Understanding a Methanogenic Benzene-Degrading Culture Using Metabolic Models Created from Metagenomic Sequences

by

Cleo Hanchen Ho

A thesis submitted in conformity with the requirements for the degree of Master of Science

Graduate Department of Chemical Engineering and Applied Chemistry

University of Toronto

Abstract

Understanding a Methanogenic Benzene-Degrading Culture Using Metabolic Models Created from Metagenomic Sequences

Master of Applied Science

2013

Cleo Hanchen Ho

Department of Chemical Engineering and Applied Chemistry

University of Toronto

Metabolic models were constructed from the metagenome of a methanogenic benzene-degrading community to understand the metabolite interactions among the key microbes in the culture. The metagenomic sequences were assembled, and it was found that assembling the short DNA fragments before they were combined with longer reads can contribute to the overall lengths of the resulting sequences. The metagenome was then taxonomically classified into the domain of archaea and bacteria, and domain-specific models were built. A mathematical framework to fill metabolic gaps at the community level was then developed and applied to the benzene-degrading community model to study how metabolic gaps can be filled by via interspecies metabolite transfer, and it suggested that among other metabolites, acetate, hydrogen, formate, coenzyme A and histidine produced by the bacteria population could potentially contribute to the growth of the methanogens. The computational framework demonstrated its ability to generate testable hypotheses about microbial interactions.

Acknowledgment

I would like to thank my supervisor, Krishna Mahadevan, for his support and enthusiasm during my time here. I am also grateful to Professor Elizabeth Edwards for her guidance throughout my study. I am indebted to all my friends and colleagues in Biozone, particularly Fei Luo, Cheryl Devine, Chris Gowen, Laura Hug, Laurence Yang, Weijun Gao, Eugene Ma, Fahimeh Salimi, and Kevin Correia; this work would not have been possible without their help. I especially wish to acknowledge the inhabitants of Wallberg Room 320, for they are truly inspirational.

The funding agencies that supported this work were Ontario Graduate Scholarships, NSERC, Genome Canada, Ontario Genomics Institute, and Government of Ontario.

Lastly, I would like to thank my parents, Szu-yin and Chia-chu, for their love and support during my graduate studies. This thesis is dedicated to my family.

Table of Contents

List of Tables

List of Figures

1.0 Introduction and Literature Review

1.1 Anaerobic Benzene Degradation in Mixed Cultures

Benzene is a toxic compound naturally occurring in crude oil. It can thus disperse in the environment via anthropogenic activities such as industrial or transport discharge, storage leaks, and fuel combustion (Health Canada, 2009). As a human carcinogen, benzene poses threats to the public health when it contaminates drinking water, soil, and air. In addition, benzene is recalcitrant in the environment where oxygen is scarce due to its thermodynamic stability conferred by the aromatic structure. In soils and sediments, natural attenuation of benzene relies on the metabolism of indigenous microorganisms, which utilize benzene as their carbon and energy source and produce compounds that are less harmful, such as carbon dioxide and methane. However, the degradation process has been shown to be slow under anoxic conditions, often taking days to months to achieve a complete reduction in benzene concentration (Ulrich and Edwards, 2003; Kunapuli et al., 2008; Sakai et al., 2009). Without the reactive oxygen, benzene activation is a thermodynamically unfavourable process, and because oxygen is often the preferred electron acceptor and can be consumed rapidly by many microbes, essentially all benzene-polluted areas are devoid of oxygen. Nevertheless, if the metabolic capabilities of these microorganisms can be harnessed, accelerated, and applied to the remediation of the pollutant, benzene contamination can be controlled and abated.

Anaerobic benzene degradation has been shown to occur under various electron-accepting conditions: denitrifying (Ulrich and Edwards, 2003; van der Zaan et al., 2012), sulfate-reducing (Lovley et al., 1995; Phelps et al., 1998; Caldwell and Suflita, 2000; Musat and Widdel, 2008;

Kleinsteuber et al., 2008), iron-reducing (Rooney-Varga et al., 1999; Kunapuli et al., 2007), and methanogenic (Ulrich and Edwards, 2003; Sakai et al., 2009). The mechanism through which benzene is activated in the absence of oxygen is still unclear, and studies have focused on elucidating the metabolic pathways of anaerobic benzene degradation as well as identifying the enzymes and microorganisms involving in this process. It is hypothesized that benzene can be activated via carboxylation, methylation, and hydroxylation, yielding benzoate, toluene, and phenol (Carmona et al., 2009), respectively, and the activation mechanism may vary across different cultures. Regardless of how the benzene ring is attacked, these compounds all channel through benzoyl-CoA, a key metabolite in the degradation of aromatic compounds, into the central metabolism (Carmona et al., 2009). So far, four pure cultures, *Azoarcus* and *Dechloromonas sp.*, have been reported for nitrate-reducing benzene degradation (Kasai et al., 2006; Coates et al., 2001), and *Ferroglobus placidus* (Holmes et al., 2011), *Geobacter metallireducens*, and *Geobacter* strain Ben (Zhang et al., 2012) have been associated with anaerobic benzene degradation under iron-reducing conditions. However, the isolation of key organisms is still difficult because in some enrichment cultures, benzene decomposition is enabled only by the concerted effort of the diverse populations in the consortia.

In energy-limiting yet phylogenetically versatile communities, such as those listed in Table 1, microorganisms develop syntrophic relationships to catabolize benzene. It is believed that the benzene degraders break down benzene into intermediates such as acetate and hydrogen, which are then consumed by other species in the community for growth or energy. The downstream organisms utilize the intermediates and thus keep the concentration of the fermentation products

Table 1. Benzene-Degrading Cultures and Their Community Structures

Benzene-Degrading Consortium	Community Structure		Suggested Role(s) of Other OTUs	Reference
	Putative Benzene Degrader	Other OTUs		
Sulfate-reducing lava granules in groundwater	Cryptanaerobacter Pelotomaculum	Desulfobacca	Acetate utilizer	Kleinsteuber et al., 2008
		Syntrophus	Aromatics utilizer Acetate utilizer	
		Magnetobacterium	Unknown	
Sulfate-reducing culture	Desulfobacraceae SB-21	Deltaproteobacteria	Sulfate reducer	Phelps et al., 1998; Oka et al., 2008
		Gammaproteobacteria	Unknown	
		Epsilonproteobacteria	H_2 scavenger	
		Cytophagles	Unknown	
Sulfate-reducing culture	Pelotomaculum-like Gram-positive bacteria (dominant)	Clone BpC43 Clone BpC52 (Clostridiaceae)	Unknown	Abu Laban et al., 2009
		Clone BpC108 (Syntrophaceae)	Unknown	
Nitrate-reducing chemostat	Peptococcaceae	Rhodocyclaceae	Nitrate-reducing secondary digester	van der Zaan et al., 2011
		Burkolderiaceae	Benzene utilizer?	
		Cholorobi	Ammonium oxidizer	
Iron-reducing culture	Peptococcaceae BF1	Uncultured Desulfobulbaceae	Iron reducer using H_2	Kunapuli et al., 2007
Methanogenic culture	OTU OR-M2?	OTU OR-M1	Syntrophic bacteria using intermediates	Ulrich and Edwards, 2003
		OTU OR-M7 OTU OR-M8 OUT OR-M9	Hydrogen-utilizing methanogens	
		OTU OR-M6	Acetoclastic methanogens	
Methanogenic culture	Unknown	Aquificae	Benzene degraders?	Chang et al., 2008
		Firmucutes		
		Bacteroidetes	Unknown	
		Thermotogae	Unknown	
		Methanosarcinales	Acetate utilizer	
		Methanomicrobiales	H_2 utilizer	
Methanogenic culture	Deltaproteobacterium Hasda-A	Firmicutes	Fermenters Acetogens	Sakai et al., 2009
		Methanosarcinales	Acetate utilizer	
		Methanomicrobiales	H_2 utilizer	

low enough for the benzene degraders to carry out the decomposition continually. The highly complex interactions among the microorganisms make it difficult to obtain a pure culture with the desired metabolic activities. In this case, metagenome sequencing has been employed as an alternative method to evaluate the metabolic potentials of the uncultured organisms and to understand the communities as a whole. Metagenomics is the study that allows the direct examination of the uncultured in their natural environment via DNA sequences, and with sequencing technologies becoming more accessible and developed, it is used increasingly and extensively. Likewise, the work presented in this thesis is founded upon the metagenome of a methanogenic benzene-degrading community.

1.2 Anaerobic Degradation of Benzene by the Methanogenic Enrichment Culture OR

The benzene-degrading culture studied here originated from sediment samples reported by Nales et al. (1998). In their work, groundwater and soil were collected from six petroleum-contaminated sites. Microcosms were then made for each site, and every microcosm was supplied with a source of electron acceptor: sulfate, nitrate, or iron (III). The microcosms were amended with either benzene or ^{14}C-labelled benzene. Continuous monitoring of the concentration of benzene and electron acceptors showed that benzene degradation might be coupled to the reduction of sulfate, nitrate, and iron (III) in certain microcosms. However, no significant degradation was observed for the methanogenic microcosms. For the microcosms where benzene degradation was significant, most ^{14}C-benzene ended up as ^{14}C-CO$_2$. It is noteworthy that a long lag time before the first sign of benzene degradation was observed for all benzene-consuming microcosms; the minimum lag time ranged from 8 to 300 days and was site-dependent. Also, the maximum rate of sustained degradation ranged from 1.1 to 56 µM/day, depending on the microcosm treatment and the sample origin.

In the study by Ulrich and Edwards (2003), enrichment cultures were prepared by successively transferring the content of the microcosms used by Nales et al. into a defined medium. Again, benzene was provided as the only carbon and energy source. It was found that some cultures were able to switch their electron acceptors from sulfate to carbon dioxide and thus produced methane, although in the original microcosms no methanogenesis was observed along with significant benzene degradation. The minimum doubling time for the methanogenic culture was estimated to be 30 days (Ulrich 2004). The phylogenetic composition of the methanogenic cultures was determined by a 16S rRNA clone library; the cultures were composed of both bacteria and archaea. Particularly, the cultures were dominated by two bacterial OTUs (Operational Taxonomic Units): OR-M1 and OR-M2. A phylogenetic characterization classified OR-M1 with the *Desulfosporosinus sp*., which are sulfate-reducing bacteria using donors such as pyruvate, ethanol, and lactate, while OR-M2 was clustered with *Desulfobacterium anilini*, a microbe able to degrade phenylamine and phenol. As a result, the authors postulated that the initial attack of benzene ring could be performed by OR-M2, and syntrophic interactions might be present between OR-M2 and other species, which could metabolize the products secreted by OR-M2 or remove other substrates that inhibit benzene degradation.

The metagenome reads used in this work were sequenced using Sanger and Roche 454 platforms by the Joint Genome Institute (JGI), and a paired-end assembly was generated. Based on the number of raw reads mapped to 16S rRNA gene sequences, the organism abundance of the metagenome was determined to be 57% of OR-M2 like *δ-Proteobacteria*, related to *Syntrophobacterales*, and 9.5% of *Methanosaeta concilii sp.*, which is an acetoclastic methanogen (Devine, 2013). In addition, 2% of the raw sequences was classified as *Methanomicrobiales spp.*, which were believed to be hydrogenotrophic, while other clones distantly related to *Thermodesulfatator sp.* and *Geobacteraceae sp*. were identified at 3.8% and

3.4%, respectively. Furthermore, with proteomic approaches, Devine et al. (2011) identified the expressed genes in the anaerobic benzoyl-CoA catabolic pathway during active benzene decomposition. Recently, the metagenome of the methanogenic benzene-degrading culture was sequenced again using Illumina sequencing technology, and the metagenome was estimated to be 84% of OR-M2-like *δ-Proteobacteria* and 9% of *Methanoregula sp*. (Devine 2013). It was therefore established that the consortium was dominated by a group of *δ-Proteobacteria* and hydrogenotrophic as well as acetoclastic methanogens.

In this thesis, only the raw reads sequenced earlier using Sanger and 454 methods were used. These sequences passed through the computational pipeline shown in Figure 1 to create genome-scale models. In a nutshell, the DNA reads were assembled, taxonomically classified, annotated, and transformed into mathematical models. These models were then interrogated for the abilities to generate testable hypotheses.

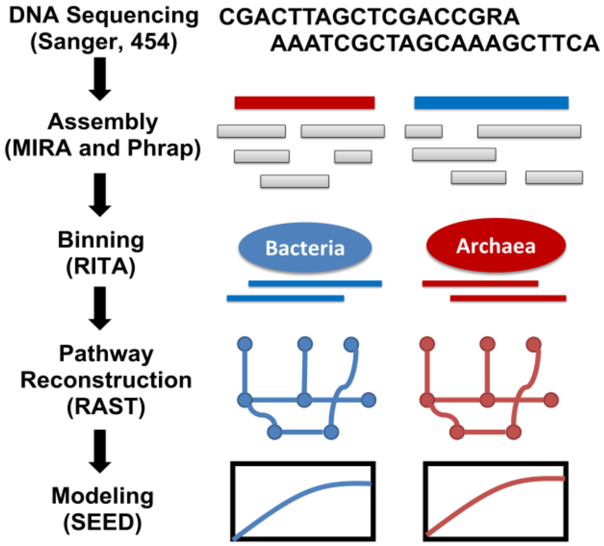

Figure 1. Network Reconstruction of the Metagenome of a Methanogenic Benzene-Degrading Culture

1.3 Metagenome Assembly

Assembly is a computational process that aligns and joins short sequence reads to generate longer sequences such that they can be used to infer function or phylogeny. As a whole, these longer sequences, or contigs, should represent the genome of interest. The reads are generated by platforms that determine the nucleotides sequences in given DNA or RNA samples. Due to the chemistry adopted in each platform, the reads generated by different technologies may vary greatly in length. For example, Sanger sequencing produces reads of 1000 - 1200 bp (Zhang et al., 2011), while next-generation technologies may give 300-500 bp in the case of Roche 454 or around 100 bp for Illumina sequencing (Kunin et al., 2008). Essentially, by aligning reads and detecting sequence overlaps among them, an assembler joins the fragments greedily until no similar sequences are left in the reads pool; alternatively, it can use a graph approach that aims to find the fewest paths by branching from one read to another via their overlaps (Pop, 2009).

Several metagenomes have been assembled, with environments ranging from mine biofilms, corals, to the human gut (Tyson et al., 2004; Meyer et al., 2009, Qin et al., 2010). Because the capability of an assembler relies partially on how it can manage and traverse through the overlaps in a pool of reads, metagenome data can complicate the process by presenting the assembler with reads from multiple organisms at often substantially different abundance. Due to the sequence similarities exhibited by the organisms, the program may mis-join the reads originating from different regions within the metagenome (Kunin et al., 2008). As pointed out by Tyson et al. (2004), if the environment of interest has fewer organisms, it is possible to reconstruct the genomes for each community member. It is therefore important to keep the heterogeneity of metagenomic data in mind when interpreting the assembly results, as it is the first step in generating metabolic models of the benzene-degrading community.

1.4 Sequence Annotation and Network Reconstruction

Genome-scale models are built based upon metabolic networks derived from genomic sequences. The process of model building, known as network reconstruction, is a bottom-up approach that integrates information of biochemical activities, substrate specificities, reaction stoichiometry and reversibility, and genome sequences (Durot et al. 2009). Typically, reactions from well-established pathways in the organisms of interest are included in the model, and they often serve as the foundation of a draft model. For these known metabolic pathways, the usual data sources are literature and public collections of enzymes and reactions; the latter includes databases like GenBank (Benson et al., 2013), IMG (Markowitz et al., 2012), and KEGG (Ogata et al., 1999), to name a few. Furthermore, phenotypic data, containing growth and flux measurements, can add to biochemical activities both qualitatively and quantitatively. With enzymatic assays and databases such as BRENDA (Schomburg et al., 2000), MetaCyc (Caspi et al., 2008), and TransportDB (Ren et al., 2004), substrate specificities can be determined for the enzymes in the model. For example, enzymes which act on a wide spectrum of substrate may be linked to multiple reactions; in addition, some enzymes, though capable of catalyzing similar reactions and accepting the same primary substrates, require different coenzymes to catalyze similar reactions, and it is essential to distinguish them as such differences are often organism specific (Reed et al. 2006; Feist et al. 2009).

For organisms that are less understood, however, the information derived from literature and databases could be limited, so the reconstruction of metabolic models must rely heavily on annotation, the interpretation of sequences, and the comparison with other genomes. The recent development in automated computational tools for inferring genes and gene products from DNA sequences has allowed the rapid creation of draft metabolic network (Meyer et al., 2008; Aziz et al., 2008). Briefly, open reading frames are identified by gene-calling algorithms, and gene products are assigned by comparing the target sequence to previously annotated proteins based on homology. Finally, cellular reactions are inferred from the annotated genomes, and a growth reaction is assigned from a biomass composition database.

1.5 Genome-Scale Model and Its Curation

Modeling has been applied to a wide range of studies ever since it emerged as a way to describe and predict biological phenomena mathematically. From representing enzyme kinetics to simulating bacteria batch growth, models have shown to be a promising tool for generating hypotheses from what is known, and when stoichiometric models were introduced, they demonstrated potentials in explaining pathway utilization and microbial physiology (Varma and Palsson, 1994). The models used in earlier studies, however, were limited to the *a priori* knowledge of metabolism, and it was not until later when genome sequencing and annotation became accessible that genome-scale models were constructed and used to a greater extent (Price et al., 2004).

A genome-scale model is, in essence, a curated network of metabolic reactions (Price et al., 2004). Ideally, such a model reflects the metabolism of an organism and represents all the biochemical conversions that could happen in the cells given a defined environment. Aside from cellular reactions, a genome-scale model also contains a growth reaction, in which biomass is synthesized from various biomass precursors, and exchange reactions, each of which is a route for a compound to travel between the extracellular and cellular compartments. In mathematical terms, a model takes the form of a matrix, where each column is a reaction and each row a metabolite, with the matrix components being the stoichiometric coefficients (Figure 2).

$$\mathrm{S} = \left[\begin{array}{cccccc} & \mathrm{rxn\ 1} & \mathrm{rxn\ 2} & \cdots & \mathrm{rxn\ (N-1)} & \mathrm{rxn\ N} \\ \mathrm{met\ 1} & -3 & 0 & \cdots & 0 & 0 \\ \mathrm{met\ 2} & -1 & 2 & \cdots & 1 & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathrm{met\ (M-1)} & 0 & 0 & \cdots & 0 & 0 \\ \mathrm{met\ M} & 1 & 0 & \cdots & 0 & 0 \end{array}\right]$$

Figure 2. An Example of Stoichiometric Matrix of with N Reactions and M Metabolites

As the output of network reconstruction, a draft model may sustain *in silico* growth in the complete medium, which often contains many amino acids as extra carbon source. Nevertheless, draft models often contain metabolic gaps, which may result from true metabolic incapacities or the loss of genetic information through sampling, sequencing events, and annotation. These gaps may hinder the synthesis of essential metabolites in the model, which inevitably leads to no cell growth in the minimal medium.

To date, there are two programs used to settle such metabolic gaps in genome-scale models by adding reactions from an external database: GapFind/GapFill (Kumar et al., 2007) and GrowMatch (Kumar and Maranas, 2009). The GapFind/GapFill algorithm is a two-stage optimization method that restores connectivity in a metabolic network to resolve no-production metabolites, or compounds that cannot be produced or transported into the model given the current set of metabolic reactions. By requiring each metabolite to be produced at a minimum level via at least one reaction, the GapFind formulation can identify those metabolites that do not meet these constraints; these metabolites are the no-production metabolites. Subsequently, to restore the connectivity of a no-production metabolite i*, GapFill (Figure 3) fills the gap by adding reactions from a database containing the reversible counterparts of the model reactions and other MetaCyc reactions that are foreign to the model. Whether a non-native reaction is incorporated into the model is determined by the binary variable γ_j, which indicates the addition of reaction j when equal to 1 and the exclusion when it is 0. Finally, the GapFill algorithm attempts to make the least changes to the model by minimizing the number of added reactions, or the sum of γ_j. The authors showed that GapFind/GapFill could resolve, to a certain extent, the disconnected metabolites in the *E. coli* model iJR904 and the *S. cerevisiae* model iND750, and it achieved so by incorporating external reactions, reversing the direction of existing reactions, and adding transport terms for the no-production metabolites. However, the framework does not guarantee *in silico* growth as it only links metabolites back to the network, so its application could be limited only to well-curated models.

$Minimize \quad \sum_{j \in Database} \gamma_j$	**(1)** Minimizing the number of reactions added from the database
such that	
$S_{i^*j} \cdot v_j \geq \delta - M\left(1 - \omega_{i^*j}\right) \quad \forall j \mid S_{ij} \neq 0$ $S_{i^*j} \cdot v_j \leq M\omega_{i^*j} \quad \forall j \mid S_{ij} \neq 0$	**(2)** When the reaction j producing metabolite i* is active (ω_{i^*j}=1), at least δ units and no more than M units of metabolite i* must be produced
$\sum_j \omega_{i^*j} \geq 1$	**(3)** For each no-production metabolite i*, there must be at least one reaction producing it
$\sum_j S_{ij} \cdot v_j \geq 0 \quad \forall i \in N$	**(4)** The mass balance for all metabolites assumes either the accumulation in cytosol (Sv > 0) or steady state (Sv = 0)
$LB_j \leq v_j \leq UB_j \quad \forall j \in Model$ $LB_j \cdot \gamma_j \leq v_j \leq UB_j \cdot \gamma_j \quad \forall j \in Database$	**(5)** Each reaction flux is constrained between its lower and upper bounds
$\omega_{ij} \in \{0, 1\} \quad \forall i, j$	**(6)** If ω_{ij}=1, reaction j producing metabolite i is active and 0 otherwise
$\gamma_j \in \{0, 1\} \quad \forall j \in Database$	**(7)** If γ_j=1, the Database reaction j is added to the model and 0 otherwis

Figure 3. Formulation of GapFill as Described by Kumar et al. (2007)

On the other hand, the gap-filling in GrowMatch, also known as NGG GrowMatch, is tied to biomass production by correcting the discrepancy between a **N**o-**G**rowth model and the observed *in vivo* **G**rowth given the same medium composition. GrowMatch replaced Constraint (2), (3), and (6) in Figure 3 with a positive biomass production rate greater than a minimum threshold ($v_{biomass} \geq v_{biomass,min}$) and user-specified constraints for energy maintenance ($v_{atp} = v^{atp}$) and substrate uptakes ($v_{uptake} = v^{uptake}$). The objective function and other constraints remained. Similar to GapFill, GrowMatch generated hypotheses to fill metabolic gaps in draft or mutant models. Both GapFill and GrowMatch necessitate the use of MetaCyc or KEGG database; these resources, however, maintain a distinctive naming convention for all compounds, which hampers their direct applications on models created from other databases. Also, the two methods cannot fill gaps at the community level because they were designed to handle a single model.

1.6 Flux Balance Analysis

There are many ways to interrogate a model, but in this thesis the emphasis is placed on the use of Flux Balance Analysis (FBA), by which the metabolic flows in the reaction network can be quantified under certain assumptions. Briefly, it is assumed that steady state is maintained in the organism due to the high turnovers of metabolites, such that the concentration of every metabolite remains constant; in addition, each reaction flux is constrained between a lower and upper bound, which reflect the minimum and maximum rates of that reaction, respectively. Since most reaction rates are unknown, the values of the bounds are, by convention, large in magnitude and specified arbitrarily. In this fashion, FBA maximizes the flow through the growth reaction by varying other fluxes in the network such that their values meet the steady-state constraints and are confined within the imposed bounds (Orth et al., 2010). The methodology has been established as an effective tool in differentiating the most capable metabolic state of an organism from those states it cannot achieve (Edwards and Palsson, 2000).

1.7 Genome-Scale Models and Its Uses in Understanding Biological Systems

The flux balance approach is convenient because details about enzyme kinetics are not required to make predictions, but it was not until later when the whole genome of many organisms became available that it was used along with genome-scale models. Amid the sequencing efforts, the first genome-scale model, *E. coli* MG1655, was constructed and assessed via FBA by Edwards and Palsson (2000). In this gene deletion study, the authors probed the metabolic potential of *E. coli* by identifying a set of essential genes in the central metabolism and comparing it with growth phenotype data. As a result, they demonstrated that the growth predictions of the genome-scale model were accurate in 86% of the mutants constructed.

Following this work, genome-scale models were created to address an array of topics, which range from physiology characterization (Mahadevan et al., 2006; Çakır et al., 2007), industrial strain optimization (Park et al., 2007), to microbial interactions (Stolyar et al., 2007; Taffs et al., 2009; Freilich et al., 2011; Zomorrodi and Maranas, 2012). In these studies, however, the metabolic networks were all constructed from the genomes of isolates; in other words, the reactions were either derived from a single species or incorporated from representative genomes selectively. Yet, the need for modeling complex mixed cultures using reconstructed metagenome is unequivocal, as most microorganisms live in association with others and interact with the environment, which results in the observed phenotypes of the consortia. So far, community modeling has been performed by merging the genome-scale models of related organisms (Stolyar et al., 2007; Wintermute and Silver, 2010; Zhuang et al., 2011). The idea of a ‘pan-metabolic network’ was thus highlighted, and ‘metagenome-scale models’ were called for in recent years (dos Santos et al., 2012; Borenstein, 2012). Up until now, constraint-based modeling has yet to be applied to any metagenomic system.

Although a syntrophic relationship was suggested between the primary benzene degrader and the other microorganisms in the methanogenic culture, the microbial interactions in the community have not been examined closely. This is mainly due to the difficulties in capturing metabolic flows using experiments since the culture of interest is associated with low concentration of metabolites. Therefore, organism-specific models created based on the metagenome sequences are used instead to help elucidate other potential interactions.

In this thesis, a model based on a metagenome of a methanogenic benzene-degrading consortium is presented. The acquired metagenomic sequences were assembled, binned, and annotated, and from the reconstructed network, stoichiometric models were built. The models were then curated for cell growth on the specified minimal medium, both individually and collectively. In addition, a method to fill metabolic gaps at the community level was developed and tried on a faux *E. coli* syntrophic model before the application on the metagenome model. As described previously, benzene degraders live in syntrophy with hydrogenotrophic and acetoclastic methanogens and other microbes consuming the fermentation products. However, the presence of other metabolite exchange cannot be ruled out. The metagenome model built here, in combination with constraint-based modeling, provides an opportunity for the discovery of unknown interactions.

This approach is expected to help deepen the understanding of the metabolic behaviours in the community and ultimately generate hypotheses that can be tested in the laboratory.

2.0 Objectives

Although modeling has been applied to microbial communities, the models employed so far have been derived from fully-sequenced single genomes. In addition, an automated method that resolves metabolic gaps at the community level, allowing the exchange of metabolites among various organisms, is yet to be developed. In this thesis, metabolic models were created from the metagenome of a methanogenic benzene-degrading culture, and it was hypothesized that using the metagenome-scale models, the mathematical framework developed here can enumerate potential metabolite exchanges between the bacteria and archaea species in the benzene-degrading community.

In particular, the objectives of this thesis are as follows.

