and J.J. Heijnen, *Metabolic-flux analysis of Saccharomyces cerevisiae CEN.PK113-7D based on mass isotopomer measurements of ^{13}C-labeled primary metabolites.* FEMS Yeast Res, 2005. **5**(6-7): p. 559-568.
352. Mashego, M.R., L. Wu, J.C. Van Dam, C. Ras, J.L. Vinke, W.A. Van Winden, W.M. Van Gulik, and J.J. Heijnen, *MIRACLE: mass isotopomer ratio analysis of U-^{13}C-labeled extracts. A new method for accurate quantification of changes in concentrations of intracellular metabolites.* Biotechnol Bioeng, 2004. **85**(6): p. 620-628.
353. Theobald, U., W. Mailinger, M. Reuss, and M. Rizzi, *In vivo analysis of glucose-induced fast changes in yeast adenine nucleotide pool applying a rapid sampling technique.* Anal Biochem, 1993. **214**(1): p. 31-37.

354. Nöh, K., K. Grönke, B. Luo, R. Takors, M. Oldiges, and W. Wiechert, *Metabolic flux analysis at ultra short time scale: Isotopically mon-stationary ^{13}C labeling experiments*. J Biotechnol, 2007. **129**(2): p. 249-267.
355. Young, J.D., J.L. Walther, M.R. Antoniewicz, H. Yoo, and G. Stephanopoulos, *An elementary metabolite unit (EMU) based method of isotopically nonstationary flux analysis.* Biotechnol Bioeng, 2007. **99**(3): p. 686-699.
356. Nöh, K., A. Wahl, and W. Wiechert, *Computational tools for isotopically instationary ^{13}C labeling experiments under metabolic steady state conditions.* Metab Eng, 2006. **8**(6): p. 554-577.
357. Schaub, J., K. Mauch, and M. Reuss, *Metabolic flux analysis in Escherichia coli by integrating isotopic dynamic and isotopic stationary ^{13}C labeling data.* Biotechnol Bioeng, 2008. **99**(5): p. 1170-1185.
358. Maier, K., U. Hofmann, M. Reuss, and K. Mauch, *Identification of metabolic fluxes in hepatic cells from transient ^{13}C-labeling experiments: Part II. Flux estimation.* Biotechnol Bioeng, 2008. **100**(2): p. 355-370.
359. Robert, M., T. Soga, and M. Tomita, *E. coli metabolomics: Capturing the complexity of a "simple" model*, in *Metabolomics*, J. Nielsen and M.C. Jewett, Editors. 2007, Springer-Verlag: Berlin, Germany. p. 189-234.
360. Zamboni, N., *Toward metabolome-based ^{13}C flux analysis: A universal tool for measuring in vivo metabolic activity*, in *Metabolomics*, J. Nielsen and M.C. Jewett, Editors. 2007, Springer-Verlag: Berlin, Germany. p. 129-157.
361. Ishii, N., K. Nakahigashi, T. Baba, M. Robert, T. Soga, A. Kanai, T. Hirasawa, M. Naba, K. Hirai, A. Hoque, P.Y. Ho, Y. Kakazu, K. Sugawara, S. Igarashi, S. Harada, T. Masuda, N. Sugiyama, T. Togashi, M. Hasegawa, Y. Takai, K. Yugi, K. Arakawa, N. Iwata, Y. Toya, Y. Nakayama, T. Nishioka, K. Shimizu, H. Mori, and M. Tomita, *Multiple high-throughput analyses monitor the response of E. coli to perturbations.* Science, 2007. **316**(5824): p. 593-597.
362. Chang, Y., P.F. Suthers, and C.D. Maranas, *Idnentification of optimal measurement sets for complete flux elucidation in MFA experiments.* Biotechnol Bioeng, 2008. **100**(6): p. 1039-49.
363. Yang, T.H., O. Frick, and E. Heinzle, *Hybrid optimization for ^{13}C metabolic flux analysis using systems parametrized by compactification.* BMC Syst Biol, 2008. **2**: p. 29.
364. ILOG, *ILOG CPLEX 11.0 User's Manual*. 2007, Sunnyvale, CA: ILOG, Inc.
365. ILOG, *ILOG CPLEX C++ API 11.0 Reference Manual*. 2007, Gentilly, France: ILOG SA.
366. Wiechert, W., *Modeling and simulation: tools for metabolic engineering.* J Biotechnol, 2002. **94**(1): p. 37-63.
367. Nöh, K. and W. Wiechert, *Experimental design principles for isotopically instationary ^{13}C labeling experiments.* Biotechnol Bioeng, 2006. **94**(2): p. 234-251.
368. MathWorks, *MATLAB Technical Documentation*. 2007, Natick: The MathWorks Inc.
369. Tawarmalani, M. and N.V. Sahinidis, *Global optimization of mixed-integer nonlinear programs: A theoretical and computational study.* Math Program, Ser. A, 2004. **99**(3): p. 563-591.