1. To construct organism-specific models based on the metagenomic sequences of a methanogenic benzene-degrading community

2. To establish a suite of computational tools for the curation of metabolic models created from metagenome sequences.

3. To provide hypotheses about the metabolic interactions among the members in the syntrophic community

The procedures taken towards all objectives are documented in Chapter 3. The outcome and discussion of Objective 1 are documented in Chapter 4-5, with Objective 2 addressed in Chapter 3.1, 3.5-3.7, Appendix D, and Appendix F. Finally, Objective 3 is covered in Chapter 6.

3.0 Method

The DNA sequences used in this thesis were acquired from the Joint Genome Institute (JGI). The DNA was sequenced by Sanger and Roche 454 sequencing platforms. Subsequently, these sequences followed through a series of manipulation outlined in Figure 1 and eventually yielded metabolic models representing the methanogenic benzene-degrading culture. In this chapter, each component of the computational pipeline is described.

3.1 *De novo* Assembly of Metagenomic Sequences of Varied Lengths

The nucleotide sequences of the methanogenic culture were assembled by JGI using Newbler, a program developed for the *de novo* assembly of 454 reads. However, the assembled sequences were overall short, with the largest contig being 26,536 bp. Therefore, in order to improve the contig length, assembly strategies (Abegunde, 2010) were applied, and two software packages, MIRA3 (Mimicking Intelligent Read Assembly version 3.2.0, Chevreax et al, 1999) and PHRAP (Gordon et al., 1998), were used to assemble the raw DNA reads. Before the assembly, 454 sequences were pre-processed by SSAHA2 (Ning et al., 2001), and the resulting reads were provided as an input to MIRA3. Although paired-end reads could be assembled in MIRA3, the software package was unable to scaffold and therefore could not link nor orient contigs using the distance information of the pairs. Auxiliary inputs such as library size and clipping information were also included in the assembly. The cloning vector sequences pUC18 were removed from the Sanger raw reads. As shown in Figure 4, the first strategy assembled all raw reads simultaneously using MIRA3, with and without technology-dependent parameters, yielding Trial 1 and 2, respectively. The second strategy consisted of two stages: a pre-assembly of 454 reads and the main assembly, in which 454 contigs were combined with Sanger reads by either MIRA or PHRAP (Trial 3 and 4). An assembly was selected based on statistical criteria and used in the following computational analyses.

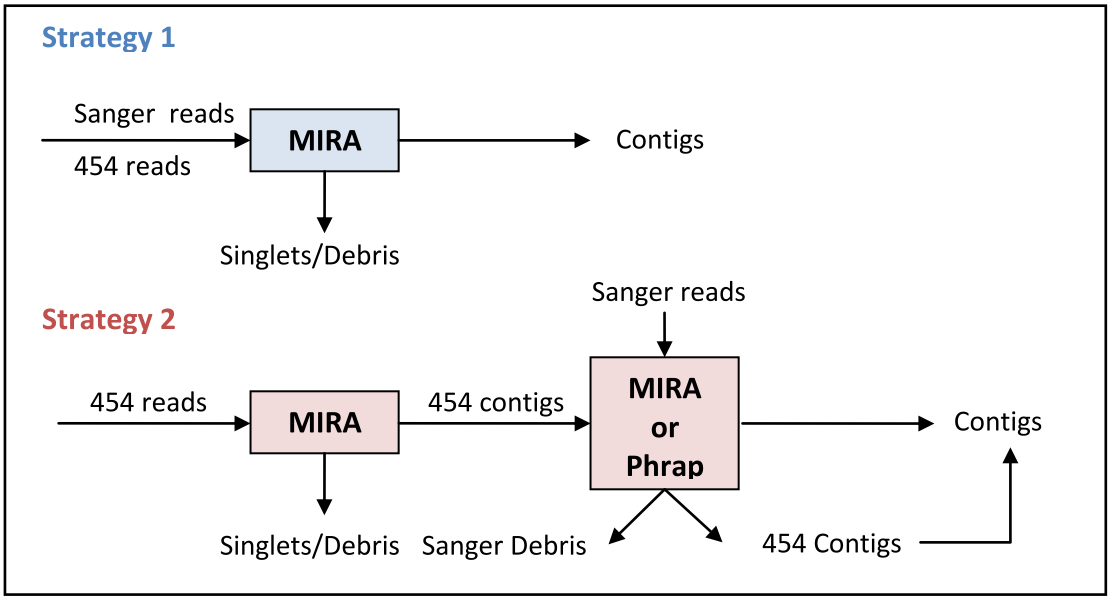

Figure 4. Assembly Strategies to Combine DNA Reads Generated Using Sanger and 454 Sequencing Technology

3.2 Identifying Contigs Related to Benzoyl-CoA Degradation Pathway

It was shown that the genes related to anaerobic benzoyl-CoA degradation were expressed during active benzene degradation (Devine et al., 2011). As a way to ensure the desired metabolic functions were present in the selected assembly, Trial 4 was screened computationally for the gene products of this pathway. The reference sequences of benzoyl-CoA degradation were chosen based on the discussion by Carmona et al. (2009) and obtained from the Protein Database of NCBI for the following organisms: *Rhodopseudomonas palustris* CGA 009 (accession number NC_005296), *Magnetospirillum magneticum* AMB-1 (accession number NC_007626), *Thauera aromatica* (accession number AJ224959), *Azoarcus sp.* Strain CIB (accession number AF515816), *Azoarcus sp.* Strain EbN1 (accession number NC_007759), *Geobacter metallireducens* GS-15 (accession number NC_007517), and *Syntrophus aciditrophicus* SB (accession number NC_007759). These genes were used to query the assembly by TBLASTN: if the E-value of a match was lower than e^{-40}, the gene product and hence the suggested metabolic function was considered present in the microbial community.

3.3 Binning the Metagenome by RITA and MG-RAST

The metagenome sequences were submitted for binning to Rob Beiko at the Dalhousie University. Binning is the computational process of classifying sequences into various taxa by comparing them to annotated sequences in a reference database. The benzene metagenome was binned by the classifier RITA, or Rapid Identification of Taxonomic Assignments (MacDonald et al., 2012). Essentially, the software package compared the contigs to the references based on sequence homology and nucleotide composition and assigned each sequence to a biological rank, which was set to genus. All available algorithms, including Naive-Bayes, D-BLASTN (Discontinuous MEGABLAST), BLASTN, BLASTX, were employed in the process. The maximum BLAST threshold was set to e^{-10}, and the minimum ratio difference between any two BLAST algorithms was 20. The accuracy of RITA is discussed in detail by MacDonald et al. (2012); briefly, there are 7 confidence groups: 1a, 1b, 2a, 2b, 3a, 3b, and 4, with Group 1a representing the most confident classifications and Group 4 being the least confident. The abundance counts of the bins were calculated based on the number of read bases used to build each classified contig. In addition, the Organism Abundance profile of MG-RAST (Meyer et al., 2008) was used to compare the results given by RITA. Finally, the sequences are grouped into bins according to their assigned domains rather than genera as previously determined (see

Section 5.1-5.2). As a result, the metagenome was organized into two bins, archaea and bacteria, for which metabolic networks were subsequently reconstructed.

3.4 Metabolic Reconstruction and Draft Model Creation

The genes and gene products were deduced for the archaea and bacteria bins using RAST (Rapid Annotation using Subsystem Technology, Aziz et al., 2008). Following the annotation, two draft models were created by the Model SEED (Overbeek et al., 2005), where putative reactions and metabolites were derived. *In silico* growth of the models in the SEED-defined complete medium was ensured by adding reactions from SEED's reaction database.

3.5 Individual Model Curation: Filling Metabolic Gaps in the Archaea and Bacteria Models

In order to produce a model capable of growth in the given mineral medium and to account for the known metabolic capabilities of the microbes, a Linear Programming problem was formulated based on GrowMatch (Kumar et al., 2007) to add reactions from an external reaction database to fill in the metabolic gaps in the model. Utilizing functions in the COBRA Toolbox 2.0 (Schellenberger et al., 2011), the algorithm (Figure 5) was coded in MATLAB and solved via CPLEX Interactive Optimizer.

As shown Figure 5, the framework attempted to add the least reactions ($\sum y_j$) from the database to keep the changes to the model minimum. Again, the model was maintained at steady state ($S_{ij}v_j$=0), and each reaction flux was bounded between -1000 and 1000 unless reversibility information was available ($LB_j \leq v_j \leq UB_j$). Every reaction in the database was assigned a binary variable y_j, which when equal to 1 signified the addition of the reaction and 0 the exclusion. The minimum growth rate was specified to be 10 units ($V_{\text{biomass}} \geq 10$), and if the organism was known to uptake or export certain metabolites, such information was specified to demand at least 10 units of fluxes for the exchange or transport reactions ($V_{\text{secretion}} \geq 10$ or $V_{\text{uptake}} \leq -10$).

$$\boxed{\begin{array}{l}
Minimize \sum_{j \in Datablase} y_j \\
s.t. \\
S_{ij} \cdot v_j = 0,\ \forall i \in M, \forall j \in N \\
LB_j \leq v_j \leq UB_j,\ \forall j \in Model \\
y_j \cdot LB_j \leq v_j \leq y_j \cdot UB_j,\ \forall j \in Database \\
v_{biomass} \geq 10 \\
v_{\mathrm{sec}retion>} \geq 10 \\
v_{uptake} \leq -10 \\
y_j \in \{0,1\}
\end{array}}$$

Figure 5. Mathematical Formulation of the Gap-Filling Algorithm for Restoring Growth in a Single Model

The gap-filling for the bacteria model was tried four times to accommodate the possible metabolite secretions: formate, acetate, hydrogen, and methanol, which are all known substrates for methanogenesis. For instance, to restore cell growth while forcing acetate secretion, $V_{\text{secretion}}$ shown in Figure 5 was constrained to be greater than 10 units for acetate. For the bacteria model, benzoate was provided as the sole carbon and energy source, even though the community is not known to grow on this substrate perhaps due to the lack of benzoate-CoA ligase or a benzoate transporter (Devine, 2013). Nevertheless, benzoate was chosen as the surrogate of benzene to channel carbons and electrons into the model because the pathway though which benzene could be activated was still unclear, and benzoate was proposed as one of the candidate compounds to which benzene was converted upon ring activation. The flux of benzoate uptake was forced to be at least 10 units as a gap-filling constraint. To ensure that benzoate would always be channeled to benzoyl-CoA, the enzyme benzoate-CoA ligase was added to the model and tagged as a reaction not associated with genetic evidence. Subsequently, the archaea model was gap-filled with acetate as the substrate for methanogenic growth. The uptake of acetate was restricted to be less than or equal to -10 units, and the export of methane was set to be greater than or equal to 10 units.

3.6 Model Curation: Filling Metabolic Gaps in an Artificial *E. coli* Community

The formulation to fill gaps at the community level is mathematically similar to that introduced in Section 3.5; it minimizes the total number of reactions added to the community while attempting to meet constraints in the individual organisms (Figure 7). The metabolites in each model as well as in its database were tagged to differentiate the cellular compartment in which the reactions take place. For example, the notation '[c]_2' refers to the intracellular metabolites in Model 2. Therefore, the intracellular metabolites in a model would not interact with those in the other model unless they were transported outside to the medium environment, where no distinction of compartment was made (Figure 6). In this way, the models were connected only by

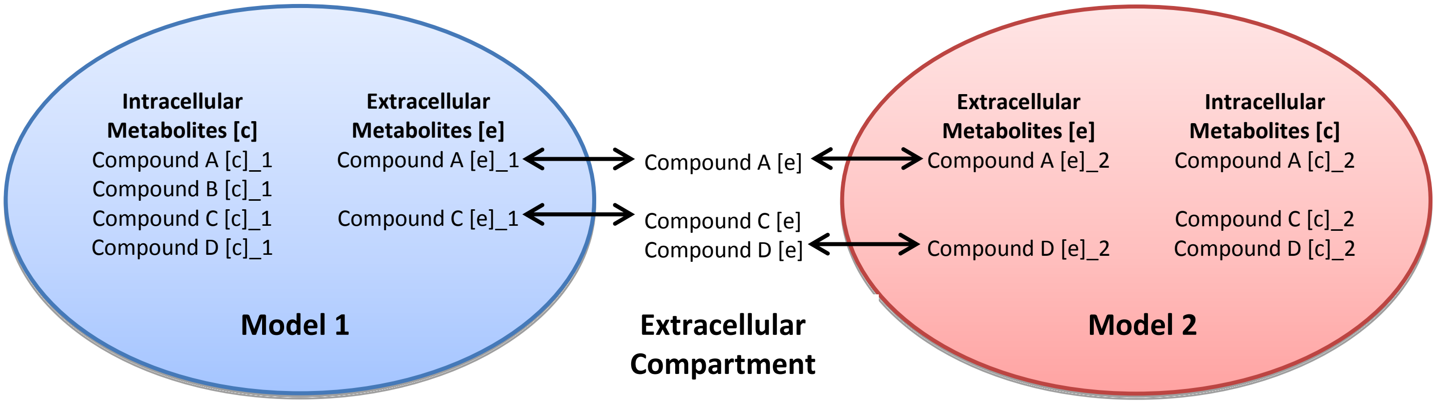

Figure 6. Conceptual Visualization of a Community Model with Two Organisms

extracellular compounds which were tagged with a suffix ‘[e]’. Since multiple models were considered, each model had a corresponding database to draw reactions from during the gap-filling process.

$$Minimize \quad \sum_{h \in Database1} y_h + \sum_{j \in Database2} y_j$$

$$s.t.$$

$$\left[\begin{array}{l} S_{gh} \cdot v_h = 0 \\ LB_h \le v_h \le UB_h,\ h \in Model\ 1 \\ y_h \cdot LB_h \le v_h \le y_h \cdot UB_h,\ h \in Database\,1 \\ v_{bio}^{1} \ge 0.9 \\ y_h \in \{0,1\} \end{array}\right]_{Organism\,1}$$

$$\left[\begin{array}{l} S_{ij} \cdot v_j = 0 \\ LB_j \le v_j \le UB_j,\ j \in Model\ 2 \\ y_j \cdot LB_j \le v_j \le y_j \cdot UB_j,\ j \in Database\,2 \\ v_{bio}^{2} \ge 0.05 \\ y_j \in \left\{0,1\right\} \end{array}\right]_{Organism2}$$

Figure 7. Gap-Filling Formulation at the Community Level for Two *Escherichia coli* Models

An artificial *Escherichia coli* community was constructed using two identical core *E. coli* models to test the expanded gap-filing algorithm (Figure 8). Modifications were made such that the first model (Ecoli1) took a role functionally similar to the benzene degraders in the methanogenic culture, except that it consumed only glucose as the carbon and energy source. Specifically, to imitate the incomplete nature of draft models generated by automated pipelines, the 8th step of glycolysis, phosphoglycerate mutase, was deleted. Additionally, reactions were deleted in TCA cycle and pyruvate metabolism to limit the model’s ability to grow while producing acetate. The second *E. coli* model (Ecoli2) was modified to be the designated acetate utilizer, so aside from the five reactions that were also deleted in Ecoli1, the glucose phosphotransferase system (PTS) was removed to restrict the uptake of glucose, and citrate synthase was deleted to further hamper the metabolism of acetyl-CoA, to which the uptaken acetate was channeled. As a result, the community consisted of an organism consuming glucose and handing off acetate and another that utilized acetate and ultimately produced carbon dioxide as the end product. The outcomes of these manipulations were two models, both unable to grow in the defined medium and possessing designated yet distinctive roles in the *in silico* community. Based on the growth simulation of the *E. coli* models before the deletions were made, the minimum growth rate was set to be 0.9 and 0.05 mmol·gDW^{-1}·hr^{-1} for Ecoli 1 and Ecoli 2, respectively, in the gap-filling

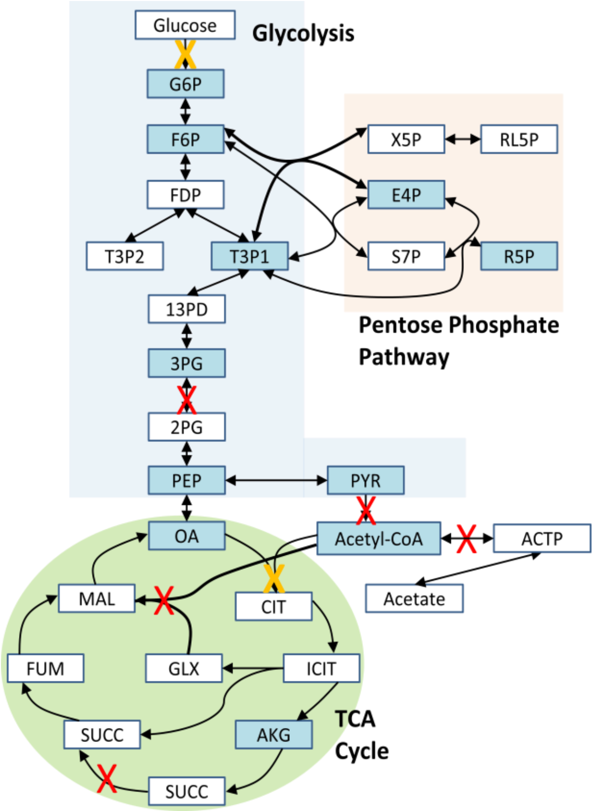

Figure 8. Deletions in the Faux *E. coli* Community. Blue boxes: biomass precursors; Red crosses: reactions deleted in both Ecoli1 and Ecoli2; Yellow crosses: reactions deleted uniquely in Ecoli2

formulation. The problem and the simulations were solved with COBRA Toolbox 2.0 and CPLEX Optimizer.

3.7 Growth Curation for Models of the Methanogenic Benzene-Degrading Culture

The same framework used in Section 3.6 was applied to models created from the metagenome of the methanogenic benzene-degrading culture, with Model 1 being the bacteria bin and Model 2 the archaea bin. The minimum growth rate thresholds ranged from 1 to 10 units. The bounds of the exchange reactions were set according to the decision tree in Figure 9. For the community gap-filling, exchange reactions were defined separately. The flux of ‘medium exchange,’ which takes the form of ‘met[e] <=> ,‘ represents the rate at which the compound met[e] leaves or enters the medium; the flux of ‘extracellular transport,’ with a formula ‘met[e] <=> met[e]_ID,’ is the rate at which met[e] is consumed or secreted by Model ID from the extracellular

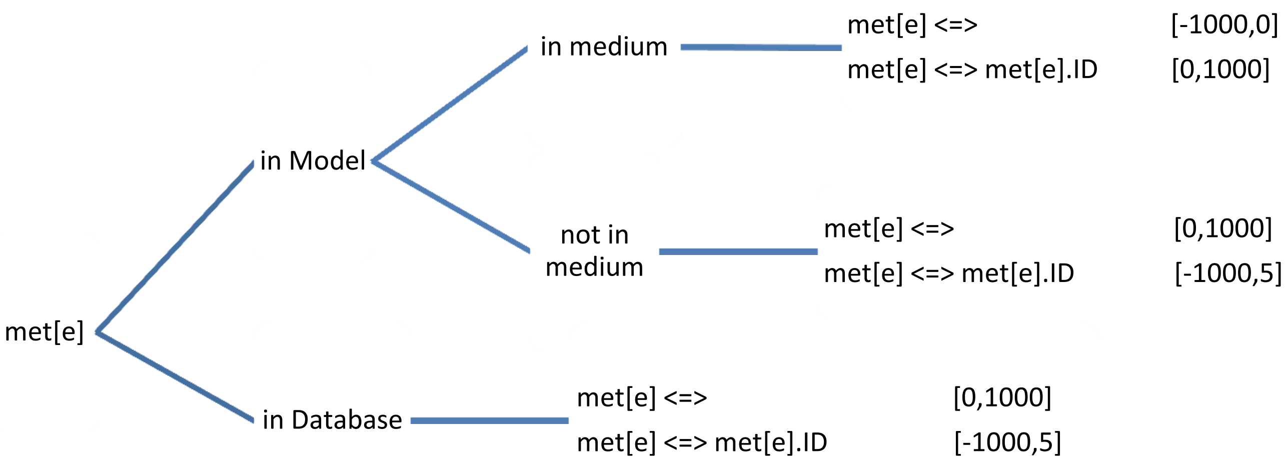

Figure 9. Decision Tree for the Bounds of Model and Medium Exchange Reactions Pertaining to an Extracellular Metabolite

compartment. This setting helped visualize how a metabolite was partitioned into each model when resources overlapped. All the extracellular metabolites that were not part of the mineral medium were allowed to be secreted into the medium at a maximum rate of 1000 units or consumed at a maximum rate of 5 units, whether they were already in the model or to be incorporated by the gap-filling algorithm. Constraining the uptake rates at 5 units served to prevent the overconsumption of those metabolites not provided in the defined medium, while providing an opportunity for other metabolite interactions. For future work, to help elucidate other key intermediates, the flux bounds of transport exchanges can be adjusted to [-1000, 1000]. Finally, while restoring growth of the two domain-specific models, the community gap-filling considered three exchange scenarios: acetate, hydrogen, and a case where no intermediate was specified. When acetate was the key exchange between the bacteria and archaea models, the acetate uptake flux of the archaea was set to be 10 units at least, and the same goes with the hydrogen scenario. In all three cases, benzoate uptake of the bacteria model and methane production of the archaea model were both constrained to be greater than or equal to 10 units in magnitude.

3.8 Preprocessing the Reference Reaction Database

The two reaction databases used in this thesis were obtained from Christopher Henry, one of the developers of the Model SEED; the Gap-Filling Database contained the reactions which SEED used to fill metabolic gaps, and the Whole Database consisted of all cellular and exchange reactions in SEED. Initially, the Gap-Filling Database was used for the growth curation of both bacteria and archaea models; however, for the bacteria bin, no feasible solution was returned presumably because the database, which has fewer reactions in comparison with the Whole

Database, did not contain the reactions needed to restore growth. It was therefore decided that the Whole Database be used for only the bacteria model.

Before the Whole Database was used to fill the gaps in a model, a Flux Variability Analysis (FVA) was performed to reduce the size of the database and hence computational intensity. Specifically, FVA determined the maximum and minimum flux of each reaction in the database, given the biomass equation and the specified extracellular environment. Those reactions with maximum and minimum fluxes of zero were then removed from the database since they could not carry any flux and contribute to cell growth given the conditions. For the bacteria bin, this approach downsized the database by 34%, bringing the total number of reactions to 7985. For the case of the *E. coli* community, the Gap-Filling database was reduced by 42% with a final size of 6100 reactions.

4.0 Sequence Assembly

4.1 Results: Pre-Assembly of Short Sequences Results in Longer Contigs

The assemblies were analyzed and selected for subsequent analyses based on four metrics: number of debris sequences, number of contigs with length greater than 5000 bp, large N50, and the total consensus (Table 2). The basic information of each assembly is summarized in Table A1 (Appendix A), and the contig length distribution of each trial is covered by Figure A5. Trial 0 was the assembly performed by JGI using Newbler on the same data, and the goal here was to produce longer sequences than those given by this assembly. As shown in Table 2, all trials performed with strategies did not assemble more reads when compared to the Newbler assembly, as indicated by the number of debris. The assemblies were also compared based on the number of contigs greater than a length threshold; the more contigs an assembly had greater than 5000 bp, the more genetic information it promised. Because the metagenome contains many repeated sequences possibly caused by strain variation, having more contigs will allow the preservation of more genomes at various abundances. This criterion is in contrast with the conventional view where fewer contigs also means fewer gaps in a single genome, given a similar debris size (Chaisson and Pevzer, 2008). In this case, both strategies could produce more contigs larger than the 5000 bp than Newbler did, and the assemblies completed by Strategy 2 generally gave more contigs that were longer when compared to those produced by Strategy 1 (Figure A5).

The traditional metric N50 was given a lower weight in these metagenome assemblies because it is affected by the intrinsic differences between the assembly algorithms. Nevertheless, all trails performed with assembly strategies had higher N50 compared to the Newbler assembly. For the assemblies completed by Strategy 2, the inclusion of the 454 contigs which were not assembled in the second stage contributed to lower N50 values; when these relatively short contigs were considered, the value of N50 reduced by 44% and 25% for Trial 3 and 4 respectively. Finally, in terms of total consensus, Trial 1, in which the differences between Sanger and 454 reads were disregarded, was the smallest of all, and Trial 4 stood out as the largest with 28 million bases, which were assumed to have the most genetic information. In summary, if both Sanger and Roche 454 reads are used, assembling the 454 reads first leads to longer sequences for this metagenome. Based on the criteria in Table 2, the MIRA-Phrap assembly (Trial 4) was selected for further metagenomic analyses.

Table 2. Statistical Comparison of Assembly Strategies

Trial #	0	1	2	3	4
Strategy	N/A	1	1	2	2
Assembler	Newbler	MIRA	MIRA	MIRA-MIRA	MIRA-Phrap
# contigs > 500 bp	12888	4820	9701	12654	13853
# contigs > 5000 bp	369	625	792	612	931
Large N50 (bp)	1782	3685	3077	2260	2619
# debris (million)	0.27	0.35	0.42	0.40	0.44
Total consensus (Mbp)	18	13	21	23	28

A targeted BLAST search then ensured the desired metabolic capabilities were carried through in the selected assembly. Since the benzoyl-CoA degradation pathway was expressed during benzene catabolism, the genes involved in this pathway were used as a reference to extract relevant sequences. As shown in Table A1 and A2 (Appendix A), contigs showing sequence similarity to benzoyl-CoA reductase and the enzymes of modified β-oxidation (acyl-CoA hydratase, hydroxylacyl-CoA dehydrogenase, and oxoacyl-CoA hydrolase) were indeed present.

4.2 Discussion: Evaluating and Comparing Assemblies

The rationale of using a hybrid approach on the benzene metagenome came from the inclination to preserve all sequences to help with the creation of organism-specific models. In order to achieve this, the genomes on which the models are based must be as complete as possible. The idea was also influenced by Abegunde's work (2010), in which the performance of hybrid assemblies was evaluated. As assemblers, PHRAP and MIRA were acknowledged for the ability to handle 454 reads, and Abegunde (2010) suggested that strategies be devised when data were generated from different sequencing sources. Specifically, if raw reads vary markedly in length, short sequences should be assembled first before being combined with longer ones. Surely for certain genomes, hybrid assemblies combining Sanger and pyrosequenced data were proven effective (Goldberg et al., 2006). In this study, the authors assembled six marine microbial genomes using Sanger reads sequenced at various coverage and 454 pseudoreads. The pseudoreads were generated by assembling 454 raw reads and shredding the resulting contigs into longer fragments, which mimicked the lengths of Sanger reads. This strategy was shown to increase contig length and reduce genome gaps because 454 data, insensitive to cloning bias, complemented Sanger reads by adding to low coverage regions. On the other hand, Sanger sequences could compensate 454 reads when polymorphic areas and large sequencing gaps prevail.

In this thesis, the criteria for choosing an assembly were loosely based on the metrics used in the short-read assembly of *E. coli* and *S. pneumoniae* (Chaisson and Pevzner, 2008). In the publication, assemblies were compared using the number of long contigs, total bases in long contigs, and N50. However, these criteria focus on the size of an assembly and do not answer to contig quality. The number of debris was included in the statistics due to the initial goal of assembling more reads given more sequenced data, i.e. Sanger along with 454, but the number of assembled sequences could be dependent on the sampled environment, and it is likely that more input data does not result in larger assembly size or higher coverage (Chain, 2011). Finally, total consensus was included as a criterion simply because of the necessity to preserve more genomic evidence for the metabolic models.

It is acknowledged that to fully validate an assembly, diversified methods and iterations are required (Pop, 2009). To address qualities, contig coverge and number of mis-assembled regions should be added to the comparison. For example, with a visualization tool, it is possible to point out regions with unusually high coverage due to repeated segments, which are common in metagenomic data. Moreover, probability-based metric, such as one that leverages between extending and cutting off a contig, can also serve as an unsupervised method to assess the quality of assembled sequences. Specifically, when presented with a pool of nearly identical sequence reads, an assembler can either extend a contig using these fragments or simply disconnect, resulting in more and perhaps redundant sequences in the final output. In fact, an evaluation metric addressing this issue was recently proposed (Laserson et al., 2011). In this scoring system, a high penalty is given when a read joins a contig region that is not yet aligned by other reads, and such costs in score can prompt the assembler to instead use that read to build another contig. Such a system may compromise on the length of the contigs as well as the number of sequences ending up in the output, but it may help overcome the varied abundance and heterogeneity in metagenome by retaining more sequences for the minor organisms. Another unsupervised metric, Feature Response Curve (FRC), was introduced by Narzisi and Mishra (2011) to address the lack of quality-indicative criteria. The FRC associates genome coverage with several error features, such as k-mer frequencies, polymorphic alignments, and mate-pair checking, and allows the users to evaluate assemblies along with other popular metrics like N50 and number of contigs.

In this section, it was found that assembling the short read first improved the overall length of the output sequences. Moreover, the inclusion of more sequenced data did not necessarily lead to more reads being assembled. In general, metagenomes are more difficult to assemble than a single genome due to issues such as incomplete sampling, sequencing bias, low coverage, strain variation, and the risk of falsely joining sequences from different species (Wooley et al., 2010). These problems were all assumed to be present to a certain degree in the metagenome of the methanogenic benzene-degrading culture because the DNA samples were taken at various time points and from multiple methanogenic cultures. In addition, it is recognized that an assembly with longer contigs is not necessarily a better assembly or one with more biological meanings, which could only be deduced from functional and phylogenetic annotations. Therefore, an accurate assembly can only be generated iteratively by the use of a wide variety of validation tools.

Recently, the metagenome of the same culture was sequenced again using Illumina sequencing technology. The DNA sample was taken from a culture at a single time point to minimize the effects of heterogeneity on the sequenced reads. Multiple assemblies based on this new data were produced by varying kmer lengths and the size of the input raw reads (Devine, 2013), and it was reported that decreasing kmer lengths facilitated the assembly of those sequences originated from low abundance organisms, while a higher kmer length might be more suitable for assembling high abundance DNAs. Essentially, due to the varied organism abundance in the metagenome, there was not a single assembly universally optimal for all genomes in the benzene-degrading consortium; however, bins with high confidence level can be extracted from these assemblies to represent the groups of microbes in the community. Therefore, it is recommended that these assemblies be passed through the computational pipeline developed in this thesis for a more refined community model.

5.0 Results and Discussions: Taxonomic Classification and Model Creation of the Methanogenic Benzene-Degrading Culture

5.1 Results: Binning the Metagenomic Sequences from the Methanogenic Culture

Binning the benzene metagenome entailed assigning sequences in the chosen assembly into taxonomic bins at the rank of genus. The number of sequences submitted to RITA was 17482, of which 16324 were binned. However, 81% of the classified contigs fell into the group with the lowest confidence or Group 4 as described by MacDonald et al. (2012), and only 3% was binned with the highest confidence level (Group 1a). In addition, binning by RITA resulted in contigs spreading across as many as 567 bins with only a few sequences in each, which made it difficult to create genus-specific metabolic models.

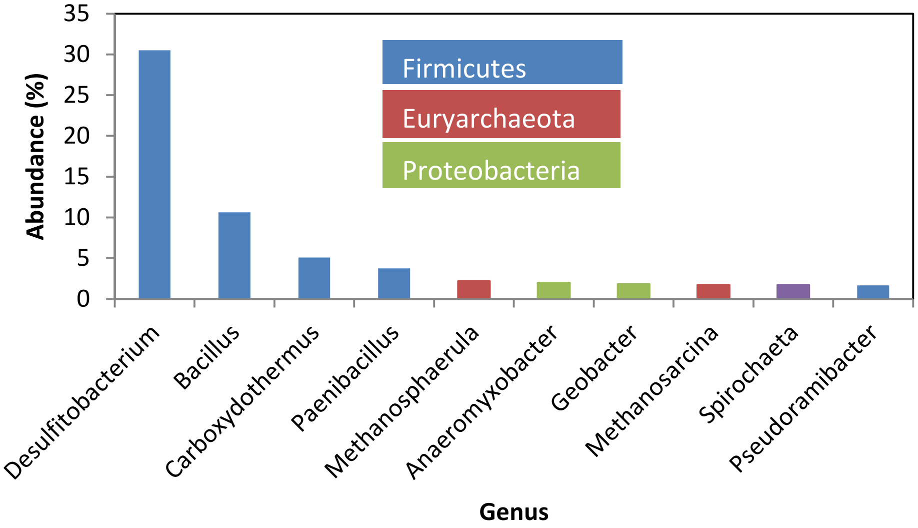

Figure 10. Taxonomic Classification by RITA: Ten Bins with Most Read Bases in the Benzene-Degrading Culture

A closer examination on the bins with higher abundances based on number of base pairs can be seen in Figure 10, where the ten largest genera predicted by RITA are listed. Among the ten bins, five belonged to the phylum *Firmicutes*, which, according to RITA, made up 52% of the culture. However, based on qPCR surveys and 16S rRNA analysis performed previously, the methanogenic culture is dominated by *δ-Proteobacteria* (OR-M2), so ideally, the majority of the sequences should be of the phylum *Proteobacteria*. However, along with those in the top ten bins, only 13% of the binned sequences were determined as *Proteobacteria* by RITA, and the contigs with 16S rRNA sequences of OR-M2 were found to classify under *Firmicutes*. The same data was submitted to MG-RAST, an online resource for functional and phylogenetic analysis, and it was found that at the phylum level, 37% of the sequences were *Proteobacteria*, 17% *Firmicutes*, and 18% *Euryarchaeota*. It is noteworthy that MG-RAST detects gene products and ribosomal RNA genes in the contigs and calculates the organism abundance by summing up the number of detected features that are linked to taxonomic ranks. Nonetheless, it was evident that the approach taken by RITA was not effective for the data here as it was to other metagenomes.

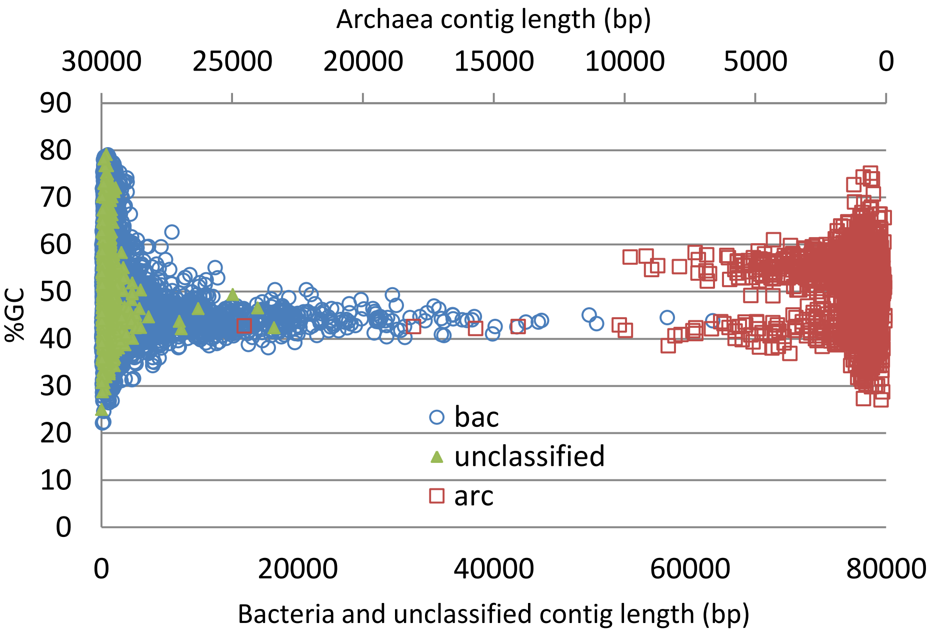

Figure 11. Domain-Specific Metagenome Assembly of the Methanogenic Benzene-Degrading Culture

The construction of organism-specific models relies on binning to segregate the sequences into their respective taxonomic group, from which separate metabolic networks are reconstructed. Therefore, to increase the confidence level, the sequences were organized at the level of domain rather than any lower biological ranks. In this case, the metagenome consisted of 8.7% of archaea and 90% of bacteria according to RITA. On the other hand, based on the taxonomic information of MG-RAST, archaea and bacteria accounted for 19 and 79% of the metagenome, respectively. Again, in MG-RAST, a contig could be counted multiple times or not counted at all, depending on whether the sequence contained features indicative of taxonomy. According to the 16S rRNA analysis performed previously on the same sequencing data, the community was composed of 14% of archaea and 86% of bacteria based on total raw reads (Devine, 2013).

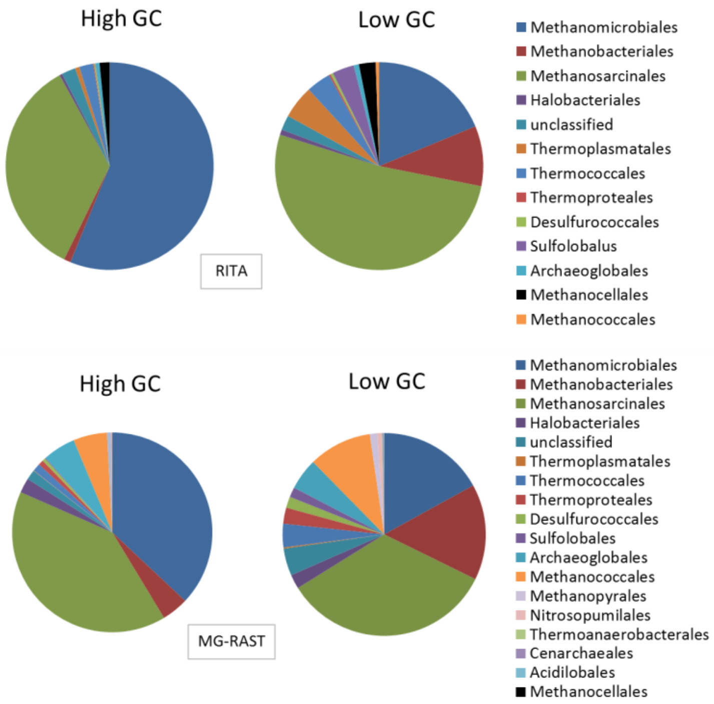

Figure 12. A Taxonomic Comparison of the High- and Low-GC Groups in the Archaea Bin

With the domain information, the assembly were visualized further as shown in Figure 11 and Figure A1 in Appendix A. Although the sequences did not form any discernible trend at various levels of coverage (Figure A1), the GC content of most contigs in the bacteria bin fell between 40 to 52%, while those in the archaea bin appeared to centre around GC content of 55% and 42%. A closer examination on the high- and low-GC peak in the archaea bin yielded Figure 12. Most contigs with a higher GC content were classified as *Methanomicrobiales* and *Methanosarcinales*. While the low-GC group also observed this aspect, it appeared more enriched by the presence of *Methanobacteriales* and other diverse members. A similar trend was also seen in the data analyzed by MG-RAST.

5.2 Discussion: Binning Novel Sequences at the Rank of Genus

The discrepancy between RITA's prediction and the experimental data at the genus level could be attributed to the novelty of the benzene-degrading culture and the algorithm being a supervised approach. Due to the absence of similar organisms or proteins in the database, RITA had no choice but to classify the sequences into distantly related lineages. Moreover, when there were no homologous sequences in the reference database, RITA inevitably relied more on composition-based algorithm, which caused false classification because homology was always deemed a stronger piece of evidence for a match than composition. In addition, similar composition profiles, such as codon usage frequency, could be shared by distinct organism; however, whether *Firmicutes* and *δ-Proteobacteria* have similar genomic contents is yet to be verified. These helped explain why a low confidence level was associated with more than eighty percent of the metagenome, for it was binned based mostly on genomic composition. Using the rank of genus in the classification might have been too ambitious a choice because it forced the classifier to assign a sequence to the lowest possible taxon, while a higher rank would have sufficed. However, the confidence level could be increased by organizing the sequences at a higher rank, as shown by the agreement between RITA's domain classification and the 16S analysis.

In the archaea bin, most sequences originated from methanogenic species, *Methanomicrobiales* and *Methanosarcinales,* which are termed Class II methanogens (Anderson et al., 2009), as well as *Methanobacteriales,* which belongs to Class I methanogens. Based on Figure 12, both binning platforms recognized the presence of all three orders. It is important to note that rather than phylogenetically analyzing the metagenome, the purpose of binning was to organize the sequences into groups so that each bin reflected the corresponding genomes as well as its functional role in the benzene-degrading community. The downstream microbes, such as methanogens, consume fermentation products and hydrogen, and therefore in this case, one should expect that ultimately in the archaea model, there will be reactions from both acetoclastic (*Methanosarcinales*) and hydrogenotrophic methanogens (*Methanomicrobiales* and *Methanobacteriales*).

Through binning, the metagenome was simply split into two groups at the domain level, representing the bacterial and archaeal species in the benzene-degrading enrichment culture. This was to increase the confidence level associated with each classified sequence. The challenge of creating rank-specific models for the benzene-degrading culture was clear; if binning could not be performed with certain confidence, then models built from the bins would not represent the community members faithfully. For future recommendations, it is necessary to assess the role of composition-based algorithm in binning metagenome, it is also essential to select the rank of bins in keeping with the extent to which a community is characterized. Particularly, sequence homology should be emphasized over genomic composition; in this case, using RITA, one can switch the composition binning algorithm in the first two stages with the BLAST algorithms originally used in later stages, such that homology is given a greater priority and that the composition-based Naïve-Bayes can only interfere when similarity-based BLAST algorithms fail to classify a sequence. On the other hand, an unsupervised approach may be used instead to bypass the database bias against novel metagenomes. In fact, the new Illumina assembly recently obtained was binned based on kmer coverage and tetranucleotide frequency clusters (Devine, 2013), and the agreement between these unsupervised methods helped the creation of taxonomic bins at various organism abundances. Furthermore, network reconstruction can be performed for these bins, which supposedly contain sequences from organisms of similar abundance, and this in turn will increase the accuracy and confidence level of the organism-specific models.

5.3 Results and Discussions: Generating Draft Metabolic Models

The two bins were submitted to RAST for annotation (Figure A2-A3). For the archaea bin, most gene products were linked the metabolism of carbohydrates, amino acids, DNA, RNA, and protein. While those in the bacteria bin shared a similar distribution, the metabolism of cofactors and vitamins also accounted for a significant portion in the subsystem coverage. Metabolic models were then reconstructed directly from the annotated bins in Model SEED (Table 3).

The draft bacteria model had 1020 metabolites and 1080 reactions, of which 48 were not gene-associated because they were added by SEED for growth in the defined rich medium. In the archaea model, there were 518 compounds and 436 reactions, and likewise, 58 were added to allow growth. However, reactions in the benzoyl-CoA catabolic pathway were not fully included in the bacteria model, with only 3 reactions, leading to the formation of acetyl-CoA in the lower benzoyl-CoA pathway, being present. Why the Model SEED excluded these reactions from the model reconstruction was unknown. In SEED's original paper, Henry et al. stated that "reactions are included in the preliminary model if one or more of the functional roles associated with these reactions in the SEED have been assigned to one or more of the genes in the annotated genome." While most genes in the benzoyl-CoA degradation pathway were annotated by RAST, they did not transform into reactions in SEED, perhaps due to the lack of related proteins or reactions in the SEED database. Similarly, some reactions of the methanogenesis pathway were neglected in the archaea model. Because the annotation process of tools as such RAST and SEED is often not transparent, users must exercise cautions as known metabolic capabilities could be lost in the draft models created. In this case, the upper and lower benzoyl-CoA pathways were manually added to the bacteria model, and in the archaea, gaps in the methanogenesis pathway were filled in the same manner.

Table 3. Metabolic Reconstruction of the Methanogenic Benzene-Degrading Community

Model	**Bacteria**	**Archaea**
Size (Mbp)	24.3	2.8
Protein features	15616	2563
Identified protein features	9912	1726
Reactions	1080	436
Gene-associated	1032 (96%)	378 (87%)
Gap-filled (by SEED)	48 (4%)	58 (13%)
Compounds	1020	518

5.4 Results: Curating Individual Draft Models Based on Growth

Although the auto-completion step in SEED added reactions to each model to allow growth in the complete medium, both models were unable to produce biomass in the environment corresponding to the mineral medium used in the laboratory. For metabolic models, the inability to synthesize biomass implies that some reactants in the assigned biomass equations cannot be made using existing pathways; in other words, some biomass precursors are missing. As shown in Table 4, many amino acids and cofactors could not be produced in the archaea model, and in the bacteria model, compounds essential to the biosynthesis of cell wall and lipids also could not be synthesized.

The goal of filling metabolic gaps in the individual models, rather than the two models collectively, was to exhaustively identify the relationship between the gaps and the reactions. This was done in the absence of interactions with another group of organisms. In other words, the non-native reactions suggested by the algorithm were not added to the models permanently; instead, they were used to infer gaps in the metabolic network. The reactions were only incorporated temporarily into the draft model for Flux Balance Analyses, from which growth rates were compared relatively.

For the curation of the bacteria model, the minimum growth rate was set to 10 units in the gap-filling algorithm, and potential metabolic capabilities were also defined; specifically, four excretion scenarios were considered: formate, acetate, hydrogen, and methanol. In all scenarios, benzoate was provided as the sole carbon and energy source, whose exchange flux ranged from -1000 to 0. The results are summarized in Table 5-6 and B1-B2 in Appendix B.

Table 4. List of Missing Biomass Precursors of the Draft Models in Minimal Medium

	Amino Acid		Cofactors and Cell Wall Components	
Archaea	Aspartate	Valine	CoA	Pyridoxal phosphate
	Phenylalanine	Serine	NADP	S-Adenosyl-L-methionine
	Asparagine	Isoleucine	FAD	Spermidine
	Glutamine	Glycine	NAD	Bactoprenyl diphosphate
	Tryptophan	Tyrosine	CTP	Peptidoglycan polymer
	Lysine	Glutamate	GTP	Acyl-carrier protein
	Leucine	Cysteine	UTP	
	Histidine	Arginine	dATP	
	Proline	Methionine	ATP	
	Threonine	Alanine	TTP	
Bacteria	Cofactor	Amino Acid	Lipid Synthesis	
	Pyridoxal phosphate	Methionine	Dianteisoheptadecanoyl-phosphatidylethanolamine	
	Tetrahydrofolate	Cysteine	Phosphatidylglycerol dioctadecanoyl	
	5-Methyltetrahydrofolate	Tryptophan	phosphatidylethanolamine dioctadecanoyl	
	2-Demethylmenaquinone 8	Lysine	Diisoheptadecanoylphosphatidyl-ethanolamine	
	S-Adenosyl-L-methionine	Isoleucine	Diisoheptadecanoyl-phosphatidylglycerol	
	Spermidine	Leucine	Dianteisoheptadecanoyl-phosphatidylglycerol	
	Menaquinone 8	Histidine	Isoheptadecanoylcardiolipin	
	10-Formyltetrahydrofolate	Threonine		
	CoA		Cell Wall Synthesis	
	ATP		Stearoylcardiolipin	
	TTP		Anteisoheptadecanoyl-cardiolipin	
	TPP		Peptidoglycan polymer	
	FAD			
	NADP			
	Acyl-carrier protein			

As shown in Table 5, when formate was secreted, 13 reactions must be added to the model to support growth. Two reactions (F1 and F2) from the thiamine metabolism restored the production of thiamin diphosphate, and F3 and F7 were added mainly for lipid and cell wall synthesis. Reaction F4 was highly connected in the network because it was essential for the production of methionine, cysteine, S-adenosylmethionine, and coenzyme A, while reaction F5 filled a gap in peptidoglycan synthesis and contributed also to the production of lysine. Reactions F11, F12, and F13 were added to balance the byproducts generated in spermidine synthesis; particularly, F13 was suggested by the algorithm in all secretion scenarios (i.e., A13, H13, and M15).

Table 5. Reactions Added to the Bacteria Model with Enforced Formate Secretion

Reaction ID	Reaction Name	Subsystem
F1	Thiamin-ABC transport	Transport
F2	ATP-thiamine phosphotransferase	Thiamine metabolism
F3	L-Threonine acetaldehyde-lyase	Glycine, serine and threonine metabolism
F4	Hydrogen-sulfide ferredoxin oxidoreductase	Sulfur metabolism
F5	R04519 transferase	Peptidoglycan biosynthesis
F6	Hydrogen transport	Transport
F7	C18 fatty acid biosynthesis	None
F8	Transport of formate	Transport
F9	Transport of benzoate (cytoplasm)	Transport
F10	Transport of benzoate (periplasm)	Transport
F11	5-Methylthioadenosine transport	Transport
F12	Exchange of urea	Exchange
F13	Exchange of 5-Methylthioadenosine	Exchange

All other missing biomass precursors listed in Table 4 were fixed by incorporating the transport reactions of benzoate, F9 and F10, indicating that the draft model was unable to grow mainly because it lacked a route to import the main carbon and energy source from the medium. In other secretion scenarios, transport reactions of benzoate served the same purpose, restoring the connectivity of most missing biomass precursors.

In the second scenario where acetate secretion was demanded (Table 6), acetate excretion via proton symport (A1) was the only acetate-related reaction suggested by the algorithm, implying that acetate-producing reactions were already in the draft model, but means to export it were required. Reaction A2 was responsible for the synthesis of threonine, isoleucine, and compounds involving in lipid metabolism. Similarly, Reaction A3 and A8 served to resolve gaps in thiamine metabolism to yield thiamine diphosphate, and A4, a component of sulfur metabolism, enabled first the sulfide-dependent production of cysteine, which then led to methionine. Spermidine synthesis was again completed by A5 and A6, and its byproducts were secreted via A12 and A13.

Table 6. Reactions Added to the Bacteria Model with Enforced Acetate Secretion

Reaction ID	Reaction Name	Subsystem
A1	Acetate transport in/out via proton symport	Transport
A2	(R)-2-Methyl-3-oxopropanoyl-CoA 2-epimerase	Propanoate metabolism
A3	ATP-thiamine phosphotransferase	Thiamine metabolism
A4	Hydrogen-sulfide NADP+ oxidoreductase	Sulfur metabolism
A5	Urea carbon-dioxide ligase	Urea cycle and metabolism of amino groups
A6	R04519 transferase	Peptidoglycan biosynthesis
A7	NADH dehydrogenase	None
A8	Thiamine transport in via proton symport	Transport
A9	b-ketoacyl synthetase (n-C181)	None
A10	Transport of benzoate (cytoplasm)	Transport
A11	Transport of benzoate (periplasm)	Transport
A12	5-Methylthioadenosine transport	Transport
A13	Exchange of 5-Methylthioadenosine	Exchange

Notably, Reaction A7 did not contribute to any biomass precursor directly but was coupled to the transport of benzoate via periplasmic protons. It is unlikely that A7 and the transport reactions of benzoate are used by the benzene-degrading culture since the culture does not grow on benzoate. However providing the model with a carbon and energy source was necessary, so these reactions were accepted for the time being. For the last two scenarios where hydrogen (Table B1 in Appendix B) and methanol (Table B2 in Appendix B) production were demanded, the reaction profiles followed a similar pattern as those presented in Table 5-6, and the diffusion term for hydrogen and methanol was added in each case, as required by the constraint of metabolic capabilities. Based on the fact that only an export term was added, the draft model indeed had gene-associated reactions able to give off hydrogen, but it did not seem equally capable in the production of methanol because two added reactions, M4 and M5, became the pathway through which methanol was generated in the model (Table B2).

Gap-filling is a computationally intensive process in the curation of draft models. In all cases above, it took at least 1 day to a week before the algorithm found the optimal solution or aborted the calculation upon a user-specified deadline; in the latter case, the algorithm reported the best solution in the solution pool at the point of termination. In light of the similar reaction profiles recommended by the algorithm and given the running time, the gap-filling for the archaea model was performed only with acetate as the carbon and energy source.

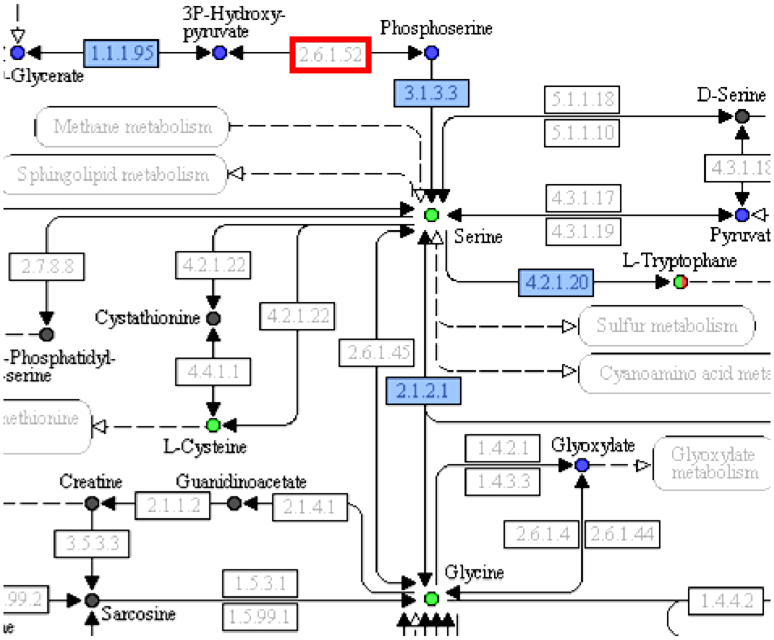

Figure 13. Filling a Gap in the Synthesis of Glycine and Serine. Red box: Reaction 5, EC 2.6.1.52, added by algorithm; Blue boxes: gene-associated reactions in the draft archaea model; White boxes: Reactions not in the model; Blue Dots: Represented metabolites in the model; Green dots: Biomass precursors; Red dot: Transportable compound; Grey dots: Metabolites not in model.

In total, 53 reactions were proposed by the algorithm (Table B3) for the archaea model. It appeared that the archaea model had more metabolic gaps compared to the bacteria model, which was a large network assumed to contain many functionally redundant reactions (Figure A4). The gap-filling process for the archaea model almost always reached the termination deadline; in other words, the solution reported here might not be optimal. Nonetheless, to minimize the changes made to the model, some solutions contained reactions that were highly connected, as indicated by the number of biomass precursors associated with them; for example, 3-Phosphoserine-2-oxoglutarate aminotransferase (Reaction 5 in Table B3, EC 2.6.1.52) filled a gap in the metabolism of serine, which lead to the production of other downstream amino acids (Figure 13). Moreover, transport terms for important carbon compounds, methane as well as acetate, were added to connect the intracellular and extracellular compartments.