370. Antoniewicz, M.R., D.F. Kraynie, L.A. Laffend, J. González-Lergier, J.K. Kelleher, and G. Stephanopoulos, *Metabolic flux analysis in a nonstationary system: Fed-batch fermentation of a high yielding strain of E. coli producing 1,3-propanediol.* Metab Eng, 2007. **9**(3): p. 277-292.
371. Holmberg, A., *On the practical identifiability of microbial growth models incorporating Michaelis-Menten type nonlinearities.* Math Biosci, 1982. **62**(1): p. 23-43.
372. Vanrolleghem, P.A., M. van Daele, and D. Dochain, *Practical identifiability of a biokinetic model of activated sludge respiration.* Water Res, 1995. **29**(11): p. 2561-2570.

373. Burgard, A.P., E.V. Nikolaev, C.H. Schilling, and C.D. Maranas, *Flux coupling analysis of genome-scale metabolic network reconstructions.* Genome Res, 2004. **14**(2): p. 301-312.
374. Nikolaev, E.V., A.P. Burgard, and C.D. Maranas, *Elucidation and structural analysis of conserved pools for genome-scale metabolic reconstructions.* Biophys J, 2005. **88**(1): p. 37-49.
375. Weitzel, M., W. Wiechert, and K. Nöh, *The topology of metabolic isotope labeling networks.* BMC Bioinformatics, 2007. **8**: p. 315.
376. Selivanov, V.A., J. Puigjaner, A. Sillero, J.J. Centelles, A. Ramos-Montoya, P.W. Lee, and M. Cascante, *An optimized algorithm for flux estimation from isotopomer distribution in glucose metabolites.* Bioinformatics, 2004. **20**(18): p. 3387-3397.
377. Wahl, S.A., K. Nöh, and W. Wiechert, *13C labeling experiments at metabolic nonstationary conditions: An exploratory study.* BMC Bioinformatics, 2008. **9**: p. 152.
378. Alvarez-Vasquez, F., Y.A. Hannun, and E.O. Voit, *Dynamics of positional enrichment: Theoretical development and application to carbon labeling in Zymomonas mobilis.* Biochem Eng J, 2008. **40**(1): p. 157-174.
379. Hoque, M.A., H. Ushiyama, M. Tomita, and K. Shimizu, *Dynamic responses of the intracellular metabolite concentrations of the wild type and pykA mutant Escherichia coli against pulse addition of glucose or NH_3 under those limiting continuous cultures.* Biochem Eng J, 2005. **26**(1): p. 38-49.
380. Wittmann, C., M. Hans, W.A. van Winden, C. Ras, and J.J. Heijnen, *Dynamics of intracellular metabolites of glycolysis and TCA cycle during cell-cycle-related oscillation in Saccharomyces cerevisiae.* Biotechnol Bioeng, 2005. **89**(7): p. 839-847.
381. Wittmann, C., M. Hans, and E. Heinzle, *In vivo analysis of intracellular amino acid labeling by GC/MS.* Anal Biochem, 2002. **307**(2): p. 379-382.
382. Atsumi, S. and J.C. Liao, *Metabolic engineering for advanced biofuels production from Escherichia coli.* Curr Opin Biotechnol, 2008. **19**(5): p. 414-9.
383. Janssen, P.H., *Propanol as an end product of threonine fermentation.* Arch Microbiol, 2004. **182**(6): p. 482-6.
384. Jones, D.T. and D.R. Woods, *Acetone-butanol fermentation revisited.* Microbiol Rev, 1986. **50**(4): p. 484-524.
385. Nair, R.V. and E.T. Papoutsakis, *Expression of plasmid-encoded aad in Clostridium acetobutylicum M5 restores vigorous butanol production.* J Bacteriol, 1994. **176**(18): p. 5843-6.
386. Formanek, J., R. Mackie, and H.P. Blaschek, *Enhanced Butanol Production by Clostridium beijerinckii BA101 Grown in Semidefined P2 Medium Containing 6 Percent Maltodextrin or Glucose.* Appl Environ Microbiol, 1997. **63**(6): p. 2306-2310.
387. Lin, Y.L. and H.P. Blaschek, *Butanol Production by a Butanol-Tolerant Strain of Clostridium acetobutylicum in Extruded Corn Broth.* Appl Environ Microbiol, 1983. **45**(3): p. 966-973.
388. Hanai, T., S. Atsumi, and J.C. Liao, *Engineered synthetic pathway for isopropanol production in Escherichia coli.* Appl Environ Microbiol, 2007. **73**(24): p. 7814-8.
389. Yan, Y. and J.C. Liao, *Engineering metabolic systems for production of advanced fuels.* J Ind Microbiol Biotechnol, 2009. **36**(4): p. 471-9.