5.5 Discussion: Curating Individual Draft Models Based on Growth

Filling gaps in metabolic pathways can be performed either manually by tracing the missing metabolite in the network or automatically using computational programs. While the former is conceptually easy, it is a tedious task that relies heavily on literature and physiology data. Due to the scale of the two models, biomass production was restored using computational methods. The mathematical framework adopted in the previous section, however, was not the first of its kind because it was formulated based on the algorithm GrowMatch (Kumar and Maranas, 2009). Although it was possible to apply GrowMatch on both models, such a task was deemed operationally difficult due to the unique naming convention of the metabolites in SEED; many compounds and reactions in SEED were not even given names in the KEGG database used by GrowMatch. Therefore, the gap-filling framework proposed here is different from GrowMatch that it adheres to the SEED's naming convention, and it integrates user-defined metabolic capabilities such as the secretions of certain compounds.

As shown in Table 7, the number of added reaction was similar in all cases. Given the benzoate

Table 7. Comparison of the Gap-Filled Bacteria Model Producing Various Intermediates

Secreted Metabolite	Formate	Acetate	Hydrogen	Methanol
Reactions added	13	13	13	15
Growth rate (units)	135	135	88	92
Benzoate uptake rate (units)	722	876	605	640

uptake rates, the growth rate in the case of formate-producing bacteria was the highest compared to that of the other three scenarios. In terms of carbon flow, this is unlikely because in order for formate to arise from acetyl-CoA, the end product of the lower benzoyl-CoA degradation pathway, acetyl-CoA must incorporate CO_2 to become pyruvate which then turns into formate. This series of reactions does not benefit the bacteria energetically because no ATP is generated. On the other hand, in the acetate-secreting case, ATP can be generated via substrate-level phosphorylation when acetyl-CoA is converted into acetate. The higher growth rate associated with formate secretion could be explained by Reaction F8 (Table 5), in which formate was exported to the periplasm along with a proton; this proton then aided the transport of benzoate in Reaction F9, providing the model with more carbon source. Therefore, the growth rates presented in Table 7 could only be interpreted relatively because the models were not constrained thermodynamically. When hydrogen was the only fermentation product, the growth rate was low compared to other scenarios. Whether the benzene degrader is able to shift its metabolism to preferentially give a single fermentation product is unknown, but biologically, the lower growth rate may indicate that fermenting benzoate to only hydrogen is not a rewarding strategy for the bacteria in terms of growth. It might suggest a weak hydrogen producer. Typically, hydrogen is produced along with other acids, mainly to regenerate electron carriers, but the yield may vary in different species (Nath and Das, 2004). Although the synthesis pathways for formate, acetate, and hydrogen are already represented in the model, the production routes of these compounds need to be examined before further interpretations can be made.

The gap-filling framework provided hypotheses about the metabolic gaps in the models and possible ways to fill them, and indeed more efforts were needed to verify these reactions. For the archaea model, the proposed reaction list was compared to the only publicly available methanogen models (Feist et al., 2006; Kumar et al., 2011), *Methanosarcina barkeri* iAF692 and *M. acetivorans* iVS941. Of the 53 suggested reactions, 32 and 33 reactions were shared by iAF692 and iVS941, respectively . Most of these shared reactions were gene-associated in the published models. The validity of the proposed reactions could also be glimpsed through their positions in the corresponding pathways. If a metabolic gap consisted of a single reaction instead of a series of reactions, it was more likely that the added reaction belonged to the model. For example, a gap in the lysine biosynthesis was filled by the enzyme aspartate-4-semialdehyde hydrolyase (EC 4.2.1.52) catalyzing Reaction 22 (Figure 14). The reaction was flanked by other downstream and upstream reactions which were gene-associated, and the metabolites it involved, aspartate-4-semialdehyde and dihydrodipicolinate, were also affiliated with other reactions in the network.

On the other hand, the reaction profiles (Table 5-6 and B1-B2) of the bacteria model were compared to the SEED model of *Geobacter metallireducens* GS-15, a *δ-Proteobacterium* known to metabolize various aromatic compounds under anaerobic conditions and was also reported for the degradation of benzene (Zhang et al., 2012). However, in each scenario, no more than 4 reactions were found in GS-15, and most shared reactions were not supported by genomic evidence in the *Geobacter* model. Finally, about half of the reactions in the four profiles were sided by neighboring reactions, with the other half consisting mainly of transport reactions and those not classified under any functional category.

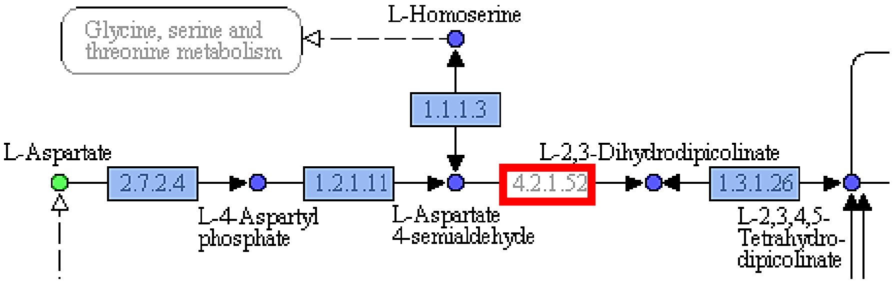
Glycine, serine and threonine metabolism
L-Homoserine
1.1.1.3
L-Aspartate
2.7.2.4
1.2.1.11
4.2.1.52
L-2,3-Dihydrodipicolinate
1.3.1.26
L-4-Aspartyl phosphate
L-Aspartate 4-semialdehyde
L-2,3,4,5-Tetrahydro-dipicolinate

Figure 14. Reaction 22 (EC. 4.2.1.25) in Lysine Biosynthesis. Blue: Gene-associated reactions in the archaea model. Red: Reaction 22, added by the algorithm

Whether the added reactions belong to these models is yet to be validated, bioinformatically or experimentally. In the construction of genome-scale models, reactions that are gene-associated are regarded more convincing than those added based on gap-filling results. When a missing reaction is pinpointed by a gap-filling algorithm in a well-represented functional subsystem, it might be useful to look for mis-annotation or sequences neglected by model-building programs. Should strong genomic evidence be found, corresponding reactions should be incorporated into the model. Nevertheless, metabolic gaps in the bacteria and archaea model of the benzene-degrading culture were partly elucidated, and hypotheses about how growth could be restored in the individual models by non-native reactions were generated.

6.0 Results and Discussions: Gap-Filling at the Community-Level

The idea of gap-filling at the community level stemmed from the role that metabolite exchange could play in completing the pathways of either group of organism and thus allowing community-wide growth. In the methanogenic benzene-degrading consortium, it was unclear if there were more potential exchanged components beyond acetate and hydrogen. To answer these questions, the mathematical framework in Chapter 5 was expanded to accommodate multiple models. The algorithm was first applied to an artificial *E. coli* community as a proof of concept and then used on the models created from the benzene metagenome. As a result, hypotheses about other transferable compounds in the methanogenic culture were given.

6.1 A Test Case: Filling Gaps in an Artificial Community Model of *E. coli*

A community model of two *Escherichia coli* strains was constructed to imitate the two bins in the benzene-degrading culture so that one model acted as the primary consumer and the other as a downstream organism. The two models were constructed separately from an *E. coli* model having 72 reactions, representing the organism's central metabolism. Before the deletion of reactions, the original model grew at 1.26 mmol·gDW^{-1}·hr^{-1}, consuming glucose, phosphate, and oxygen at 10, 4.6, and 5 mmol·gDW^{-1}·hr^{-1}, respectively. In total 12 reactions were deleted from the community model (Figure 8), and as a result, neither model could grow in the minimal medium, lacking the biomass precursors in Table 8. It was expected that their implicit roles would resume after incorporating the reactions suggested by the gap-filling framework.

Table 8. Missing Biomass Precursors in the *E. coli* Community

Ecoli 1	Ecoli 2
NAD	ATP
3-phosphoglycerate	NAD
Phosphoenoylpyruvate	Acetyl-CoA
Pyruvate	NADPH
Acetyl-CoA	alpha-ketoglutarate
NADPH	
Oxalolactate	
alpha-ketoglutarate	

The gap-filling results revealed four reactions to be added: one to Ecoli1 and three to Ecoli2. As shown in Figure 15, after including these reactions, growth was restored with Ecoli1 consuming glucose and excreting acetate, which was then used by Ecoli2. Despite deletions in glycolysis and pyruvate metabolism, the algorithm added F6P-phosphoketolase, which produces E4P and acetylphosphate.

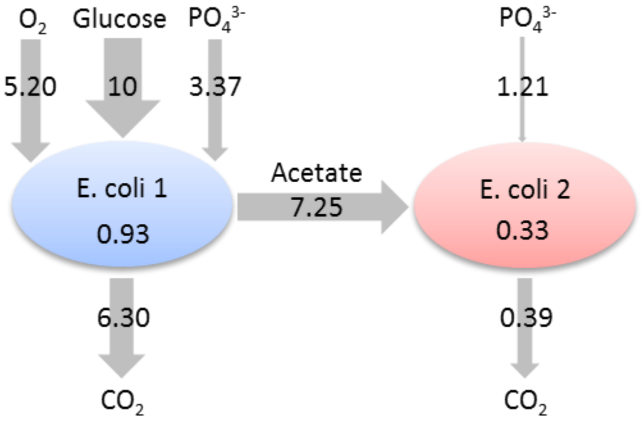

Figure 15. Growth Schematic of the Faux *E. coli* Community after Gap-Filling

The acetylphosphate was then converted into acetate by the acetate kinase reaction already in Ecoli1 (Figure 16). In addition, the reactions in Entner-Doudoroff Pathway were turned on in Ecoli1 such that pyruvate could be made from glucose-6-phosphate to sustain growth. In Ecoli 2, citrate oxaloacetate lyase was added to replace the role of citrate synthase, and pyruvate dehydrogenase reconnected acetyl-CoA with pyruvate, which then replenished those biomass precursors in the upper part of glycolysis. As a result, both models grew at sub-optimal rates, 0.93 and 0.33 mmol·gDW^{-1}·hr^{-1}, because Ecoli1 had a metabolic duty to fulfill, i.e., secreting acetate, and Ecoli 2 had to live with substrate limitation.

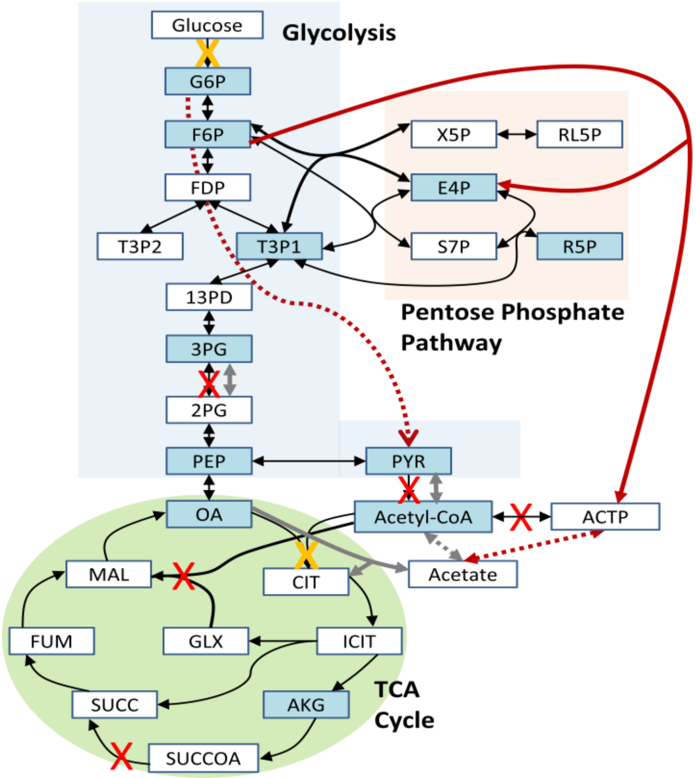

Figure 16. Gap-Filled *E. coli* Communiy Model. Red solid lines: reactions added to Ecoli1. Red dotted lines: reactions turned on in Ecoli1; Grey solid lines: reactions added to Ecoli2; Grey dotted lines: reactions turned on in Ecoli2; Blue boxes: biomass precursors

This example demonstrated the efficacy of the expanded gap-filling framework. It suggested that metabolic gaps in one organism could be filled by reactions in another via metabolite exchange. In particular, acetate secretion by the primary digester (Ecoli1) and acetate uptake by the downstream organism (Ecoli2) were not enforced as constraints during the gap-filling process; rather, the algorithm arrived at this decision such that it could use a minimum number of reactions to restore growth of both parties. It also showed how minimal changes to a gene-associated model could be maintained by adding the fewest possible reactions and turning on existing ones. As long as routes to synthesize highly connected metabolites are present, growth can be restored. It is noteworthy that if the optimal solutions are inadequate for the users, other solutions can be enumerated at the cost of adding more reactions; in other words, the solution pool might have contained the set of 12 reactions deleted initially during the creation of the *E. coli* community model. The computational method was shown to be useful in the curation of community model and in discovering possible metabolic interactions, and it was subsequently applied to the benzene metagenomic model.

6.2 Results and Discussions: Gap-Filling the Benzene-Degrading Metagenome-Scale Model

6.2.1 Description of the Gap-Filling Conditions

Before the decision tree shown in Section 3.7 was formed, the model exchanges of those metabolites not in the minimal medium were set to have flux bounds of [-1000, 1000]. This range suggested that for a given compound, it could be either secreted into or utilized from the culture medium by the model at a maximum rate of 1000. Same as the *E. coli* test case, the interspecies electron trafficking was not specified as a constraint, in hope that the gap-filling algorithm would capture the expected metabolic interactions upon meeting the requirement of positive growth rates in both bacteria and archaea model. However, although only 40 reactions were added in this trial, acetate and hydrogen were not detected as part of metabolite exchanges between the two domains. Instead, numerous exchanges of dipeptides, amino acids, and cofactors were found (Table C1 in Appendix C). Therefore, the bounds for model exchanges were adjusted to [-1000, 5] such that the organisms could consume at most 5 units of compounds that were not in the minimal medium. Also, the constraint for acetate and hydrogen exchange was enforced separately in subsequent trials by requiring at least 10 units of each being consumed by the archaea model.

6.2.2 Quantitative Results Generated by Community Gap-Filling

Table 9 shows the major carbon fluxes and exchanges in the scenario where acetate is specified as the key intermediate. The FBA results are reported for the top three solutions predicted by the gap-filling algorithm; the number of reactions added was 80, 80, and 82 in Solution 1 (Table C4), 2, and 3, respectively. In the FBA, the objective function was the sum of growth rates of the two models, and the specified benzoate degradation rate, 2.7 $\mathrm{mmol \cdot gDW^{-1} \cdot month^{-1}}$, was calculated from an actual benzene degradation of 0.2 $\mathrm{mg \cdot L^{-1} \cdot day^{-1}}$ and a cell concentration of 10^8 cells/mL. In addition, the lower bound of the archaea biomass equation was specified to be non-negative, and acetate, hydrogen, formate, CO_2, and methane were allowed to accumulate in the medium. Without the lower bound imposed on the growth of archaea, its biomass equation could not carry any flux. The flux values of other exchanges are listed in Table C2-C3 in Appendix C.

Immediately evident from the flux distribution in each solution was that benzoate was consumed only by the bacteria model and that methane was produced only by the archaea. Furthermore, the archaea used acetate secreted by the bacteria, as indicated by the acetate exchange fluxes in opposite signs. The growth rate of the bacteria was higher compared to that of archaea in all solutions. The community growth rate, 0.51 month^{-1}, was lower than the actual growth rate 0.69 month^{-1}, calculated from ln(2) divided by a doubling time of 30 days. Although unlikely, it was also shown that the archaea could excrete trehalose as a byproduct of growth. In terms of medium components, there was a marked difference in the uptake rate of thiamin and folate between the two models, and riboflavin was consumed by both. In addition, metal ions such as copper, cobalt, and magnesium were also utilized by both groups of organisms (Table C2-C3).

Table 9. Carbon Exchange Fluxes of the Gap-Filled Bacteria and Archaea Model with Acetate as the Key Intermediate

Exchange Fluxes (mmol·gDW^{-1}·month^{-1})	Bacteria			Archaea		
	Solution 1	Solution 2	Solution 3	Solution 1	Solution 2	Solution 3
Growth (month^{-1})	0.503	0.503	0.505	9.80E-03	9.80E-03	9.80E-03
Benzoate	2.7	2.7	2.7	0	0	0
Acetate	-1.42	-1.42	-1.43	1.42	1.42	1.43
Methane	0	0	0	-1.26	-1.26	-1.33
CO2	-0.524	-0.524	-0.533	-1.28	-1.28	-1.35
Thiamin	2.29E-03	2.29E-03	2.30E-03	0	0	0
Folate	6.88E-03	6.88E-03	6.90E-03	0	0	0
Trehalose	0	0	0	-2.44E-02	-2.44E-02	-1.96E-02

However, ratios between the carbon fluxes diverged significantly from the theoretical values. First, given the benzoate degradation rate of 2.7 mmol·gDW^{-1}·month^{-1}, the predicted growth yield of the community, 1.58 (g cell/g benzoate), was high compared to the theoretical yield of 0.07 g cells/g benzoate, calculated from an fs value of 0.05 (Calculation in Appendix E). In addition, benzoate was not stoichiometrically converted to acetate at the ratio of 1 mole benzene to 3.75 mole of acetate. A further examination on the flux distribution showed that given 1 mole of benzoate, although 3 moles of acetyl-CoA were indeed produced at the end of the benzoyl-CoA lower pathway, not all acetyl-CoAs were shuttled out as acetate to the archaea, as shown by the acetate exchange fluxes in Table 9. As a result, the flux ratio of methane to benzoate also deviated from the theoretical value. A separate analysis indicated that the majority of acetyl-CoA, although converted into acetate, eventually entered pentose phosphate pathway, suggesting that these carbons might have been used for the synthesis of biomass precursors instead. On the other hand, although in archaea acetate was stoichiometrically converted to methane and carbon dioxide, the growth yield on acetate, 0.12 (g cell/g acetate), also exceeded the theoretical value of 0.04 (g cell/g acetate). These observations suggest that the model is not ready for any quantitative predictions about the microbial community, although metabolic gaps are now resolved.

Table 10. Carbon Exchange Fluxes of the Gap-Filled Bacteria and Archaea Model with Hydrogen as the Key Intermediate

Exchange Fluxes (mmol·gDW^{-1}·month^{-1})	Bacteria			Archaea		
	Solution 1	Solution 2	Solution 3	Solution 1	Solution 2	Solution 3
Growth (month^{-1})	0.326	0.326	0.324	0.0322	0.0322	0.497
Benzoate	2.7	2.7	2.7	0	0	0
Formate	-2.7	-2.7	-2.7	2.7	2.7	2.7
H2	-3.03	-3.03	-3.24	3.03	3.03	3.24
Methane	0	0	0	-1.37	-1.37	-3.36
Thiamin	1.51E-03	1.51E-03	1.48E-03	0	0	0
Folate	4.53E-03	4.53E-03	4.43E-03	0	0	0
H2O2	0	0	0	-2.54	-2.54	-8.10

In the scenario of hydrogen transfer (Table 10), the optimal solution added 68 reactions (Table C5) to the model. The lower bound on the archaea biomass equation was still enforced. The predicted yield of 1.12 (g cell/g benzoate) was still high compared to the theoretical value of 0.07

(g cell/g benzoate). Similar to the acetate scenario, complications reported for carbon ratios were present as well. Moreover, an initial simulation showed that hydrogen could accumulate in the medium; in a syntrophic association, the methanogens should consume all the hydrogen produced by the benzene degraders such that the partial pressure of hydrogen is kept low in the community. However, the model did not reflect this phenomenon, so hydrogen accumulation was constrained to be zero in all the solutions reported in Table 10. Other constraints remain the same as the previous scenario. Nonetheless, although hydrogen was uptaken by the archaea model in all solutions, it contributed to the growth of archaea only in the third solution, resulting in an unrealistically growth rate of the archaea at 0.50 month^{-1}. Given the hydrogen and formate uptake rates in Solution 3 and using fs values of 0.26 eeq cells/eeq H_2 and 0.25 eeq cells/eeq formate, the theoretical growth of the methanogens should be no more than $1.70 \cdot 10^{-2}$ month^{-1}. The growth of methanogens in Solution 1 and 2 were supported solely by formate, with which the theoretical growth rate was calculated to be $7.5 \cdot 10^{-3}$ month^{-1}. However, formate production by the benzene degrader was not enforced during gap-filling; instead, it was an event that came with the gap-filled model when an explicit carbon source for the archaea was not specified during the gap-filling process. This is interesting as both formate and hydrogen are known substrates to methanogenesis. Also in Solutions 1 and 2, the ratio of growth rate of bacteria to archaea was 10, and this value was in line with the population counts based on total DNA reads: 90% of bacteria and 8.7% of archaea. It was shown by visual examination under microscopy that the bacteria dominating the methanogenic culture were smaller in size than the archaea (Devine, 2013), which may explain the faster growth rate of bacteria. Finally, the medium constituents consumed are listed in Table C3, and again, thiamin and folate were consumed only by the bacteria. The production of hydrogen peroxide in the archaea was considered an artifact of the gap-filling process (see Discussion).

The results show that including the non-native reactions proposed by the gap-filling algorithm leads to biomass production in both domains of the benzene-degrading community. From the flux balance analysis, one can also see how the medium components contribute to the growth of the two populations. The general interactions between the organisms and with the extracellular environment could be captured, although not at a quantitative level. The carbon imbalance can be attributed to the presence of thermodynamically inconsistent reactions in the models; these reactions form cycles and enable the unlimited production of ATP (Price et al., 2002), leading to most carbons being directed toward biomass production while the model can freely access the ATPs generated by cycles. This phenomenon was observed in both models; upon the maximization of ATP maintenance in all solutions, the flux of this reaction approached 1000 $\mathrm{mmol \cdot gDW^{-1} \cdot month^{-1}}$, the arbitrarily set upper bound. However, in reality the ATP maintenance flux never reaches its maximum because the amount of energy available to the cells is always limited. The stoichiometric imbalance can be corrected if cycles are removed (see Discussion).

6.2.3 Qualitative Hypotheses Generated by Community Gap-Filling

In the acetate scenario where 80 reactions were added to model (i.e. Solution 1), the metabolic gaps in the bacteria and archaea model were filled by 30 and 50 reactions, respectively (Table C4). There was also an overlap with the results from individual gap-filling; in total 30 reactions were suggested previously when each model was gap-filled independently. Since the metabolic gaps remained unchanged, it was assumed that these reactions served to restore the connectivity of the same metabolites for which they were previously responsible. Most missing biomass precursors were fixed by the inclusion of benzoate transport to the bacteria bin and acetate uptake by the archaea bin. Gaps in the biosynthesis of amino acids, cell wall, lipids, and cofactors were resolved accordingly. Interestingly, 37 reactions, not seen previously, were transport or exchange reactions related to 9 compounds listed in Figure 17. Aside from acetate being transferred from the bacteria to archaea (omitted in Figure 17 due to larger flux and similar magnitudes in all solutions), these metabolites also contributed to the growth of the methanogens model in trace amount. Among these metabolites, pyridoxal, a compound in vitamin B6 metabolism, had the greatest flux values in all three solutions shown, followed by the exchange of xanthine, AMP, methionine, phenylalanine, and CoA.

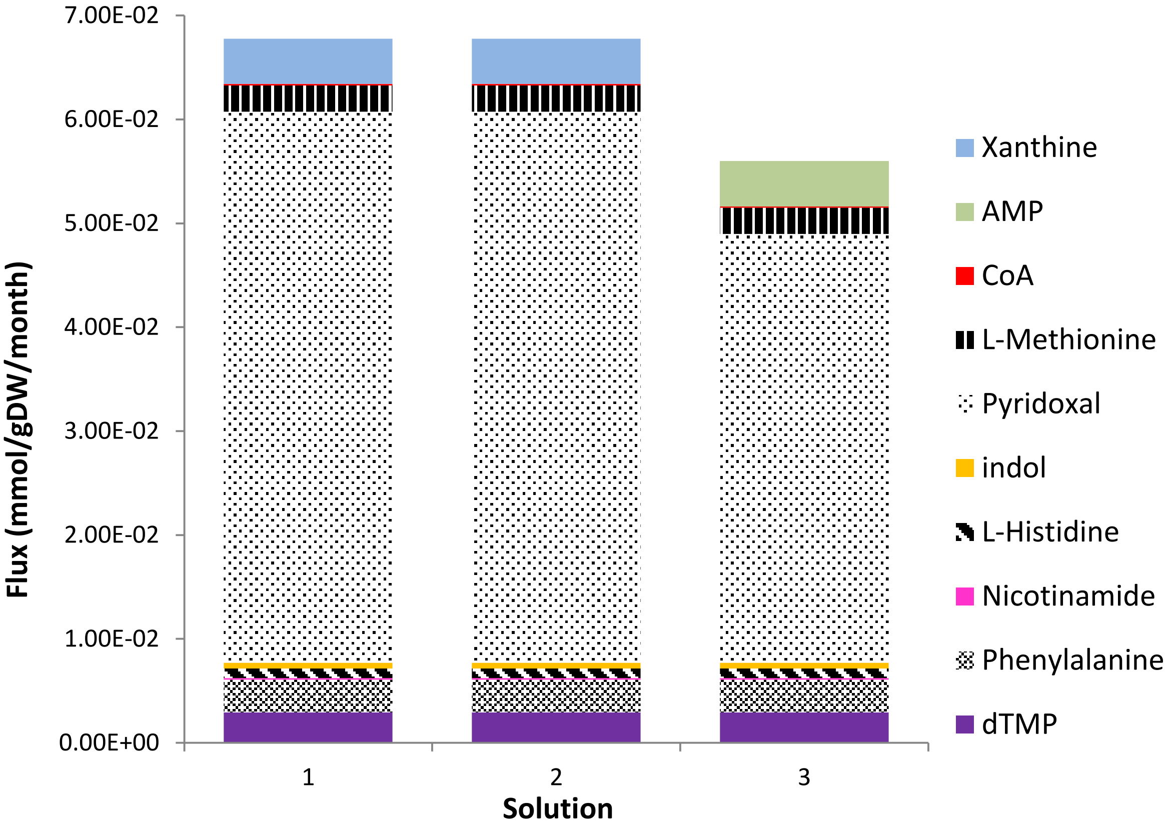

Figure 17. Flux Distribution of Putative Metabolites Transferred from the Bacteria to Archaea in Benzene-Degrading Community with Acetate as the Interspecies Intermediate

When hydrogen or formate was the key intermediate, 18 and 50 reactions were added to the bacteria and archaea bin, respectively (Table C5). Among these 68 reactions, 39 were shared by the reaction profiles generated in the individual gap-filling process. Based on their functional categories, the other 29 reactions also enabled the production of missing precursors. It is noteworthy that even with the same gap-filling constraints and parameters, the framework may suggest different reaction profiles due to the ways solutions are searched in the CPLEX Optimizer. Hence it is common that distinctive reactions may mend the same metabolic gap. Nonetheless, while intracellular metabolic gaps were resolved within each model, similar to the acetate scenario, the bacteria model also incorporated transport reactions to excrete coenzyme A and histidine that facilitated the growth of archaea (Figure 18). Interestingly, the metabolite trafficking was always uni-directional in both secretion scenarios, from the primary degrader to the downstream organism. This phenomenon has been shown to be a common in syntrophic associations (Wintermute and Silver, 2010; Freilich et al., 2011). The hypothesis that the

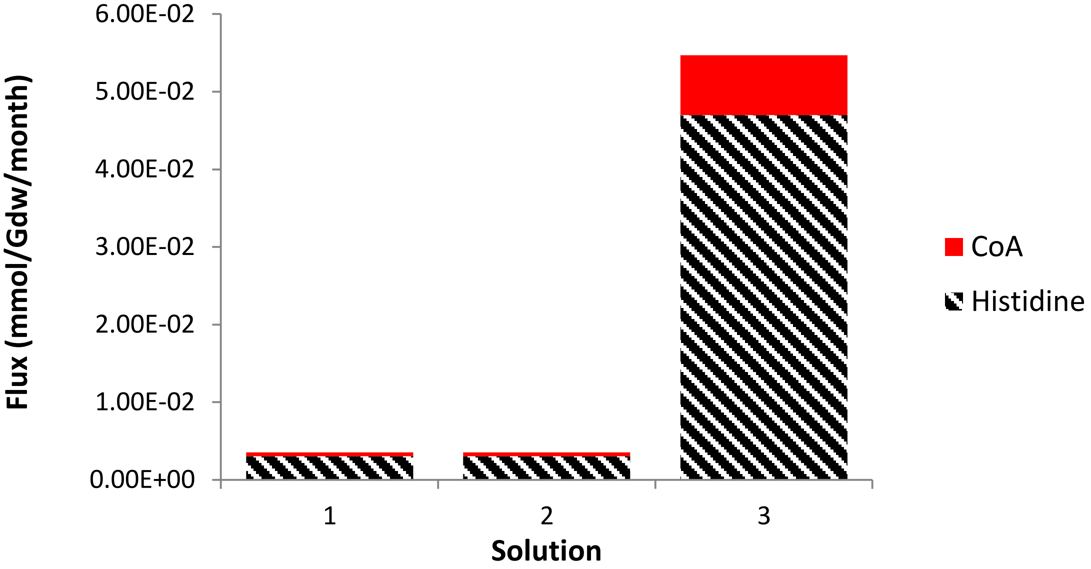

Figure 18. Flux Distribution of Putative Metabolites Transferred from the Bacteria to Archaea with Hydrogen and Formate as the Main Interspecies Intermediates

benzene degraders and methanogens may interact via the exchange of acetate, hydrogen, formate, CoA, and histidine therefore is generated, because the deletion of related interactions results in no cellular growth in the community model.

6.2.3 Discussion: Community Model and Potential Metabolic Interactions

Community Gap-Filling in Comparison with Other Community Modeling Programs

The community gap-filling algorithm developed here is the first known framework that resolves metabolic gaps at a community level. Moreover, it was applied to a ‘metagenome-scale’ model. The computational tool is comparable to OptCom, a method to model microbial communities (Zomorrodi and Maranas, 2012). Given the model of interactions, the OptCom framework is able to predict interspecies flux distribution and generates hypotheses about ecology when experimental data are provided. This also explains why well-understood microbial communities were employed in demonstrating the capabilities of OptCom. However, the purpose of community gap-filling is different because it takes advantage of the incompleteness of draft models and tries to describe why the microorganisms cannot live in isolation, by enumerating all the possible mechanisms via which they interact. At the same time, the framework is also different from gap-filling programs such as GrowMatch and GapFill since it deals with the growth of multiple organisms.

A Survey on the Transferred Compounds

According to the outcome of community gap-filling, several metabolites were proposed to be exported by the bacteria and subsequently consumed by the archaea species in the mixed culture.

Coenzyme A, a crucial cofactor for many cellular enzymes, was among them. The cofactor participates in the central carbon metabolism and fatty acid synthesis (Genschel, 2004) and is made exclusively from pantothenate, a precursor uptaken by all bacteria species (Leonardi, 2005). CoA biosynthesis is thought to be present in certain archaeal genomes and is important for carrying the acyl group in acetate for acetoclastic methanogens. Xanthine, on the other hand, does not seem to play any role in methanogenesis; instead, it takes part in purine metabolism, and its production and transport are regulated by purines such as guanine (Nudler and Mironov, 2004), so it may contribute to the methanogens in this regard. Methionine, on the other hand, has been related as a stimulus of methanogenesis in estuarine sediments due to the elevated production of methane in methionine-supplemented samples (Oremland and Polcin, 1982); it is therefore postulated that methionine may contribute as an intermediate in the pathway of methane formation.

The exchange flux of pyridoxal, a derivative of vitamin B6, carried the largest flux among others because it was used to make pyridoxal 5-phosphate, a biomass precursor in the archaea model. The inclusion of this component in the archaeal biomass equation perhaps is not without a reason; in fact, in *Methanosarcina* and *Methanospirillum*, serine hydroxymethyltransferase, which delivers carbons to purine biosynthesis, was shown to be dependent on pyridoxal phosphate (Ferry, 1993; Lin and Sparling, 1998). The archaea model obtained indole from bacteria to produce tryptophan, also a biomass precursor. Indole is only one reaction away from tryptophan, and this reaction is supported by metagenomic sequences. Histidine and phenylalanine are amino acids required for the production of biomass, and instances of interspecies amino acids exchange have been demonstrated only in yeasts (Klitgord and Segrè, 2010). The synthesis pathway of histidine, although present, is quite fragmented in the archaea bin, which may explain the interspecies transfer because completing the histidine synthesis pathway may entail making significant changes to the model. On the other hand, while nicotinamide or nicotinate transport might be a potential addition to the model since nicotinic acid was provided in the mineral medium, it was unlikely that it be transferred from the bacteria to archaea because consuming it from the medium was more direct. Nicotinamide was used by the archaea bin to synthesize NAD and NADP. Finally, the dTMP uptaken was used to make dTTP in the archaea model in the metabolism of pyrimidine.

The metagenome model indeed contained gene-associated reactions to synthesize or utilize these compounds because the corresponding anabolic and catabolic pathways were not added by the gap-filling framework. It can be argued that the bacteria supply these substances for the methanogens such that benzene degradation and hence the benzene degraders can benefit from the removal of fermentation products by other organisms in the culture. However, whether the bacteria have the ability to export these metabolites or if the archaea can uptake them with various modes of transport is still questioned. To elucidate these metabolic capabilities in the model, the metagenome needs to be bioinformatically mined for transport proteins.

The Hydrogen Scenario and Its Metabolic Implications

The methanogens constitute the hydrogen-consuming population in the benzene-degrading community. Because the hydrogen partial pressure is kept low by these organisms, benzene decomposition is able to proceed (Rakoczy et al, 2011). From the results of community gap-filling, although not required as a constraint, formate was produced by the bacteria and subsequently metabolized by the methanogens using gene-associated reactions. This points to the possibility that formate may also be a key intermediate that facilitates anaerobic benzene degradation in the OR community. Formate is associated particularly with the methanogens without cytochromes (Thauer et al., 2008) and is catabolised also in hydrogenotrophic methanogenesis. Although hydrogen alone should support the methanogens, formate may contribute to the syntrophic growth as well.

Filling the gaps in this scenario, however, led to the accumulation of hydrogen peroxide in the medium, which was the result of adding the cytochrome peroxidase reaction. Why the algorithm decided to incorporate this reaction was traced back as follows. In the model, cytochrome peroxidase produced hydrogen peroxide upon the reduction of cytochrome c. The oxidized cytochrome c consumed by in this reaction originated from other reactions that were coupled to ubiquinone reduction. The ubiquinone here was produced by succinate dehydrogenase (FADH2 + ubiquinone <=> FAD + ubiquinol) and pyruvate oxidase (pyruvate + ubiquinone + H2O <=> acetate + CO2 + ubiquinol). In the case of pyruvate oxidase, this reaction served to channel carbons to growth-associated processes, but since methanogens do not have ubiquinone (Thauer et al., 1977), pyruvate synthase instead (pyruvate + 2 oxidized ferredoxin + CoA <=> acetyl-CoA + 2 reduced ferredoxin + CO2 + H^+) served as the major anabolic enzyme to direct carbons in some methanogens (Shieh and Whitman, 1987; Ladopo and Whitman, 1990; Furdui and Ragsdale, 2000). Pyruvate oxidase might have been mis-annotated since in archaea reaction of this type is mediated by ferredoxin rather than ubiquinone. In the models of *Methanosarcina barkeri* and *M. acetivorans* (Feist et al., 2006; Kumar et al., 2011), only pyruvate synthase was included. Similarly, FAD reduction can perhaps be mediated by ferredoxin instead. However, FAD and FADH2 were not included nor accounted for in these published models, although both compounds are represented in methanogens. In summary, the hydrogen peroxide production is a result of the model needing a route to regenerate electron carriers used extensively in the central metabolism, but this carrier, ubiquinone, is a product of mis-annotation.

Refining the community model

The complications that influence the quality of genome-scale models impact community models likewise. For example, the reversibility of reactions can affect the flux distribution significantly (Reed et al., 2006). The reversibility and substrate specificity of the metagenomic model were extracted from SEED, and validating the information may require an automated approach due to the scale of these models. In addition, it is necessary to verify the authenticity of metabolic gaps; in this case, a gap may result from mis-annotation, biases in selecting reactions, unannotated ORFs, and finally true metabolic incapacity. As mentioned previously, a substantial part of the benzene metagenome was not functionally classified. On top of the missing reactions, one must also keep in mind that regulatory mechanisms are not specified in such models, so the flux distribution is the result of all reactions being active. Finally, although growth rate is now a well-accepted biological objective for genome-scale models, it may not be for a community with a specific function such as benzene degradation. In the syntrophy study performed by Taffs et al. (2009), a complex consortium that alters its metabolic pattern according to the availability of light was analyzed using representative genome scale models, all without objective functions. Instead, they simply enumerated all the possible interactions among the microbes by only stoichiometry and steady state constraints. This approach can potentially be applied to the benzene-degrading culture as it bypasses the need for a biological objective. Nevertheless, the fitness indicator of the benzene-degrading community should be explored; in the meantime, if the growth of both archaea and bacteria are to serve as the objective function, a weighted sum can be given to account for the relative abundance of these community members.

7.0 Conclusions and Recommendations for Future Work

A metagenome model was constructed based on the sequences derived from a methanogenic benzene-degrading community. First, the DNA reads of the metagenome were assembled with strategies to overcome the marked difference in lengths, and it was found that pre-assembling the short reads could improve the overall lengths of the resulting contigs. However, other than searching for the sequences of known gene in the contigs, a method to address the quality of the selected assembly should be applied. The metagenome was taxonomically organized, and the confidence level of the organization was increased by creating bins at the domain level rather than genus. Based on the binning information, the presence of both acetoclastic and hydrogenotrophic methanogens was supported. The reconstructed draft models revealed metabolic gaps in amino acids and cofactors, but most gaps could be resolved by the addition of necessary transport reactions. In addition, the ability of the bacteria population to produce formate, acetate, and hydrogen were supported by the fact that no synthesis pathways but only export reactions of these compounds were added by the gap-filling framework. The likelihood that the added reactions actually belong to the models was probed by a comparison with the published models of similar organisms and examining the reaction neighborhood. For the remainders not shared or classified under any functional subsystem, such as transport reactions, genomic and experimental evidence are needed for their validations. The gap-filling framework was expanded to accommodate community-wide metabolic gaps, and its capability was demonstrated via a synthetic *E. coli* community. The same framework then suggested putative metabolic interactions between the bacteria and archaea in the benzene-degrading community. Specifically, it hypothesized the uni-directional transfer of CoA and histidine, among other metabolites. In addition, it postulated that formate might also be a key intermediate in the community. The results also highlighted the effects of mis-annotation and the importance of thermodynamic constraints, necessitating model refinement and re-annotation.

Recommendations for Community Modeling and Correcting Stoichiometric Imbalance

After the community gap-filling, many quantitative discrepancies between the model prediction and experimental data were found. The high growth yields and low fluxes of CO_2 and methane evolution can be attributed to the fact that thermodynamic constraints are not in place yet, both in terms of growth yields and reaction reversibilities. In addition, the reconstructed network contains a significant number of redundant reactions, supposedly from a variety of microbes present in the community. Figure A4 shows the number of reactions in the bacteria model discussed in this thesis and other models created from the new Illumina data. As a result, the bacteria model can preferentially utilize reactions originated from any bacteria species, not just those of the benzene degraders. Moreover, redundant reactions can form thermodynamically inconsistent cycles such as those described by Price et al. (2002); these cycles can lead to the unlimited production of ATP in the model. As shown in Appendix D, the presence of cycles is confirmed in both bacteria and archaea models. When ATP becomes a free resource, benzoate consumed is inevitably channelled to biomass synthesis rather than energy generation via the export of fermentation products, causing the predicted high growth yields and low levels of acetate and formate secretion. However, this issue can be resolved by implementing thermodynamic constraints and removing the reaction cycles, such that benzoate can be channeled to ATP synthesis. Using the algorithm REMI, an attempt to correct thermodynamic inconsistency by removing cycles was made before the models were gap-filled individually

(Appendix D), but it became immediately evident that the task was counterproductive because the reactions eliminated may later be added by the gap-filling algorithm for growth. However, with the community growth restored now, REMI algorithm can be applied to the metagenome model to remove redundant reactions that are generating free ATPs, thus correcting the carbon imbalance.

Alternatively, community gap-filling can be applied to the models created from the newly sequenced metagenome of the methanogenic benzene-degrading community. The assembly from this sequencing event was binned based mainly on read depth, and draft models representing OTU OR-M2 and organisms with high and low abundance in the community were generated (Figure A4). In terms of the total number of reactions, the OR-M2 model is highly condensed compared to the bacteria model used in this thesis as it represents the core set of reactions used by the primary benzene degraders. In addition, a model most likely associated with *Methanoregula* was produced from the Illumina assembly. With these new models, community gap-filling may be performed with greater confidence since the assembly was based on a single DNA sample, binning was performed using an unsupervised approach, and most reactions were derived from the target organisms.

General Recommendations

To improve any models, iterations and refinements are key. For the metagenome models presented in this thesis, the refinement process can begin at network reconstruction. For instance, many gene products were not incorporated or identified by the annotation platform, and a bi-directional sequence search may help with the discovery of more reactions or correction of existing ones, such as the pyruvate oxidase mentioned previously. It is also a more transparent approach as the user will be familiar with the criteria based on which a reaction is derived. Certainly the list of reactions suggested by the gap-filling algorithm needs to be validated bioinformatically, especially for the transport proteins. On the other hand, gap-filling simulation can be run for simultaneous exchange of acetate, hydrogen, and formate. Most importantly, the thermodynamic constraints in the benzene-degrading culture must be reflected in the community model. Specifically, energy-yielding, excess reaction cycles can be identified and removed using publicly available methods or that in Appendix D; however, *in silico* growth must be maintained, and so are reactions with strong genomic evidence. The discrepancy between the predictions and the real outcome of experiments always reveals the way to develop a model, which then generates more hypotheses to be validated. It is in this loop that a model exists and improves. While precautions must be taken in interpreting the results, modeling is still a powerful tool for discovering and formulating ideas.

References

Abegunde, Tejumoluwa. (2010). Comparison of DNA sequence assembly algorithms using mixed data sources. Retrieved from University of Saskatchewan Electronic Theses and Dissertations. http://ecommons.usask.ca/handle/10388/etd-04142010-133341.

Abu Laban, N., Selesi, D., Jobelius, C., & Meckenstock, R. U. (2009). Anaerobic benzene degradation by Gram-positive sulfate-reducing bacteria. *FEMS microbiology ecology*, *68*(3), 300–311. doi:10.1111/j.1574-6941.2009.00672.x

Anderson, I., Ulrich, L. E., Lupa, B., Susanti, D., Porat, I., Hooper, S. D., Lykidis, A., et al. (2009). Genomic characterization of methanomicrobiales reveals three classes of methanogens. *PloS one*, *4*(6), e5797. doi:10.1371/journal.pone.0005797

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC genomics*, *9*, 75. doi:10.1186/1471-2164-9-75

Bas M Van Der Zaan, F. T. S. A. J. M. S. C. M. P. W. M. D. V. H. S. A. A. M. L. J. G. (2012). Anaerobic benzene degradation under denitrifying conditions: Peptococcaceae as dominant benzene degraders and evidence for a syntrophic process. *Environmental Microbiology*, *14*(5), 1171 – 1181. Retrieved from http://resolver.scholarsportal.info/resolve/14622912/v14i0005/1171_abdudcaefasp.xml

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic acids research*, *41*(Database issue), D36–42. doi:10.1093/nar/gks1195

Borenstein, E. (2012). Computational systems biology and in silico modeling of the human microbiome. *Briefings in bioinformatics*, *13*(6), 769–80. doi:10.1093/bib/bbs022

Branco Dos Santos, F., De Vos, W. M., & Teusink, B. (2012). Towards metagenome-scale models for industrial applications-the case of Lactic Acid Bacteria. *Current opinion in biotechnology*. doi:10.1016/j.copbio.2012.11.003

Cakir, T., Efe, C., Dikicioglu, D., Hortaçsu, A., Kirdar, B., & Oliver, S. G. (2007). Flux balance analysis of a genome-scale yeast model constrained by exometabolomic data allows metabolic system identification of genetically different strains. *Biotechnology progress*, *23*(2), 320–6. doi:10.1021/bp060272r

Caldwell, M. E., & Suflita, J. M. (2000). Detection of Phenol and Benzoate as Intermediates of Anaerobic Benzene Biodegradation under Different Terminal Electron-Accepting Conditions. *Environmental Science & Technology*, *34*(7), 1216–1220. doi:10.1021/es990849j

Carmona, M., Zamarro, M. T., Blázquez, B., Durante-Rodríguez, G., Juárez, J. F., Valderrama, J. A., Barragán, M. J. L., et al. (2009). Anaerobic catabolism of aromatic compounds: a genetic and genomic view. *Microbiology and molecular biology reviews : MMBR*, *73*(1), 71–133. doi:10.1128/MMBR.00021-08

Caspi, R., Foerster, H., Fulcher, C. A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., et al. (2008). The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic acids research*, *36*(Database issue), D623–31. doi:10.1093/nar/gkm900

Chain, Patrick. (2011). *Metagenomic assembly challenges and tools* [PowerPoint Slides]. Retrieved from Metagenomics Methods and Data Analysis Workshop, Ontario Genomics Institute. http://www.ontariogenomics.ca/event/2011-10-11/711

Chaisson, M. J., & Pevzner, P. A. (2008). Short read fragment assembly of bacterial genomes. *Genome research*, *18*(2), 324–30. doi:10.1101/gr.7088808

Chang, W., Um, Y., & Pulliam Holoman, T. R. (2005). Molecular characterization of anaerobic microbial communities from benzene-degrading sediments under methanogenic conditions. *Biotechnology progress*, *21*(6), 1789–94. doi:10.1021/bp050250p

Chevreux, B., Wetter, T. and Suhai, S. (1999): Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB)* 99, pp. 45-56.

Coates, J. D., Chakraborty, R., Lack, J. G., O'Connor, S. M., Cole, K. A., Bender, K. S., & Achenbach, L. A. (2001). Anaerobic benzene oxidation coupled to nitrate reduction in pure culture by two strains of Dechloromonas. *Nature*, *411*(6841), 1039–43. doi:10.1038/35082545

Devine, C. E., Gitiafroz, R., and Edwards, E.A. (2011). Metabolic pathways, genes and enzymes in anaerobic benzene-degrading cultures: from "omics" to application, Batelle International Symposium on Bioremediation and SustainableTechnologies. Reno, NV.

Devine, C. E. (2013). Identification of key organisms, genes and pathways in benzene-degrading methanogenic cultures. In prep. University of Toronto, ON

Durot, M., Bourguignon, P.-Y., & Schachter, V. (2009). Genome-scale models of bacterial metabolism: reconstruction and applications. (V. De Lorenzo, Ed.)*FEMS Microbiology Reviews*, *33*(1), 164–190. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2704943&tool=pmcentrez&rendertype=abstract

Edwards, J., & Palsson, B. (2000). The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proceedings of the National Academy of*

Sciences of the United States of America, *97(10)*, 5528–5533. Retrieved from http://www.pnas.org/content/97/10/5528.short

Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L., & Palsson, B. Ø. (2009). Reconstruction of biochemical networks in microorganisms. *Nature reviews. Microbiology*, *7*(2), 129–43. doi:10.1038/nrmicro1949

Feist, A. M., Scholten, J. C. M., Palsson, B. Ø., Brockman, F. J., & Ideker, T. (2006). Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri. *Molecular systems biology*, *2*, 2006.0004. doi:10.1038/msb4100046

Ferry, J.G. (1993). Methanogenesis: Ecology, Physiology, Biochemistry & Genetics (Ed.). London: Chapman and Hall

Freilich, S., Zarecki, R., Eilam, O., Segal, E. S., Henry, C. S., Kupiec, M., Gophna, U., et al. (2011). Competitive and cooperative metabolic interactions in bacterial communities. *Nature communications*, *2*, 589. doi:10.1038/ncomms1597

Furdui, C., & Ragsdale, S. W. (2000). The role of pyruvate ferredoxin oxidoreductase in pyruvate synthesis during autotrophic growth by the Wood-Ljungdahl pathway. *The Journal of biological chemistry*, *275*(37), 28494–9. doi:10.1074/jbc.M003291200

Genschel, U. (2004). Coenzyme A biosynthesis: reconstruction of the pathway in archaea and an evolutionary scenario based on comparative genomics. *Molecular biology and evolution*, *21*(7), 1242–51. doi:10.1093/molbev/msh119

Goldberg, S. M. D., Johnson, J., Busam, D., Feldblyum, T., Ferriera, S., Friedman, R., Halpern, A., et al. (2006). A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(30), 11240–5. doi:10.1073/pnas.0604351103

Gordon, D., Abajian, C., & Green, P. (1998). Consed: a graphical tool for sequence finishing. *Genome research*, *8*(3), 195–202. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9521923

Health Canada. (June, 2009). Environmental and Workplace Health: Benzene Guideline Technical Document: Guidelines for Canadian Drinking Water Quality. Retrieved from http://www.hc-sc.gc.ca/ewh-semt/pubs/water-eau/benzene/index-eng.php

Holmes, D. E., Risso, C., Smith, J. A., & Lovley, D. R. (2011). Anaerobic oxidation of benzene by the hyperthermophilic archaeon Ferroglobus placidus. *Applied and environmental microbiology*, *77*(17), 5926–33. doi:10.1128/AEM.05452-11

Kasai, Y., Takahata, Y., Manefield, M., & Watanabe, K. (2006). RNA-based stable isotope probing and isolation of anaerobic benzene-degrading bacteria from gasoline-contaminated groundwater. *Applied and environmental microbiology*, *72*(5), 3586–92. doi:10.1128/AEM.72.5.3586-3592.2006

Kleinsteuber, S., Schleinitz, K. M., Breitfeld, J., Harms, H., Richnow, H. H., & Vogt, C. (2008). Molecular characterization of bacterial communities mineralizing benzene under sulfate-reducing conditions. *FEMS microbiology ecology*, *66*(1), 143–57. doi:10.1111/j.1574-6941.2008.00536.x

Klitgord, N., & Segrè, D. (2010). Environments that induce synthetic microbial ecosystems. *PLoS computational biology*, *6*(11), e1001002. doi:10.1371/journal.pcbi.1001002

Kumar, VS., Dasika, M. S., & Maranas, C. D. (2007). Optimization based automated curation of metabolic reconstructions. BMC bioinformatics, 8(1), 212. doi:10.1186/1471-2105-8-212

Kumar, VS., Ferry, J. G., & Maranas, C. D. (2011). Metabolic reconstruction of the archaeon methanogen Methanosarcina Acetivorans. BMC systems biology, 5(1), 28. doi:10.1186/1752-0509-5-28

Kumar, V. S., & Maranas, C. D. (2009). GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. (C. A. Ouzounis, Ed.)*PLoS computational biology*, *5*(3), e1000308. doi:10.1371/journal.pcbi.1000308

Kunapuli, U., Lueders, T., & Meckenstock, R. U. (2007). The use of stable isotope probing to identify key iron-reducing microorganisms involved in anaerobic benzene degradation. *The ISME journal*, *1*(7), 643–53. doi:10.1038/ismej.2007.73

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., & Hugenholtz, P. (2008). A bioinformatician's guide to metagenomics. *Microbiology and molecular biology reviews : MMBR*, *72*(4), 557–78, Table of Contents. doi:10.1128/MMBR.00009-08

Laserson, J., Jojic, V., & Koller, D. (2011). Genovo: de novo assembly for metagenomes. *Journal of computational biology : a journal of computational molecular cell biology*, *18*(3), 429–43. doi:10.1089/cmb.2010.0244

Leonardi, R., Zhang, Y.-M., Rock, C. O., & Jackowski, S. (n.d.). Coenzyme A: back in action. *Progress in lipid research*, *44*(2-3), 125–53. doi:10.1016/j.plipres.2005.04.001

Lin, Z., & Sparling, R. (1998). Investigation of serine hydroxymethyltransferase in methanogens. *Canadian journal of microbiology*, *44*(7), 652–6. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9783425

Lovley, D., Coates, J., Woodward, J., & Phillips, E. (1995). Benzene Oxidation Coupled to Sulfate Reduction. *Appl. Envir. Microbiol.*, *61*(3), 953–958. Retrieved from http://aem.asm.org/content/61/3/953.short

MacDonald, N. J., Parks, D. H., & Beiko, R. G. (2012). Rapid identification of high-confidence taxonomic assignments for metagenomic data. *Nucleic acids research*, *40*(14), e111. doi:10.1093/nar/gks335

Mahadevan, R., Bond, D. R., Butler, J. E., Esteve-Nuñez, A., Coppi, M. V, Palsson, B. O., Schilling, C. H., et al. (2006). Characterization of metabolism in the Fe(III)-reducing organism Geobacter sulfurreducens by constraint-based modeling. *Applied and environmental microbiology*, *72*(2), 1558–68. doi:10.1128/AEM.72.2.1558-1568.2006

Markowitz, V. M., Chen, I.-M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., et al. (2012). IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic acids research*, *40*(Database issue), D115–22. doi:10.1093/nar/gkr1044

Masumoto, H., Kurisu, F., Kasuga, I., Tourlousse, D. M., & Furumai, H. (2012). Complete mineralization of benzene by a methanogenic enrichment culture and effect of putative metabolites on the degradation. *Chemosphere*, *86*(8), 822–8. doi:10.1016/j.chemosphere.2011.11.051