390. Boynton, Z.L., G.N. Bennett, and F.B. Rudolph, *Cloning, sequencing, and expression of genes encoding phosphotransacetylase and acetate kinase from Clostridium acetobutylicum ATCC 824.* Appl Environ Microbiol, 1996. **62**(8): p. 2758-66.
391. Fontaine, L., I. Meynial-Salles, L. Girbal, X. Yang, C. Croux, and P. Soucaille, *Molecular characterization and transcriptional analysis of adhE2, the gene encoding the NADH-dependent aldehyde/alcohol dehydrogenase responsible for butanol production in*

alcohologenic cultures of Clostridium acetobutylicum ATCC 824. J Bacteriol, 2002. **184**(3): p. 821-30.
392. Wiesenborn, D.P., F.B. Rudolph, and E.T. Papoutsakis, *Thiolase from Clostridium acetobutylicum ATCC 824 and Its Role in the Synthesis of Acids and Solvents.* Appl Environ Microbiol, 1988. **54**(11): p. 2717-2722.
393. Becker, D.F., J.A. Fuchs, D.K. Banfield, W.D. Funk, R.T. MacGillivray, and M.T. Stankovich, *Characterization of wild-type and an active-site mutant in Escherichia coli of short-chain acyl-CoA dehydrogenase from Megasphaera elsdenii.* Biochemistry, 1993. **32**(40): p. 10736-42.
394. O'Neill, H., S.G. Mayhew, and G. Butler, *Cloning and analysis of the genes for a novel electron-transferring flavoprotein from Megasphaera elsdenii. Expression and characterization of the recombinant protein.* J Biol Chem, 1998. **273**(33): p. 21015-24.
395. Wallace, K.K., Z.Y. Bao, H. Dai, R. Digate, G. Schuler, M.K. Speedie, and K.A. Reynolds, *Purification of crotonyl-CoA reductase from Streptomyces collinus and cloning, sequencing and expression of the corresponding gene in Escherichia coli.* Eur J Biochem, 1995. **233**(3): p. 954-62.
396. Sentheshanuganathan, S., *The mechanism of the formation of higher alcohols from amino acids by Saccharomyces cerevisiae.* Biochem J, 1960. **74**: p. 568-76.
397. Atsumi, S., A.F. Cann, M.R. Connor, C.R. Shen, K.M. Smith, M.P. Brynildsen, K.J. Chou, T. Hanai, and J.C. Liao, *Metabolic engineering of Escherichia coli for 1-butanol production.* Metab Eng, 2008. **10**(6): p. 305-11.
398. Steen, E.J., R. Chan, N. Prasad, S. Myers, C.J. Petzold, A. Redding, M. Ouellet, and J.D. Keasling, *Metabolic engineering of Saccharomyces cerevisiae for the production of n-butanol.* Microb Cell Fact, 2008. **7**: p. 36.
399. Bader, G.D., M.P. Cary, and C. Sander, *Pathguide: a pathway resource list.* Nucleic Acids Res, 2006. **34**(Database issue): p. D504-6.
400. Wang, Y., J. Xiao, T.O. Suzek, J. Zhang, J. Wang, and S.H. Bryant, *PubChem: a public information system for analyzing bioactivities of small molecules.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W623-33.
401. Clomburg, J.M. and R. Gonzalez, *Biofuel production in Escherichia coli: the role of metabolic engineering and synthetic biology.* Appl Microbiol Biotechnol. **86**(2): p. 419-34.
402. Kagamiyama, H. and H. Hayashi, *Branched-chain amino-acid aminotransferase of Escherichia coli.* Methods Enzymol, 2000. **324**: p. 103-13.
403. Smit, B.A., J.E. van Hylckama Vlieg, W.J. Engels, L. Meijer, J.T. Wouters, and G. Smit, *Identification, cloning, and characterization of a Lactococcus lactis branched-chain alpha-keto acid decarboxylase involved in flavor formation.* Appl Environ Microbiol, 2005. **71**(1): p. 303-11.
404. Perpete, P. and S. Collin, *Contribution of 3-methylthiopropionaldehyde to the worty flavor of alcohol-free beers.* J Agric Food Chem, 1999. **47**(6): p. 2374-8.
405. Shimizu, K., *Metabolic flux analysis based on 13C-labeling experiments and integration of the information with gene and protein expression patterns.* Adv Biochem Eng Biotechnol, 2004. **91**: p. 1-49.
406. Nielsen, D.R., E. Leonard, S.H. Yoon, H.C. Tseng, C. Yuan, and K.L. Prather, *Engineering alternative butanol production platforms in heterologous bacteria.* Metab Eng, 2009. **11**(4-5): p. 262-73.
407. Kanehisa, M., *The KEGG database.* Novartis Found Symp, 2002. **247**: p. 91-101; discussion 101-3, 119-28, 244-52.
408. Im, W., M.S. Lee, and C.L. Brooks, 3rd, *Generalized born model with a simple smoothing function.* J Comput Chem, 2003. **24**(14): p. 1691-702.