Meyer, E., Aglyamova, G. V, Wang, S., Buchanan-Carter, J., Abrego, D., Colbourne, J. K., Willis, B. L., et al. (2009). Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. *BMC genomics*, *10*(1), 219. doi:10.1186/1471-2164-10-219

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., et al. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, *9*(1), 386. doi:10.1186/1471-2105-9-386

Musat, F., & Widdel, F. (2008). Anaerobic degradation of benzene by a marine sulfate-reducing enrichment culture, and cell hybridization of the dominant phylotype. *Environmental microbiology*, *10*(1), 10–9. doi:10.1111/j.1462-2920.2007.01425.x

Nales, M., Butler, B. J., & Edwards, E. a. (1998). Anaerobic Benzene Biodegradation: A Microcosm Survey. *Bioremediation Journal*, *2*(2), 125–144. doi:10.1080/10889869891214268

Narzisi, G., & Mishra, B. (2011). Comparing de novo genome assembly: the long and short of it. *PloS one*, *6*(4), e19175. doi:10.1371/journal.pone.0019175

Nath, K., & Das, D. (2004). Improvement of fermentative hydrogen production: various approaches. *Applied microbiology and biotechnology*, *65*(5), 520–9. doi:10.1007/s00253-004-1644-0

Ning Z, Cox AJ and Mullikin JC. (2001). SSAHA: a fast search method for large DNA databases. *Genome research* 11;10;1725-9. DOI: 10.1101/gr.194201

Nudler, E., & Mironov, A. S. (2004). The riboswitch control of bacterial metabolism. *Trends in biochemical sciences*, *29*(1), 11–7. doi:10.1016/j.tibs.2003.11.004

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research*, *27*(1), 29–34. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=148090&tool=pmcentrez&rendertype=abstract

Oka, A. R., Phelps, C. D., McGuinness, L. M., Mumford, A., Young, L. Y., & Kerkhof, L. J. (2008). Identification of critical members in a sulfidogenic benzene-degrading consortium by DNA stable isotope probing. *Applied and environmental microbiology*, *74*(20), 6476–80. doi:10.1128/AEM.01082-08

Oremland, R. S., & Polcin, S. (1982). Methanogenesis and sulfate reduction: competitive and noncompetitive substrates in estuarine sediments. *Applied and environmental microbiology*, *44*(6), 1270–6. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=242184&tool=pmcentrez&rendertype=abstract

Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? *Nature biotechnology*, *28*(3), 245–8. doi:10.1038/nbt.1614

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V, Chuang, H.-Y., Cohoon, M., De Crécy-Lagard, V., et al. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic acids research*, *33*(17), 5691–702. doi:10.1093/nar/gki866

Park, J. H., Lee, K. H., Kim, T. Y., & Lee, S. Y. (2007). Metabolic engineering of Escherichia coli for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(19), 7797–7802. doi:10.1073/pnas.0702609104

Phelps, C. D., Kerkhof, L. J., & Young, L. Y. (1998). Molecular characterization of a sulfate-reducing consortium which mineralizes benzene. *FEMS Microbiology Ecology*, *27*(3), 269–279. doi:10.1111/j.1574-6941.1998.tb00543.x

Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Briefings in bioinformatics*, *10*(4), 354–66. doi:10.1093/bib/bbp026

Price, N. D., Famili, I., Beard, D. A., & Palsson, B. Ø. (2002). Extreme pathways and Kirchhoff's second law. *Biophysical journal*, *83*(5), 2879–82. doi:10.1016/S0006-3495(02)75297-1

Price, N. D., Reed, J. L., & Palsson, B. Ø. (2004). Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature reviews. Microbiology*, *2*(11), 886–97. doi:10.1038/nrmicro1023

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, *464*(7285), 59–65. doi:10.1038/nature08821

Rakoczy, J., Schleinitz, K. M., Müller, N., Richnow, H. H., & Vogt, C. (2011). Effects of hydrogen and acetate on benzene mineralisation under sulphate-reducing conditions. FEMS microbiology ecology, 77(2), 238–47. doi:10.1111/j.1574-6941.2011.01101.x

Reed, J. L., Famili, I., Thiele, I., & Palsson, B. O. (2006). Towards multidimensional genome annotation. *Nature reviews. Genetics*, *7*(2), 130–41. doi:10.1038/nrg1769

Ren, Q., Kang, K. H., & Paulsen, I. T. (2004). TransportDB: a relational database of cellular membrane transport systems. *Nucleic acids research*, *32*(Database issue), D284–8. doi:10.1093/nar/gkh016

Rooney-Varga, J. N., Anderson, R. T., Fraga, J. L., Ringelberg, D., & Lovley, D. R. (1999). Microbial Communities Associated with Anaerobic Benzene Degradation in a Petroleum-Contaminated Aquifer. *Appl. Envir. Microbiol.*, *65*(7), 3056–3063. Retrieved from http://aem.asm.org/content/65/7/3056.short

Sakai, N., Kurisu, F., Yagi, O., Nakajima, F. ., & Yamamoto, K. (2009). Identification of putative benzene-degrading bacteria in methanogenic enrichment cultures. *Journal of Bioscience and Bioengineering*, *108*(6), 501 – 507. Retrieved from http://resolver.scholarsportal.info/resolve/13891723/v108i0006/501_iopbbimec.xml

Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., et al. (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nature protocols*, *6*(9), 1290–307. doi:10.1038/nprot.2011.308

Schomburg, I., Hofmann, O., Baensch, C., Chang, A., & Schomburg, D. (2000). Enzyme data and metabolic information: BRENDA, a resource for research in biology, biochemistry, and medicine. *Gene Function & Disease*, *1*(3-4), 109–118. doi:10.1002/1438-826X(200010)1:3/4<109::AID-GNFD109>3.0.CO;2-O

Shieh, J. S., & Whitman, W. B. (1987). Pathway of acetate assimilation in autotrophic and heterotrophic methanococci. *Journal of bacteriology*, *169*(11), 5327–9. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=213948&tool=pmcentrez&rendertype=abstract

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome research*, *19*(6), 1117–23. doi:10.1101/gr.089532.108

Stolyar, S., Van Dien, S., Hillesland, K. L., Pinel, N., Lie, T. J., Leigh, J. A., & Stahl, D. A. (2007). Metabolic modeling of a mutualistic microbial community. *Molecular systems biology*, *3*, 92. doi:10.1038/msb4100131

Taffs, R., Aston, J. E., Brileya, K., Jay, Z., Klatt, C. G., McGlynn, S., Mallette, N., et al. (2009). In silico approaches to study mass and energy flows in microbial consortia: a syntrophic case study. *BMC systems biology*, *3*(1), 114. doi:10.1186/1752-0509-3-114

Thauer, R. K., Kaster, A.-K., Seedorf, H., Buckel, W., & Hedderich, R. (2008). Methanogenic archaea: ecologically relevant differences in energy conservation. *Nature reviews. Microbiology*, *6*(8), 579–91. doi:10.1038/nrmicro1931

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V, et al. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, *428*(6978), 37–43. doi:10.1038/nature02340

Ulrich, A. C., & Edwards, E. A. (2003). Physiological and molecular characterization of anaerobic benzene-degrading mixed cultures. *Environmental Microbiology*, *5*(2), 92–102. doi:10.1046/j.1462-2920.2003.00390.x

Ulrich, A.C. (2004). *Characterization of Benzene-Degrading Cultures* (Doctoral Dissertation). University of Toronto, Toronto, ON

Varma, A., & Palsson, B. O. (1994). Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. *Bio/Technology*, *12*(10), 994–998. doi:10.1038/nbt1094-994

Wintermute, E. H., & Silver, P. A. (2010). Emergent cooperation in microbial metabolism. *Molecular systems biology*, *6*, 407. doi:10.1038/msb.2010.66

Wooley, J. C., Godzik, A., & Friedberg, I. (2010). A primer on metagenomics. *PLoS computational biology*, *6*(2), e1000667. doi:10.1371/journal.pcbi.1000667

Zhang, J., Chiodini, R., Badr, A., & Zhang, G. (2011). The impact of next-generation sequencing on genomics. *Journal of genetics and genomics = Yi chuan xue bao*, *38*(3), 95–109. doi:10.1016/j.jgg.2011.02.003

Zhang, T., Bain, T. S., Nevin, K. P., Barlett, M. A., & Lovley, D. R. (2012). Anaerobic benzene oxidation by Geobacter species. *Applied and environmental microbiology*, *78*(23), 8304–10. doi:10.1128/AEM.02469-12

Zhuang, K., Izallalen, M., Mouser, P., Richter, H., Risso, C., Mahadevan, R., & Lovley, D. R. (2011). Genome-scale dynamic modeling of the competition between Rhodoferax and Geobacter in anoxic subsurface environments. *The ISME journal*, *5*(2), 305–16. doi:10.1038/ismej.2010.117

Zomorrodi, A. R., & Maranas, C. D. (2012). OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. (C. V. Rao, Ed.)*PLoS computational biology*, *8*(2), e1002363. doi:10.1371/journal.pcbi.1002363

Appendices

List of Figures in the Appendices

List of Tables in the Appendices

Appendix A. Supplementary Materials for Chapter 2, 3, and 5.0-5.3

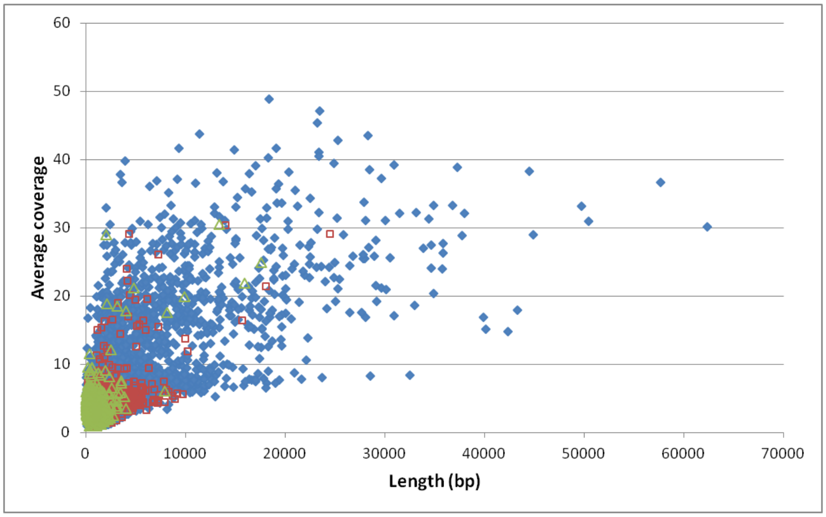

Figure A 1. Coverage of Domain-Specific Benzene Metagenome

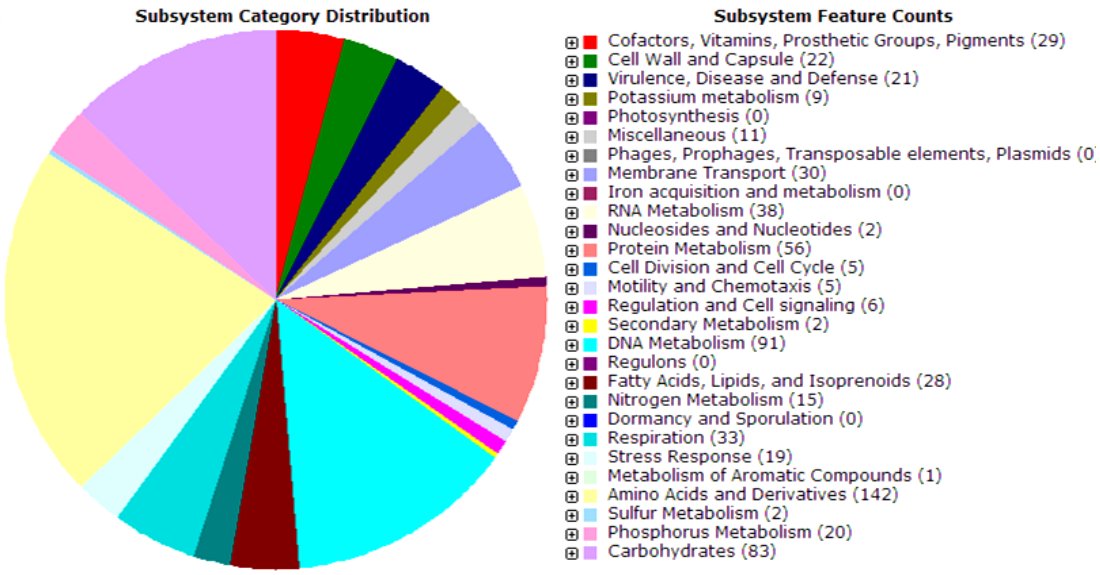

Figure A 2. Subsystems in the Archaea Bin Based on Protein Feature Counts in RAST

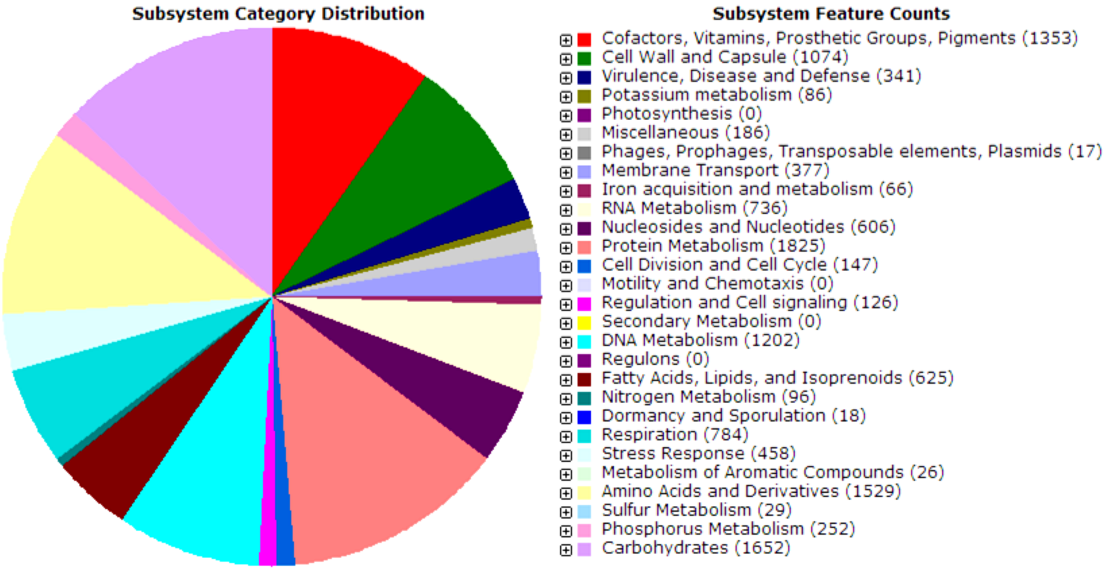

Figure A 4. Subsystems in the Bacteria Bin Based on Protein Feature Counts in RAST

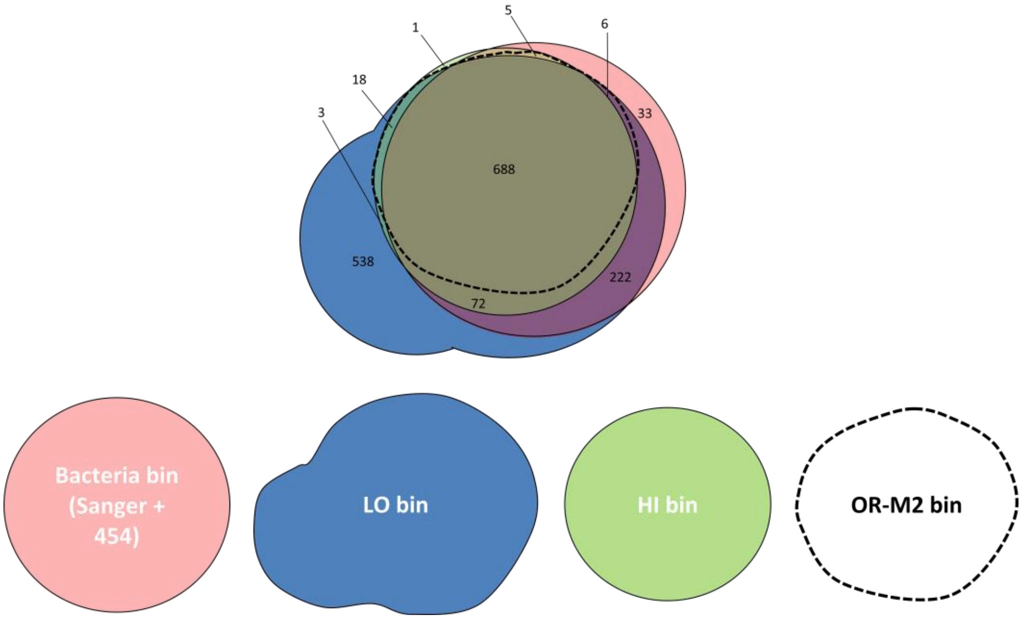

Figure A 3. Size Comparison of Reconstructed Models of the Methanogenic Benzene-Degrading Community.

Pink: Bacteria bin used in this thesis
Blue: Low abundance bin (Illumina assembly)
Green: High abundance bin (Illumina assembly)
Dotted: OR-M2 bin (Illumina assembly)

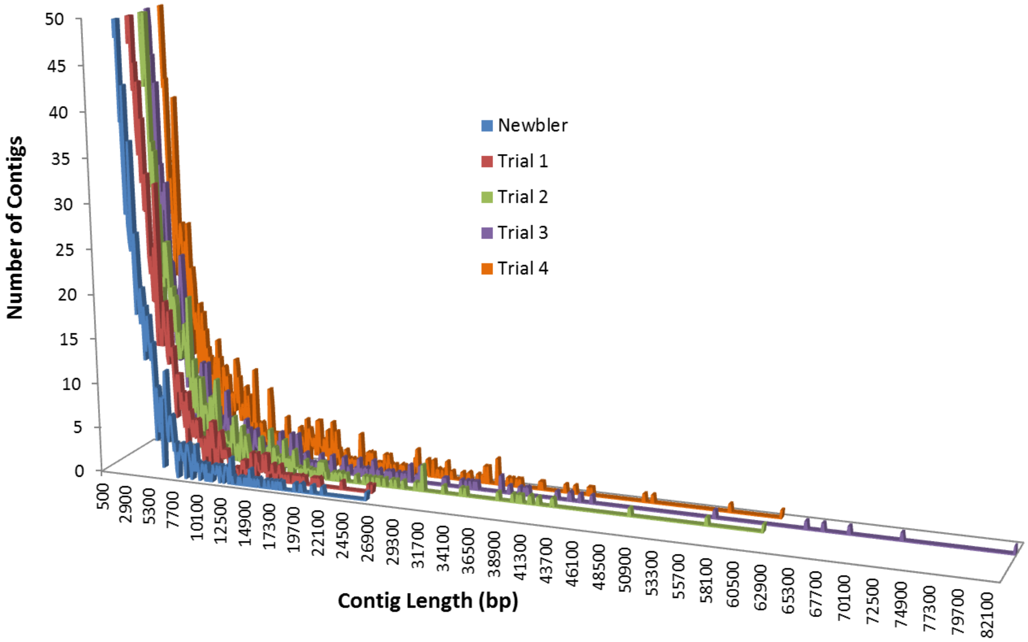

Figure A 5. Contig Length Distribution of Each Assembly

Table A 1. Summary of Input Reads to Metagenome Assembly

Sequencing Method	Trial 1	Trial 2	Trial 3	Trial 4
	Number of Input Reads			
Sanger	56663	56663	56663	56663
Roche 454	1052083	1315738	1015184	1315738
Sequencing Method	Number of Assembled Reads			
Sanger	42468	49193	42944	48278
Roche 454	607131	761145	625008	671155

Table A 2. Sequence Search for Benzoyl-CoA Reductase in the Benzene Metagenome

Contig ID	Benzoyl-CoA reductase genes			
	Azoarcus sp. EbN1	*Azoarcus sp. CIB*	*Geobacter metallireducens*	*Syntrophus aciditrophicus*
11213	bzdNQ	bzdNQ		
2130	bzdQ	bzdQ		
4645	bzdQ	bzdQ		
11627	bzdQ	bzdQ		
11859	bzdN	bzdN		
11863	bzdQ	bzdQ		
12190	bzdN	bzdN		
12282			**bamB**C**DE**FH**I**	**bamB**C**DE**FG**HI**
11568			**bam**C**DE**FH	bamC**DE**FG
1566			**bam**FH**I**	**bam**FG**HI**
11505			**bamB**C**D**	**bamB**C**D**
12006			**bam**H**I**	**bam**G**HI**
11181			**bamI**	**bamHI**
11466			bamC**DE**	**bam**C**D**
10729			**bamE**H	bam**E**G
11273			**bamE**F	bam**E**F
11714			**bamB**C	**bamB**
668			bamF	bamF
3296			**bamB**	**bamB**
11060			**bamE**	**bamE**
12248			**bamB**	**bamB**
12308			**bamI**	**bamH**
106			**bamD**	
4633				bamI
12220				bamD

Normal font: E-value is between e^{-40} and e^{-60}
Underlined: E-value is between e^{-60} and e^{-80}
Bold and underlined: E-value is between e^{-80} and zero
*For each subunit, only the top five hits are shown

Table A 3. Sequence Search for Gene Products of Modified beta-Oxidation in the Benzene Metagenome

Contig ID	Acyl-CoA hydratase						
	R. palustris	M. magneticum	T. aromatica	Azoarcus sp. EbN1	Azoarcus. sp CIB	G. metallireducens	S. aciditrophicus
1566		dch	dch			bamR	
12006		dch	dch			bamR	
12282		dch	dch			bamR	
9811				badK			
11527				badK			
11647				badK			
11868				badK			
12035				badK			
	Hydroxyacyl-CoA dehydrogenase						
11705	badH						
11916	badH						
12182	badH						
6650					bzdX		bamQ
10409		had	had		**bzdX**	bamQ	**bamQ**
10733					bzdX		bamQ
11645		had	had		**bzdX**	bamQ	**bamQ**
12230		had	had		**bzdX**	bamQ	**bamQ**
	Oxoacyl-CoA hydrolase						
4795	badI			bzdY			
11205	badI			bzdY			
11807	badI			bzdY			
12248	badI			bzdY			
4686		oah	oah		bzdY	bamA	bamA
9025		oah	**oah**		**bzdY**	bamA	bamA
11465		**oah**	**oah**		**bzdY**	**bamA**	bamA
12009		**oah**	**oah**		**bzdY**	**bamA**	bamA
12037		**oah**	**oah**		**bzdY**	**bamA**	bamA

Normal font: E-value is between e^{-40} and e^{-60}

Underlined: E-value is between e^{-60} and e^{-80}

Bold and underlined: E-value is between e^{-80} and zero

*For each subunit, only the top five hits are shown

Appendix B. Supplementary Info: Gap-Filling the Bacteria and Archaea Draft Models

Table B1. Reactions Added to the Bacteria Draft Model with Enforced Hydrogen Export

Reaction ID	Reaction Name	Subsystem
H1	Thiamin-ABC transport	Transport
H2	meso-2,6-diaminoheptanedioate NADP+ oxidoreductase	Lysine metabolism
H3	ATP-L-homoserine O-phosphotransferase	Glycine, serine and threonine metabolism
H4	Hydrogen-sulfide NADP+ oxidoreductase	Sulfur metabolism
H5	Urea carbon-dioxide ligase	Urea cycle and metabolism of amino groups
H6	ATP-thiamin pyrophosphotransferase	Thiamine metabolism
H7	Hydrogen transport	Transport
H8	Transport of benzoate	Transport
H9	Transport of benzoate	Transport
H10	Fatty acid oxidation (octadecanoateubiquinone)	None
H11	5-Methylthioadenosine transport	Transport
H12	NADH dehydrogenase (ubiquinone-8 3.5 protons)	None
H13	Exchange of 5-Methylthioadenosine	Exchange

Table B2. Reactions Added to the Bacteria Draft Model with Enforced Methanol Export

Reaction ID	Reaction Name	Subsystem
M1	meso-2,6-diaminoheptanedioate NADP+ oxidoreductase	Lysine metabolism
M2	L-Threonine acetaldehyde-lyase	Glycine, serine and threonine metabolism
M3	ATP-thiamin pyrophosphotransferase	Thiamine metabolism
M4	S-Adenosyl-L-methionine protein-L-glutamate O-methyltransferase	None
M5	Protein-L-glutamate-O4-methyl-ester acylhydrolase	None
M6	Thiamine transport in via proton symport	Transport
M7	b-ketoacyl synthetase (n-C181)	None
M8	Methanol diffusion	Transport
M9	Transport of benzoate	Transport
M10	Transport of benzoate	Transport
M11	Urea carboxylase	None
M12	5-Methylthioadenosine transport	Transport
M13	NADH dehydrogenase	None
M14	Hydrogen-sulfide ferredoxin oxidoreductase	None
M15	Exchange of 5-Methylthioadenosine	Exchange

Table B 3. List of Reactions Added to the Archaea Draft Model with Acetate as the Carbon and Energy Source

ID	SEED ID	Reaction Name	Restored Biomass Precursors
1	rxn00737	L-threonine ammonia-lyase	Isoleucine
2	rxn03437	(R)-2,3-Dihydroxy-3-methylpentanoate hydro-lyase	Isoleucine
3	rxn00902	3-Carboxy-3-hydroxy-4-methylpentanoate 3-methyl-2-oxobutanoate-lyase	Leucine
4	rxn00898	2,3-Dihydroxy-3-methylbutanoate hydro-lyase	Leucine and Valine
5	rxn02914	3-Phosphoserine2-oxoglutarate aminotransferase	L-Tryptophan, Glycine, L-Threonine, L-Serine, ATP, L-Cysteine, L-Isoleucine, S-Adenosyl-L-methionine, FAD
6	rxn05909	L-Serine hydro-lyase (adding homocysteine)	CoA, Cysteine
7	rxn03175	Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase	Histidine
8	rxn02835	1-(5-phospho-D-ribosyl)-AMP 1,6-hydrolase	Histidine
9	rxn00789	1-(5-Phospho-D-ribosyl)-ATPpyrophosphate phosphoribosyl-transferase	Histidine
10	rxn02834	Phosphoribosyl-ATP pyrophosphohydrolase	Histidine
11	rxn02160	L-Histidinol-phosphate phosphohydrolase	Histidine
12	rxn00791	N-(5-Phospho-D-ribosyl)anthranilatepyrophosphate	Tryptophan
13	rxn02508	N-(5-Phospho-beta-D-ribosyl)anthranilate ketol-isomerase	Tryptophan
14	rxn02507	1-(2-Carboxyphenylamino)-1-deoxy-D-ribulose-5-phosphate	Tryptophan
15	rxn02155	ATPnicotinamide-nucleotide adenylyltransferase	NAD, NADP
16	rxn02402	Nicotinate-nucleotidepyrophosphate phosphoribosyltransferase	NAD, NADP
17	rxn02341	N-[(R)-4-Phosphopantothenoyl]-L-cysteine carboxy-lyase	CoA
18	rxn02175	ATPpantetheine-4-phosphate adenylyltransferase	CoA
19	rxn12510	ATPpantothenate 4-phosphotransferase	CoA
20	rxn12512	(R)-4-PhosphopantothenateL-cysteine ligase	CoA
21	rxn01991	meso-2,6-diaminoheptanedioateNADP+ oxidoreductase (deaminating)	Lysine
22	rxn01644	L-Aspartate-4-semialdehyde hydro-lyase	Lysine
23	rxn00829	ATP(R)-5-diphosphomevalonate carboxy-lyase (dehydrating)	Bactoprenyl diphosphate
24	rxn02322	ATP(R)-5-phosphomevalonate phosphotransferase	Bactoprenyl diphosphate
25	rxn00178	Acetyl-CoAacetyl-CoA C-acetyltransferase	Bactoprenyl diphosphate
26	rxn01501	(R)-MevalonateNADP+ oxidoreductase (CoA acylating)	Bactoprenyl diphosphate
27	rxn01465	(S)-Dihydroorotate amidohydrolase	dTTP, CTP, UTP
28	rxn01362	Orotidine-5-phosphatepyrophosphate phosphoribosyltransferase	dTTP, CTP, UTP
29	rxn01256	Chorismate pyruvatemutase	Phenylalanine, tyrosine
30	rxn01332	PhosphoenolpyruvateD-erythrose-4-phosphate	Phenylalanine, Tyrosine, Tryptophan
31	rxn00548	D-Fructose-6-phosphate D-erythrose-4-phosphate-lyase	Phenylalanine, Tyrosine, Tryptophan
32	rxn01255	5-O-(1-Carboxyvinyl)-3-phosphoshikimate phosphate-lyase	Phenylalanine, Tyrosine, Tryptophan
33	rxn13847	Threonine dehydrogenase	Tryptophan, glycine, serine, CoA, cysteine, spermidine, FAD, GTP
34	rxn01106	2-Phospho-D-glycerate 2,3-phosphomutase	Pyridoxal phosphate, L-Phenylalanine, L-Tryptophan, TTP, L-Histidine, Glycine, L-Threonine, CTP, Peptidoglycan polymer (n subunits), L-Serine, ACP, L-Methionine, L-Cysteine, dATP, Spermidine, UTP, L-Isoleucine, S-Adenosyl-L-methionine, L-Tyrosine, GTP

ID	SEED ID	Reaction Name	Restored Biomass Precursors
35	rxn00790	5-Phosphoribosylaminepyrophosphate phosphoribosyltransferase	CoA, dATP, Spermidine, S-Adenosyl-L-methionine, FAD, GTP
36	rxn01303	Acetyl-CoAL-homoserine O-acetyltransferase	Methionine, S-adenosylmethionine, spermidine
37	rxn00623	hydrogen-sulfideNADP+ oxidoreductase	L-Methionine, CoA, L-Cysteine, Spermidine, S-Adenosyl-L-methionine
38	rxn00378	Sulfiteferricytochrome-c oxidoreductase	L-Methionine, L-Cysteine, Spermidine, S-Adenosyl-L-methionine
39	rxn13689	cytochrome-c reductase (ubiquinol8 4 protons translocated)	Methionine, cysteine, S-adenosylmethionine, spermidine
40	rxn12215	5-methyltetrahydropteroyltri-l-glutamate synthesis	Methionine, Spermidine, S-Adenosyl-L-methionine
41	rxn00429	FormaldehydeNAD+ oxidoreductase	-
42	rxn03638	Acetyl-CoAD-glucosamine-1-phosphate N-acetyltransferase	-
43	rxn13687	cytochrome-c peroxidase	-
44	rxn00293	UTPN-acetyl-alpha-D-glucosamine-1-phosphate uridylyltransferase	-
45	rxn00138	Deamino-NAD+ammonia ligase (AMP-forming)	-
46	rxn08291	D-Amino acid dehydrogenase	Alanine
47	rxn09174	Propanoyl-CoA succinate CoA-transferase	Arginine, Glutamate, Proline, Spermidine
48	rxn10471	Methane Transport	Transport
49	rxn10904	Acetate Transport	Transport
50	EX_cpd00025(e)	EX H2O2	Transport
51	rxn13085	Spermidine synthase	Spermidine
52	rxn13799	5mta transport irreversible, extracellular	Spermidine
53	EX_cpd00147(e)	EX 5-Methylthioadenosine e	Spermidine

Appendix C. Supplementary Materials of Community Gap-Filling

Table C 1. Exchange Flux Distribution of the Benzene Community Model (Intermediate Not Specified)

Exchange Reaction	Flux (umol/gDW/hr)		Exchange Reaction	Flux (umol/gDW/hr)	
	Bacteria	Archaea		Bacteria	Archaea
EX_Thiamin	4.10E-04	0	EX_gly-glu-L	1.00E+01	-1.00E+01
EX_Fe2+	4.10E-04	5.36E-03	EX_gly-asp-L	-7.77E+00	7.77E+00
EX_Cu2+	4.10E-04	5.36E-03	EX_Gly-Leu	-3.10E-01	3.10E-01
EX_Folate	1.23E-03	0	EX_ala-L-asp-L	6.14E+00	-6.14E+00
EX_NH3	6.03E+00	0	EX_Putrescine	4.88E+00	-4.88E+00
EX_H+	2.91E+00	0	EX_Urea	4.88E+00	-4.88E+00
EX_Benzoate	3.75E+00	0	EX_Gly-Phe	9.78E+00	-9.78E+00
EX_Zn2+	4.10E-04	5.36E-03	EX_PPi	7.92E-02	-7.92E-02
EX_K+	4.10E-04	5.36E-03	EX_gly-asn-L	2.82E+00	-2.82E+00
EX_Sulfate	2.10E-01	5.36E-03	EX_Ala-Leu	1.00E+01	-1.00E+01
EX_Mg	4.10E-04	5.36E-03	EX_Cys-Gly	-4.21E+00	4.21E+00
EX_Co2+	4.10E-04	5.36E-03	EX_Gly-Cys	4.14E+00	-4.14E+00
EX_Mn2+	4.10E-04	5.36E-03	EX_Spermidine	4.10E-04	-4.10E-04
EX_Cl-	4.10E-04	5.36E-03	EX_L-Isoleucine	-2.00E-01	2.00E-01
EX_fe3	4.10E-04	5.36E-03	EX_L-Valine	-2.91E-01	2.91E-01
EX_Riboflavin	8.19E-04	1.07E-02	EX_ala-L-Thr-L	2.61E-01	-2.61E-01
EX_Ca2+	4.10E-04	5.36E-03	EX_L-alanylglycine	-3.31E+00	3.31E+00
EX_ala-L-glu-L	6.41E+00	-6.41E+00	EX_L-Tryptophan	-3.90E-02	3.90E-02
EX_Gly-Tyr	-1.00E+01	1.00E+01	EX_Adenosine	-3.01E-01	3.01E-01
EX_met-L-ala-L	-1.00E+01	1.00E+01	EX_Thymidine	-2.08E-01	2.08E-01
EX_L-Lysine	-3.01E-01	3.01E-01	EX_Dephospho-CoA	-1.07E-02	1.07E-02
EX_Gly-Met	9.89E+00	-9.89E+00	EX_N-Acetyl-D-mannosamine	-1.31E-01	1.31E-01
EX_Gly-Gln	-1.00E+01	1.00E+01	EX_Palmitate	-1.07E+00	1.07E+00
EX_L-Leucine	-1.00E+01	1.00E+01	EX_Nicotinamide ribonucleotide	-1.07E-02	1.07E-02
EX_Ala-Gln	-1.00E+01	1.00E+01	EX_Phosphate	0	9.52E-01
EX_gly-pro-L	-2.08E+00	2.08E+00	EX_Methane	0	-2.92E+00
EX_Ala-His	-6.55E-02	6.55E-02	EX_CO2	0	-3.22E+00

Table C 2. Exchange Fluxes of the Benzene Community Model with Acetate as the Intermediate (mmol·gDW^{-1}·month^{-1})

Exchange Reaction	Bacteria			Archaea		
	Solution 1	Solution 2	Solution 3	Solution 1	Solution 2	Solution 3
Growth	5.60E-01	5.60E-01	5.62E-01	1.00E-02	1.00E-02	1.00E-02
Thiamin	2.29E-03	2.29E-03	2.30E-03	0	0	0
Phosphate	5.65E-01	5.65E-01	5.71E-01	8.29E-03	8.29E-03	3.96E-03
Fe2+	2.29E-03	2.29E-03	2.30E-03	7.58E-05	7.58E-05	7.58E-05
Cu2+	2.29E-03	2.29E-03	2.30E-03	7.58E-05	7.58E-05	7.58E-05
Folate	6.88E-03	6.88E-03	6.90E-03	0	0	0
Acetate	-1.42E+00	-1.42E+00	-1.43E+00	1.42E+00	1.42E+00	1.43E+00
NH3	3.95E+00	3.95E+00	3.96E+00	5.72E-02	5.72E-02	2.58E-02
H+	3.74E+00	3.74E+00	3.74E+00	1.18E-01	1.18E-01	7.35E-02
Benzoate	2.70E+00	2.70E+00	2.70E+00	0	0	0
Zn2+	2.29E-03	2.29E-03	2.30E-03	7.58E-05	7.58E-05	7.58E-05
K+	2.29E-03	2.29E-03	2.30E-03	7.58E-05	7.58E-05	7.58E-05
Sulfate	2.29E-03	2.29E-03	2.30E-03	7.58E-05	7.58E-05	7.58E-05
Mg	2.29E-03	2.29E-03	2.30E-03	7.58E-05	7.58E-05	7.58E-05
Co2+	2.29E-03	2.29E-03	2.30E-03	7.58E-05	7.58E-05	7.58E-05
CO2	-5.24E-01	-5.24E-01	-5.33E-01	-1.28E+00	-1.28E+00	-1.35E+00
Mn2+	2.29E-03	2.29E-03	2.30E-03	7.58E-05	7.58E-05	7.58E-05
Cl-	2.29E-03	2.29E-03	2.30E-03	7.58E-05	7.58E-05	7.58E-05
fe3	2.29E-03	2.29E-03	2.30E-03	7.58E-05	7.58E-05	7.58E-05
Riboflavin	4.58E-03	4.58E-03	4.60E-03	1.52E-04	1.52E-04	1.52E-04
Ca2+	2.29E-03	2.29E-03	2.30E-03	7.58E-05	7.58E-05	7.58E-05
Xanthine	-4.33E-03	-4.33E-03	0	4.33E-03	4.33E-03	0.00E+00
Methane	0	0	0	-1.26E+00	-1.26E+00	-1.33E+00
TRHL	0	0	0	-2.44E-02	-2.44E-02	-1.96E-02

Table C 3. Exchange Fluxes of the Benzene Community Model with Hydrogen and Formate as the Intermediate (mmol·gDW^{-1}·month^{-1})

Exchange Reaction	Bacteria			Archaea		
	Solution 1	Solution 2	Solution 3	Solution 1	Solution 2	Solution 3
Growth	3.26E-01	3.26E-01	3.24E-01	3.22E-02	3.22E-02	4.97E-01
Thiamin	1.51E-03	1.51E-03	1.48E-03	0	0	0
Phosphate	3.74E-01	3.74E-01	3.81E-01	3.68E-02	3.68E-02	5.70E-01
Fe2+	1.51E-03	1.51E-03	1.48E-03	2.49E-04	2.49E-04	3.85E-03
Cu2+	1.51E-03	1.51E-03	1.48E-03	2.49E-04	2.49E-04	3.85E-03
Folate	4.53E-03	4.53E-03	4.43E-03	0	0	0
NH3	3.81E+00	3.81E+00	3.90E+00	3.45E-01	3.45E-01	3.98E-02
H+	1.88E+00	1.88E+00	1.86E+00	1.03E+00	1.03E+00	0
Benzoate	2.7	2.7	2.7	0	0	0
Zn2+	1.51E-03	1.51E-03	1.48E-03	2.49E-04	2.49E-04	3.85E-03
K+	1.51E-03	1.51E-03	1.48E-03	2.49E-04	2.49E-04	3.85E-03
H2O	4.17E+00	4.17E+00	4.19E+00	0	0	0
Sulfate	1.51E-03	1.51E-03	1.48E-03	8.32E-03	8.32E-03	1.29E-01
Mg	1.51E-03	1.51E-03	1.48E-03	2.49E-04	2.49E-04	3.85E-03
Co2+	1.51E-03	1.51E-03	1.48E-03	2.49E-04	2.49E-04	3.85E-03
Mn2+	1.51E-03	1.51E-03	1.48E-03	2.49E-04	2.49E-04	3.85E-03
Cl-	1.51E-03	1.51E-03	1.48E-03	2.49E-04	2.49E-04	3.85E-03
Formate	-2.7	-2.7	-2.7	2.7	2.7	2.7
fe3	1.51E-03	1.51E-03	1.48E-03	2.49E-04	2.49E-04	3.85E-03
H2	-3.03E+00	-3.03E+00	-3.24E+00	3.03E+00	3.03E+00	3.24E+00
Riboflavin	3.02E-03	3.02E-03	2.96E-03	4.98E-04	4.98E-04	7.70E-03
Ca2+	1.51E-03	1.51E-03	1.48E-03	2.49E-04	2.49E-04	3.85E-03
Methane	0	0	0	-1.37E+00	-1.37E+00	-3.36E+00

Table C 4. Reactions Added to the Benzene Community Model When Acetate is the Key Intermediate

Note 1: Reactions in grey are shared by the reaction profiles suggested during the individual gap-filling

Note 2: The first 30 reactions are added to bacteria, with the rest included into the archaea model

#	Reaction Name	Formula	Functional Category
1	dTMP transport in/out	H+[e] + dTMP[e] <=> H+ + dTMP	Transport
2	acetate transport in/out via proton symport	ACET[e] + H+[e] <=> ACET + H+	Transport
3	HISt2	H+[e] + L-Histidine[e] <=> H+ + L-Histidine	Transport
4	L-Threonine acetaldehyde-lyase	threonine <=> Gly + AALD	Glycine, Serine, and Threonine Biosynthesis
5	ATP:thiamin pyrophosphotransferase	ATP + THI <=> AMP + TPP	Thiamine Metabolism
6	ATP:pyridoxal 5'-phosphotransferase	ATP + Pyridoxal <=> ADP + Pyridoxal phosphate	Vitamin B6 Metabolism
7	R04519 transferase	UDN-aNa-L-ala-D-glu-meso-2-6-diaminopimeloyl-D-ala-D-ala <=> Undecaprenyldiphosphate	Peptidoglycan Biosynthesis
8	L-methionine reversible transport via proton symport	Methionine[e] + H+[e] <=> Methionine + H+	Transport
9	Indole transport via proton symport	H+[e] + indole[e] <=> H+ + indole	Transport
10	Fatty acid oxidation (octadecenoate)	8 H2O + ATP + 8 NAD + 9 CoA + 7 FAD + octadecenoate <=> 8 NADH + PPi + AMP + 9 Acetyl-CoA + 7 H+ + 7 FADH2	Fatty Acid Oxidation
11	CoA transporter	CoA[e] <=> CoA	Transport
12	EX_thm	THI[e] <=> THI	Thiamine Metabolism
13	rxn11062	apo-ACP <=> ACP	Unknown
14	rxn11334	H+[e] + NICO[e] <=> H+ + NICO	Transport
15	rxn11337	H+[e] + Phenylalanine[e] <=> L-Phenylalanine + H+	Transport
16	transport of benzoate	H+[p] + Benzoate[p] <=> H+ + Benzoate	Transport
17	transport of benzoate	Benzoate[e] <=> Benzoate[p]	Transport
18	Pyridoxal transport	Pyridoxal[e] <=> Pyridoxal	Transport
19	Urea carboxylase	ATP + urea + HCO3- = ADP + phosphate + urea-1-carboxylate	Unknown
20	5mta transport	5'-Methylthioadenosine <=> 5'-Methylthioadenosine	Transport
21	NADH dehydrogenase	(10) NADH + (45) H+ + (10) Ubiquinone-8 <=> (10) NAD+ + (35) H+ + (10) Ubiquinol-8	Transport
22	EX_dTMP	dtmp[e] <=> dtmp[e]_m1	Model Exchange
23	EX_Phenylalanine	phenylalanine[e] <=> phenylalanine[e]_m1	
24	EX_Nicotinamide	NICO[e] <=> NICO[e]_m1	
25	EX_L-Histidine	Histidine[e] <=> Histidine[e]_m1	
26	EX_indol	indole[e] <=> indole_m1	
27	EX_5-Methylthioadenosine	5mta[e] <=> 5mta[e]_m1	
28	EX_Pyridoxal	Pyridoxal[e] <=> Pyridoxal[e]_m1	
29	EX_L-Methionine	methionine[e] <=> methionine[e]_m1	
30	EX_CoA	CoA[e] <=> CoA[e]_m1	
31	dTMP transport in/out via proton symport	H+[e] + dTMP[e] <=> H+ + dTMP	Transport
32	L-Methionine ABC transport	H2O + ATP + Methionine[e] <=> ADP + Phosphate + Methionine + H+	Transport
33	Acetyl-CoA:D-glucosamine-1-phosphate N-acetyltransferase	Acetyl-CoA + D-Glucosamine1-phosphate <=> CoA + H+ + N-Acetyl-D-glucosamine1-phosphate	Peptidoglycan biosynthesis
34	(R)-2,3-Dihydroxy-3-methylpentanoate hydro-lyase	2,3-Dihydroxy-3-methylvalerate <=> H2O + 3MOP	Valine, Leucine, and Isoleucine Biosynthesis
35	3-Phosphoserine:2-oxoglutarate aminotransferase	2-Oxoglutarate + phosphoserine <=> GLU + 3-Phosphonooxypyruvate	Glycine, Serine, and Threonine Biosynthesis
36	4-Hydroxyphenyllactate:NADP+ oxidoreductase	TPN + HPL <=> TPNH + H+ + 4-Hydroxyphenylpyruvate	Tyrosine Metabolism
37	4-hydroxyphenyllactate:NAD+ oxidoreductase	NAD + HPL <=> NADH + H+ + 4-Hydroxyphenylpyruvate	Tyrosine Metabolism
38	ATP:(R)-5-phosphomevalonate phosphotransferase	ATP + (R)-5-Phosphomevalonate <=> ADP + (R)-5-Diphosphomevalonate	Archaea Lipids
39	meso-2,6-diaminoheptanedioate:NADP+ oxidoreductase (deaminating)	H2O + TPN + meso-2,6-Diaminopimelate <=> TPNH + NH3 + H+ + L-2-Amino-6-oxopimelate	Lysine Biosynthesis
40	L-Aspartate-4-semialdehyde hydro-lyase (adding pyruvate and	Pyruvate + L-Aspartic 4-semialdehyde <=> 2 H2O + H+ + Dihydrodipicolinate	Lysine Biosynthesis

#	Reaction Name	Formula	Functional Category
41	XMP:pyrophosphate phosphoribosyltransferase	PPi + XMP <=> PRPP + XAN	Purine Metabolism
42	(R)-Mevalonate:NADP+ oxidoreductase (CoA acylating)	2 TPN + CoA + (R)-Mevalonate <=> 2 TPNH + 2 H+ + HMG-CoA	Archaea Lipids
43	3-Carboxy-3-hydroxy-4-methylpentanoate 3-methyl-2-oxobutanoate-lyase	CoA + H+ + 2-Isopropylmalate <=> H2O + Acetyl-CoA + 3MOB	Valine, Leucine, and Isoleucine Biosynthesis (pyruvate metabolism)
44	2,3-Dihydroxy-3-methylbutanoate hydro-lyase	2,3-Dihydroxy-isovalerate <=> H2O + 3MOB	Valine, Leucine, and Isoleucine Biosynthesis (pyruvate metabolism)
45	ATP:(R)-5-diphosphomevalonate carboxy-lyase (dehydrating)	ATP + (R)-5-Diphosphomevalonate <=> ADP + Phosphate + CO2 + Isopentenyldiphosphate	Archaea Lipids
46	O-Succinyl-L-homoserine succinate-lyase (adding cysteine)	H2O + O-Succinyl-L-homoserine <=> NH3 + Succinate + H+ + 2-Oxobutyrate	Methionine Metabolism
47	L-threonine ammonia-lyase	threonine <=> NH3 + 2-Oxobutyrate	Valine, Leucine, and Isoleucine Biosynthesis/ Glycine, Serine, Threonine Metabolism
48	L-Threonine acetaldehyde-lyase	threonine <=> Gly + AALD	Glycine, Serine, and Threonine Biosynthesis
49	L-Alanine:NAD+ oxidoreductase (deaminating)	H2O + NAD + ALA <=> NADH + NH3 + Pyruvate + H+	Reductive Carboxylate Cycle
50	Acetyl-CoA:acetyl-CoA C-acetyltransferase	2 Acetyl-CoA <=> CoA + Acetoacetyl-CoA	Pyruvate Metabolism and 6 other Subsystems
51	ATP:pyridoxal 5'-phosphotransferase	ATP + Pyridoxal <=> ADP + Pyridoxal phosphate	Vitamin B6 Metabolism
52	L-Serine hydro-lyase (adding homocysteine)	Serine + H2S <=> H2O + L-Cysteine	Glycine, Serine, and Threonine Biosynthesis
53	Propanoyl-CoA: succinate CoA-transferase	Succinate + Propanoyl-CoA <=> Succinyl-CoA + Propanoate	Vitamin B6 Metabolism
54	UDP-N-acetylglucosamine diphosphorylase	UTP + N-Acetyl-D-glucosamine1-phosphate <=> PPi + UDP-N-acetyl-D-galactosamine	Glycine, Serine, and Threonine Biosynthesis
55	xanthine reversible transport	XAN[e] <=> XAN	Transport
56	NMN transport via NMN glycohydrolase	H2O + NMN[e] <=> H+ + Ribose 5-phosphate + NICO	Transport
57	Indole transport via proton symport, reversible	H+[e] + indol[e] <=> H+ + indol	Transport
58	Methane Transport	Methane[e] <=> Methane	Transport
59	CoA transporter	CoA[e] <=> CoA	Transport
60	Nicotinamide acid uptake	NICO[e] <=> NICO	Transport
61	EX_ac_e	ACET[e] <=> ACET	Transport
62	EX_his_L_e	L-Histidine[e] <=> L-Histidine	Transport
63	rxn11337	H+[e] + Phenylalanine[e] <=> L-Phenylalanine + H+	Transport
64	5-methyltetrahydropteroyltri-l-glutamate synthesis	NADH + Serine + H+ + Tetrahydropteroyltri-L-glutamate <=> H2O + NAD + Gly + 5-Methyltetrahydropteroyltri-L-glutamate	Not classified
65	Pyridoxal transport	Pyridoxal[e] <=> Pyridoxal	Transport
66	Spermidine synthase	Putrescine + S-Adenosyl-L-methionine -> CO2 + 5-Methylthioadenosine + Spermidine	Spemidine Synthesis
67	Transport of L-threonine, mitochondrial	l-threonine[e] <=> l-threonine[m]	Transport
68	5mta transport irreversible, extracellular	5'-Methylthioadenosine <=> 5'-Methylthioadenosine	Transport
69	EX_dTMP	dtmp[e] <=> dtmp[e]_m2	Model Exchange
70	EX_Phenylalanine	phenylalanine[e] <=> phenylalanine[e]_m2	Model Exchange
71	EX_Nicotinamide	NICO[e] <=> NICO[e]_m2	Model Exchange
72	EX_L-Histidine	Histidine[e] <=> Histidine[e]_m2	Model Exchange
73	EX_indol	indole[e] <=> indole_m2	Model Exchange
74	EX_5-Methylthioadenosine	5mta[e] <=> 5mta[e]_m2	Model Exchange
75	EX_Pyridoxal	Pyridoxal[e] <=> Pyridoxal[e]_m2	Model Exchange
76	EX_L-Methionine	methionine[e] <=> methionine[e]_m2	Model Exchange
77	EX_XAN	Xanthine[e] <=> Xanthine[e]_m2	Model Exchange
78	EX_CoA	CoA[e] <=> CoA[e]_m2	Model Exchange
79	EX_TRHL	trehalose[e] <=>	Medium Exchange

#	Reaction Name	Formula	Functional Category
80	EX_5-Methylthioadenosine	5mta[e] <=>	Medium Exchange

Table C 5. Reactions Added to the Benzene Community Model When Hydrogen is the Key Intermediate

Note 1: Reactions in grey are shared by the reaction profiles suggested during the individual gap-filling

Note 2: The first 18 reactions are added to bacteria, with the rest included into the archaea model

#	Reaction Name	Formula	Functional Category
1	S-Methyl-5-thio-D-ribulose-1-phosphate hydro-lyase	methylthioribulose-1-phosphate <=> H_2O + 2,3-diketo5-methylthio-1-phosphopentane	Methionine Metabolism
2	meso-2,6-diaminoheptanedioate:NADP+ oxidoreductase (deaminating)	H_2O + TPN + meso-2,6-Diaminopimelate <=> TPNH + NH_3 + H+ + L-2-Amino-6-oxopimelate	Lysine Biosynthesis
3	ATP:thiamine phosphotransferase	ATP + THI <=> ADP + TMP	Thiamine Metabolism
4	N-Carbamoylputrescine amidohydrolase	H_2O + 2 H+ + N-Carbamoylputrescine <=> CO_2 + NH_3 + PUTR	Urea Cycle
5	Thiamine transport in via proton symport	H+[e] + THI[e] <=> H+ + THI	Transport
6	2,3-diketo-5-methylthio-1-phosphopentane degradation reaction	3 H_2O + 2,3-diketo5-methylthio-1-phosphopentane <=> Phosphate + FORM + 7 H+ + 4-methylthio 2-oxobutyrate	Not classified
7	Fatty acid oxidation (octadecanoate)	8 H_2O + ATP + 8 NAD + 9 CoA + 8 FAD + Stearate <=> 8 NADH + PPi + AMP + 9 Acetyl-CoA + 7 H+ + 8 $FADH_2$	Not classified
8	hydrogen transport	H_2 <=> H_2[e]	Model Exchange
9	CoA transporter	CoA[e] <=> CoA	Model Exchange
10	UNK2	L-Glutamine + 2 H+ + 4-methylthio 2-oxobutyrate <=> GLU + Methionine	Not classified
11	rxn11062	apo-ACP <=> ACP	Not classified
12	rxn11327	H+[e] + Histidine[e] <=> H+ + L-Histidine	Transport
13	transport of formate	FORM[e] + H+[p] <=> FORM + H+	Transport
14	transport of benzoate	H+[p] + Benzoate[p] <=> H+ + Benzoate	Transport
15	transport of benzoate	Benzoate[e] <=> Benzoate[p]	Transport
16	Threonine dehydrogenase	CoA + L-Threonine <=> Acetyl-CoA + Glycine + Hydrogen	Not classified
17	EX_Histidine	Histidine[e] <=> Histidine[e]_m1	Model Exchange
18	EX_CoA	CoA[e] <=> CoA[e]_m1	Model Exchange
19	O-phospho-L-serine:hydrogen-sulfide	H_2S + phosphoserine <=> Phosphate + H+ + L-Cysteine	Cysteine Metabolism
20	Acetyl-CoA:D-glucosamine-1-phosphate N-acetyltransferase	Acetyl-CoA + D-Glucosamine1-phosphate <=> CoA + H+ + N-Acetyl-D-glucosamine1-phosphate	Aminosugars Metabolism
21	(R)-2,3-Dihydroxy-3-methylpentanoate hydro-lyase	2,3-Dihydroxy-3-methylvalerate <=> H_2O + 3MOP	Valine, Leucine, and Isoleucine Biosynthesis
22	3-Phosphoserine:2-oxoglutarate aminotransferase	2-Oxoglutarate + phosphoserine <=> GLU + 3-Phosphonooxypyruvate	Glycine, Serine, and Threonine Metabolism
23	N-(5-Phospho-beta-D-ribosyl)anthranilate ketol-isomerase	N-5-phosphoribosyl-anthranilate <=> 1-(2-carboxyphenylamino)-1-deoxyribulose 5-phosphate	Phenylalanine, Tyrosine, and Tryptophan Biosynthesis
24	1-(2-Carboxyphenylamino)-1-deoxy-D-ribulose-5-phosphate	H+ + 1-(2-carboxyphenylamino)-1-deoxyribulose 5-phosphate <=> H_2O + CO_2 + Indoleglycerol phosphate	Phenylalanine, Tyrosine, and Tryptophan Biosynthesis
25	Nicotinate-nucleotide:pyrophosphate phosphoribosyltransferase	CO_2 + PPi + Nicotinate ribonucleotide <=> 2 H+ + PRPP + Quinolinate	Nicotinate and Nicotinamide Metabolism
26	ATP:(R)-5-phosphomevalonate phosphotransferase	ATP + (R)-5-Phosphomevalonate <=> ADP + (R)-5-Diphosphomevalonate	Archaeal Lipid Biosynthesis
27	ATP:nicotinamide-nucleotide adenylyltransferase	ATP + Nicotinate ribonucleotide <=> PPi + Deamido-NAD	Nicotinate and Nicotinamide Metabolism
28	meso-2,6-diaminoheptanedioate:NADP+ oxidoreductase (deaminating)	H_2O + TPN + meso-2,6-Diaminopimelate <=> TPNH + NH_3 + H+ + L-2-Amino-6-oxopimelate	Lysine Biosynthesis
29	L-2,4-Diaminobutanoate:pyruvate aminotransferase	Pyruvate + L-2,4-Diaminobutyrate <=> ALA + L-Aspartic 4-semialdehyde	Not classified
30	L-Aspartate-4-semialdehyde hydro-lyase (adding pyruvate and	Pyruvate + L-Aspartic 4-semialdehyde <=> 2 H_2O + H+ + Dihydrodipicolinate	Lysine Biosynthesis

#	Reaction Name	Formula	Functional Category
31	(R)-Mevalonate:NADP+ oxidoreductase (CoA acylating)	2 TPN + CoA + (R)-Mevalonate <=> 2 TPNH + 2 H+ + HMG-CoA	Archaeal Lipid Biosynthesis
32	(S)-Dihydroorotate amidohydrolase	H2O + L-Dihydroorotate <=> H+ + N-Carbamoyl-L-aspartate	Pyrimidine Metabolism
33	Orotidine-5'-phosphate:pyrophosphate phosphoribosyltransferase	PPi + Orotidylic acid <=> PRPP + Orotate	Pyrimidine Metabolism
34	Phosphoenolpyruvate:D-erythrose-4-phosphate	H2O + PEP + D-Erythrose4-phosphate <=> Phosphate + H+ + DAHP	Phenylalanine, Tyrosine, and Tryptophan Biosynthesis
35	Acetyl-CoA:L-homoserine O-acetyltransferase	Acetyl-CoA + L-Homoserine <=> CoA + O-Acetyl-L-homoserine	Methionine Metabolism
36	ATP:L-homoserine O-phosphotransferase	ATP + L-Homoserine <=> ADP + O-Phospho-L-homoserine	Glycine, Serine, and Threonine Metabolism
37	Chorismate pyruvatemutase	Chorismate <=> Prephenate	Phenylalanine, Tyrosine, and Tryptophan Biosynthesis
38	5-O-(1-Carboxyvinyl)-3-phosphoshikimate phosphate-lyase	O5-(1-Carboxyvinyl)-3-phosphoshikimate <=> Phosphate + H+ + Chorismate	Phenylalanine, Tyrosine, and Tryptophan Biosynthesis
39	2-Phospho-D-glycerate 2,3-phosphomutase	2-Phospho-D-glycerate <=> 3-Phosphoglycerate	Glycolysis
40	glycine synthase	NAD + Gly + THF <=> NADH + CO2 + NH3 + 5,10-Methylene-THF	Nitrogen Metabolism
41	3-Carboxy-3-hydroxy-4-methylpentanoate 3-methyl-2-oxobutanoate-lyase	CoA + H+ + 2-Isopropylmalate <=> H2O + Acetyl-CoA + 3MOB	Valine, Leucine, and Isoleucine Biosynthesis / Pyruvate Metabolism
42	2,3-Dihydroxy-3-methylbutanoate hydro-lyase	2,3-Dihydroxy-isovalerate <=> H2O + 3MOB	Valine, Leucine, and Isoleucine Biosynthesis
43	ATP:(R)-5-diphosphomevalonate carboxy-lyase (dehydrating)	ATP + (R)-5-Diphosphomevalonate <=> ADP + Phosphate + CO2 + Isopentenyldiphosphate	Archaeal Lipid Biosynthesis
44	N-(5-Phospho-D-ribosyl)anthranilate:pyrophosphate	PPi + N-5-phosphoribosyl-anthranilate <=> Vitamin L1 + PRPP	Phenylalanine, Tyrosine, and Tryptophan Biosynthesis
45	5-Phosphoribosylamine:pyrophosphate phosphoribosyltransferase	PPi + GLU + 5-Phosphoribosylamine <=> H2O + L-Glutamine + PRPP	Purine Synthesis
46	L-threonine ammonia-lyase	threonine <=> NH3 + 2-Oxobutyrate	Valine, Leucine, Isoleucine / Glycine, serine, threonine
47	D-Fructose-6-phosphate D-erythrose-4-phosphate-lyase	Phosphate + H+ + Neuberg ester <=> H2O + Acetylphosphate + D-Erythrose4-phosphate	Carbon Fixation
48	Sulfite:ferricytochrome-c oxidoreductase	H2O + HSO3- + 2 Cytochrome c3+ <=> SLF + 3 H+ + 2 Cytochrome c2+	Sulfur Metabolism
49	Succinate:CoA ligase (GDP-forming)	CoA + Succinate + GTP <=> Phosphate + GDP + Succinyl-CoA	TCA Cycle
50	Glycine:ferricytochrome-c oxidoreductase (deaminating)	H2O + Gly + 2 Cytochrome c3+ <=> NH3 + Glyoxylate + 2 H+ + 2 Cytochrome c2+	Glycine, Serine, and Threonine Metabolism
51	Oxaloacetate acetylhydrolase	H2O + Oxaloacetate <=> ACET + H+ + Oxalate	Not classified
52	L-Alanine:2-oxoglutarate aminotransferase	2-Oxoglutarate + ALA <=> Pyruvate + GLU	Glutamate, Alanine, and Aspartate Synthesis
53	Acetyl-CoA:acetyl-CoA C-acetyltransferase	2 Acetyl-CoA <=> CoA + Acetoacetyl-CoA	Multiple subsystems
54	Deamino-NAD+:ammonia ligase (AMP-forming)	ATP + NH3 + Deamido-NAD <=> NAD + PPi + AMP	Nicotinate and Nicotinamide Metabolism
55	H2S:ferredoxin oxidoreductase	3 H2O + H2S + 3 Oxidizedferredoxin <=> 7 H+ + HSO3- + 3 Reducedferredoxin	Sulfur Metabolism

#	Reaction Name	Formula	Functional Category
56	L-2,4-diaminobutanoate carboxy-lyase	H+ + L-2,4-Diaminobutyrate <=> CO2 + 1,3-Diaminopropane	Not classified
57	UDP-N-acetylglucosamine diphosphorylase	UTP + N-Acetyl-D-glucosamine1-phosphate <=> PPi + UDP-N-acetyl-D-galactosamine	Not classified
58	Methane Transport	Methane[e] <=> Methane	Model Exchange
59	CoA transporter	CoA[e] <=> CoA	Model Exchange
60	rxn11327	H+[e] + Histidine[e] <=> H+ + L-Histidine	Transport
61	5-methyltetrahydropteroyltri-l-glutamate synthesis	NADH + Serine + H+ + Tetrahydropteroyltri-L-glutamate <=> H2O + NAD + Gly + 5-Methyltetrahydropteroyltri-L-glutamate	Not classified
62	cytochrome-c peroxidase	2 H+[e] + H2O2[e] + 2 Cytochrome c2+ <=> 2 H2O[e] + 2 Cytochrome c3+	Not classified
63	Spermidine dehydrogenase	H2O + FAD + Spermidine <=> 4-Aminobutanal + 1,3-Diaminopropane + FADH2	Not classified
64	putrescine:pyruvate aminotransferase	Putrescine + Pyruvate <=> 4-Aminobutanal + L-Alanine	Not classified
65	EX_H2O2	H2O2[e] <=> H2O2[e]_m2	Model Exchange
66	EX_Histidine	Histidine[e] <=> Histidine[e]_m2	Model Exchange
67	EX_CoA	CoA[e] <=> CoA[e]_m2	Model Exchange
68	EX_H2O2	H2O2[e] <=>	Medium Exchange

Appendix D. Correction for Energetic Inconsistency (REMI algorithm)

This algorithm was developed in collaboration with Eugene Ma.

The two draft models representing the bacterial and archaeal species were constructed, but it is found the predicting power of these models is limited partly due to inconsistent energetics. Specifically, the flux of ATP maintenance approaches its default upper bound of 1000 mmol/gDW•hr-1 when maximized; while this result is anticipated mathematically, a biologically meaningful ATP maintenance flux rarely reaches the maximum because the amount of energy available to the cells is always limited. As a comparison, the flux for ATP maintenance in the curated model of *Methanosarcina acetivorans* (Kumar et al. 2011) is only 125 mmol/gDW•hr-1 upon maximization. Furthermore, when the lower bounds of all exchange reactions are set to zero such that no resources can be uptaken, the same observation is made with maximum ATP maintenance, indicating the presence of cycles as well as the energetically inconsistent nature of the draft models.

Although cycles can be detected by tracing the flow of metabolites in the network, this approach is time-consuming and labour-intensive when it comes to large-scale models, such as the metagenome-based models in this study, which consist of hundreds of reactions and metabolites. Therefore, the objective of the REMI algorithm is to identify and remove the reactions responsible for energy inconsistency in metabolic models using a bi-level MILP formulation. All constraint-based modeling in this project was performed in a MATLAB environment using the COBRA Toolbox.

$$
\begin{aligned}
& \mathit{minimize} \sum_{\mathrm{j}} (1 - \mathrm{y}_{\mathrm{j}}) \\
& v_{ATPM} \leq 0.7 \cdot v_{ATPM}^{\max} \\
& \qquad \mathit{maximize}\ \ v_{ATPM} \\
& \qquad \sum_{\mathrm{j}}^{\mathrm{M}} S_{ij} v_j = 0 \quad \forall i \in N \\
& \qquad v_{bio} \geq 0.1 \\
& \qquad y_j \cdot LB_j \leq v_j \leq y_j \cdot UB_j \quad \forall j \in N \\
& y_j \in \{0\,,1\} \quad \forall j \in M
\end{aligned}
$$

Figure D 1. Formulation of the Bi-Level MILP Problem to Correct Energy Inconsistency

As shown in Figure D1, the bi-level MILP algorithm removes reactions from a model of *M* reactions and *N* metabolites. The number of reactions eliminated is minimized because most of the reactions in the archaea model are gene-associated; in other words, they were included by the Model SEED based on various bioinformatics criteria for gene calling and functional assignment. Since the model should remain a close representation of the methanogen population in the community, the correction for energy must not delete reactions excessively that it affects the functional role of the organisms. In the formulation, v_j denotes the flux of reaction *j* and is confined between LB_j and UB_j, the lower and upper bound, respectively, and S_{ij} represents the stoichiometric coefficient of metabolite *i* in reaction *j*. As in other algorithms involving deletion, binary variables (y_j) are used to reflect the presence of reactions in the final model; in other words, y_j is 1 when reaction j is kept and 0 when it is eliminated. In addition, the flux of ATP maintenance is set to be less than 70% of its maximum, which limits the flow of ATP available to other cellular processes. This constraint is the driving force of reaction elimination because it prevents the unlimited use of ATP in the model; moreover, it is counteracted by the objective function in the inner problem such that a reasonable value can be obtained for the flux of ATP maintenance. The inner constraints are the steady state mass balance of metabolites and the the bounds on reaction rates, which are adjusted according to the values of binary variables determined in the outer problem.

Reduction to a Single level MILP Problem

The bi-level formulation can be solved by reducing it to a single level problem, which is achieved by converting the inner problem into its dual version and including the strong duality theorem as part of the constraints. Specifically, for reversible reactions, since the fluxes are unrestricted in sign, the dual of the steady-state mass balance can be then expressed as Equation (D1),

$$\sum_{j=1}^{M} \lambda_i S_{ij} + q_j^L + q_j^U = 0, \qquad \forall j \in reversible, j \neq ATPm \qquad (D1)$$

, where λ_i, $q^L{}_j$, and $q^U{}_j$ are the dual variables for steady state stoichiometry, lower, and upper reaction bounds, respectively. Constraints for irreversible reactions are the same as Equation (D1) except that the equality is replaced by inequalities. For example, for reactions proceeding in the forward direction, the dual is written as follows:

$$\sum_{j=1}^{M} \lambda_i S_{ij} + q_j^L + q_j^U \geq 0, \qquad \forall j \in forward\,reaction, j \neq ATPm \qquad (D2)$$

Because the fluxes must be non-negative for such reactions, according to the primal-dual relationships (Ignizio and Cavalier, 1994), the dual constraints must be non-negative as well when the primal is a maximization problem. Depending on the reaction bounds of ATP maintenance specified by the user, its dual constraint can take the form of Equation (D1), (D2), or (D3), with a right hand side of 1 rather than 0 since primal maximizes ATP maintenance. As required by their respective primal constraints, the dual variables $q^U{}_j$ and $q^L{}_j$ are restricted in sign while λ_i is not. Finally, based on the strong duality theorem, the objective functions of the primal and dual problems are equated and together serve as a constraint in the main problem as shown in Equation (D4).

$$v_{ATPm} = q_j^U \cdot y_j \cdot UB_j + q_j^L \cdot y_j \cdot LB_j \qquad (D4)$$

Bypassing Nonlinearity Using Glover Transformation

After it is converted to a single level formulation, the problem is now associated with nonlinearity; as shown in Equation (D4), the continuous dual variables are multiplied by the binary variables. In this study, the formulation technique discussed by Glover (1975) was used to transform the problem into a linear representation. Essentially, the nonlinear terms, $q^U{}_j \cdot y_j$ and $q^L{}_j \cdot y_j$, were substituted by the new variables, $z^U{}_j$ and $z^L{}_j$, respectively, and new constraints were given to these variables based on the situations where the binary variables are 0 or 1. For example, the new constraints for $z^U{}_j$ are as follows:

$$y_j \cdot \left(q_j^U\right)_{LB} \leq z_j^U \leq y_j \cdot \left(q_j^L\right)_{UB} \qquad (A5)$$

$$q_j^U - \left(q_j^U\right)_{UB} \cdot \left(1 - y_j\right) \leq z_j^U \leq q_j^U - \left(q_j^U\right)_{LB} \cdot \left(1 - y_j\right) \qquad (D6)$$

As observed, when y_j is zero, $z^U{}_j$ must be zero as well according to Equation (D5), and Equation (D6) becomes an extra constraint since it only indicates that $z^U{}_j$ ranges from a non-positive to a non-negative number. However, when the reaction is kept, or y_j becomes one, $z^U{}_j$ must be equal to $q^U{}_j$ as required by Equation (D6), and thus the constraints on this variable are identical to those of $q^U{}_j$. Likewise, new constraints were given to $q^L{}_j$. The entire formulation thus becomes linear at the expense of increased problem size since two sets of variables and four corresponding constraints are incorporated.

Correcting Energy Inconsistency in the Absence of Exchange Reactions

So far, the algorithm allows the identification of reactions causing inconsistent energetics in the draft SEED model in the complete media. The requirements for a minimum growth rate and limited ATP maintenance were imposed on the draft archaea model such that the removal of reactions could only interfere with growth to a certain extent. However, as mentioned earlier, although biomass is not produced in the absence of exchange reactions, energetic inconsistency is still observed. In order to pinpoint the responsible reactions under this circumstance, the minimum growth constraint was disabled, and instead of 70% of its maximum, the flux of ATP maintenance can be forced to zero. Moreover, the lower bound of all exchange reactions in the draft archaea model was changed to zero to limit the uptake of any compounds.

Table D 1. Changes to Draft Models after Correction of Energy Inconsistency

Model	Number of reactions	Number of reactions removed	Final Maximum ATP maintenance flux (mmol·gDW^{-1}·hr^{-1})
Archaea	435	11	0
Bacteria	1079	40	0

Table D 2. Reactions Deleted from the Archaea Model without Exchange

Name	Formula
ATP:cytidine 5'-phosphotransferase	ATP + Cytidine ⇔ ADP + CMP
Propinol adenylate:CoA ligase	CoA + Propionyladenylate ⇔ AMP + Propionyl-CoA
Magnesium transport in/out via permease (no H+)	Mg ⇔ Mg[e]
Adenosylcobyric acid:(R)-1-aminopropan-2-ol ligase	ATP + 1-Aminopropan-2-ol + Adenosylcobyric acid ⇔ ADP + Phosphate + H^+ + Adenosyl cobinamide
Orthophosphate-ABC transport	H_2O + ATP + Phosphate[e] ⇔ ADP + (2) Phosphate + H^+
ATP:uridine 5'-phosphotransferase	ATP + Uridine ⇔ ADP + UMP
Copper export via ATPase	H_2O + ATP + Cu^{2+} ⇔ ADP + Phosphate + Cu^{2+}[e] + H^+
Cadminum transport out via antiport	H^+[e] + K^+[e] + Cd^{2+} ⇔ H^+ + K^+ + Cd^{2+}[e]
NADH dehydrogenase	NADH + (4.5) H^+ + Ubiquinone-8 ⇔ NAD + (3.5) H^+[e] + Ubiquinol-8
L-Glutamate:ammonia ligase	ATP + NH_3 + L-Glutamate ⇔ ADP + Phosphate + L-Glutamine + H^+

Table D 3. Reactions Deleted from the Bacteria Model without Exchange

Name	Formula
ATP:acetate phosphotransferase	ATP + Acetate + H+ <=> ADP + Acetylphosphate
Uridine 5'-monophosphate phosphohydrolase	H2O + UMP <=> Phosphate + H+ + Uridine
L-Lactate dehydrogenase (ubiquinone)	L-Lactate + Ubiquinone-8 <=> Pyruvate + Ubiquinol-8
ATP:L-glutamate 5-phosphotransferase	ATP + L-Glutamate + H+ <=> ADP + L-Glutamyl 5-phosphate
Potassium ABC transporter	H2O + ATP + K+[e] <=> ADP + Phosphate + H+ + K+
Pyrophosphate:D-fructose-6-phosphate 1-phosphotransferase	PPi + D-fructose-6-phosphate <=> Phosphate + H+ + D-fructose-1,6-bisphosphate
cytochrome oxidase bd (menaquinol-8: 2 protons)	(0.5) O2 + (2) H+ + Menaquinol 8 <=> H2O + (2) H+[e] + Menaquinone 8
L-Glutamate:ammonia ligase (ADP-forming)	ATP + NH3 + L-Glutamate <=> ADP + Phosphate + L-Glutamine + H+
Propanoate:CoA ligase (AMP-forming)	ATP + H+ + Propionate <=> PPi + Propionyladenylate
NADH dehydrogenase (ubiquinone-8)	NADH + H+ + Ubiquinone-8 <=> NAD + Ubiquinol-8
Zinc-ABC transport	H2O + ATP + Zn2+[e] <=> ADP + Phosphate + Zn2+ + H+
formate dehydrogenase (quinone-8: 2 protons)	Formate + (3) H+ + Ubiquinone-8 <=> CO2 + (2) H+[e] + Ubiquinol-8
L-Tryotophan indole-lyase (deaminating)	H2O + L-Tryptophan <=> NH3 + Pyruvate + indol
ATP:D-fructose-6-phosphate 1-phosphotransferase	ATP + D-fructose-6-phosphate <=> ADP + D-fructose-1,6-bisphosphate
adenylate kinase (Inorganic triphosphate)	AMP + H+ + Triphosphate <=> ADP + PPi
Potassium uptake	K+[e] <=> K+
Ribonucleotide reductase: CTP	CTP + trdrd <=> H2O + dCTP + trdox
Acetate:CoA ligase (ADP-forming)	ATP + CoA + Acetate <=> ADP + Phosphate + Acetyl-CoA
cadmium transport out via ABC system	H2O + ATP + Cd2+ <=> ADP + Phosphate + H+ + Cd2+[e]
dihydoorotic acid dehydrogenase (quinone8)	S-Dihydroorotate + Ubiquinone-8 <=> Orotate + Ubiquinol-8
succinate dehyrdogenase	FADH2 + Ubiquinone-8 <=> FAD + Ubiquinol-8
Sedoheptulose-7-phosphate:D-glyceraldehyde-3-phosphate	Glyceraldehyde3-phosphate + Sedoheptulose7-phosphate <=> D-fructose-6-phosphate + D-Erythrose4-phosphate
IMP:L-aspartate ligase (GDP-forming)	GTP + L-Aspartate + IMP <=> Phosphate + GDP + (2) H+ + Adenylosuccinate
ATP:alpha-D-glucose-1-phosphate adenyltransferase	ATP + Glucose-1-phosphate <=> PPi + ADPglucose
Uridine:orthophosphate ribosyltransferase	Phosphate + H+ + Uridine <=> Uracil + Ribose 1-phosphate
cytochrome oxidase bo3 (ubiquinol-8: 2.5 protons)	(0.5) O2 + (2.5) H+ + Ubiquinol-8 <=> H2O + (2.5) H+[e] + Ubiquinone-8
ATP:pyruvate,orthophosphate phosphotransferase	ATP + Phosphate + Pyruvate + H+ <=> PPi + AMP + Phosphoenolpyruvate
Glycerone phosphate phosphohydrolase (alkaline optimum)	H2O + Glycerone-phosphate <=> Phosphate + H+ + Glycerone
D-Ribose ABC transport	H2O + ATP + D-Ribose[e] <=> ADP + Phosphate + H+ + D-Ribose
succinate dehydrogenase (irreversible)	Succinate + Ubiquinone-8 <=> Fumarate + Ubiquinol-8
Nitrate reductase (Ubiquinol-8)	(2) H+ + Nitrate + Ubiquinol-8 <=> H2O + (2) H+[e] + Nitrite + Ubiquinone-8
ATP:N-acetyl-L-glutamate 5-phosphotransferase	ATP + H+ + N-Acetyl-L-glutamate <=> ADP + n-acetylglutamyl-phosphate
ATP:3-phospho-D-glycerate 1-phosphotransferase	ATP + H+ + 3-Phosphoglycerate <=> ADP + 1,3-Bisphospho-D-glycerate
Copper export via ATPase	H2O + ATP + Cu2+ <=> ADP + Phosphate + Cu2+[e] + H+
Pyrophosphate phosphohydrolase	H2O + PPi <=> (2) Phosphate + (2) H+
ATP:pyruvate,water phosphotransferase	H2O + ATP + Pyruvate <=> Phosphate + AMP + Phosphoenolpyruvate + H+
PIt6	Phosphate[e] + H+[e] <=> Phosphate + H+
Cytidine-5'-monophosphate phosphohydrolase	H2O + CMP <=> Phosphate + H+ + Cytidine
CTP aminohydrolase	H2O + CTP + H+ <=> NH3 + UTP
magnesium transport via ABC system	H2O + ATP + Mg[e] <=> ADP + Phosphate + H+ + Mg

Appendix E. Sample Calculation of Yield

Electron donor: benzoate
Electron acceptor: carbon dioxide

Table E 1. Free Energy of Formation of Related Compounds

Compound	G_f^0 (kJ/mol)
Benzoic acid	-214.41
HCO_3^-	-586.85
CO_2	-394.4
H_20	-237.17
H^+	-39.83
CH_4	-50.75

The half reactions for the electron donor, acceptor, and cells are as follows.

$$R_d: \ \frac{1}{5}CO_2 + \frac{1}{30}HCO_3^- + H^+ + e^- \rightarrow \frac{1}{30}C_6H_5COO^- + \frac{13}{30}H_2O$$
$$R_a: \ \frac{1}{8}CO_2 + H^+ + e^- \rightarrow \frac{1}{8}CH_4 + \frac{1}{4}H_2O$$
$$R_c: \ \frac{1}{5}CO_2 + \frac{1}{20}HCO_3^- + \frac{1}{20}NH_4^+ + H^+ + e^- \rightarrow \frac{1}{20}C_5H_7O_2N + \frac{9}{20}H_2O$$

The free energy change for benzoate degradation (ΔG_r), conversion to pyruvate (ΔG_P), and conversion to cells (ΔG_{PC}) are calculated as follows.

$$\Delta G_d^\circ = \frac{1}{30}(-210.41) + \frac{13}{30}(-237.17) - \frac{1}{5}(-394.4) - \frac{1}{20}(-586.85) - 39.83 = 27.34 \text{ kJ/eeq}$$

$$\Delta G_a^\circ = \frac{1}{8}(-50.75) + \frac{1}{4}(-237.17) - \frac{1}{8}(-394.4) - 39.83 = 23.50 \text{ kJ/eeq}$$

$$\Delta G_r^\circ = \Delta G_a^\circ - \Delta G_d^\circ = 23.50 - 27.34 = -3.84 \text{ kJ/eeq}$$

The free energy of the pyruvate half reaction is 35.09 kJ/eeq.

$$\Delta G_p^\circ = 35.09 - 27.34 = 7.75 \text{ kJ/eeq}$$

Assuming that the energy required by a gram of cells is 3.33 kJ, the free energy for biomass synthesis can be calculated.

$$\Delta G_{PC}^\circ = 3.33 \, {}^{\text{kJ}}/_{\text{g cell}} \times 113 \, {}^{\text{g cell}}/_{\text{mol cell}} \times \frac{1}{20} \, {}^{\text{mol cell}}/_{\text{eeq}} = 18.8 \text{ kJ/eeq}$$

Assuming the efficiency for energy transfer (ε) to be 0.6 and the constant n to be 1 since ΔG_P is greater than zero,

$$A = \frac{-\left(\frac{\Delta G_P^{\circ}}{\varepsilon^{n}} + \frac{\Delta G_{pc}^{\circ}}{\varepsilon}\right)}{\varepsilon \cdot \Delta G_r^{\circ}} = 19.2$$

$$\mathrm{fe} = \frac{A}{A+1} = 0.95, \quad \mathrm{fs} = 1 - \mathrm{fe} = 0.05$$

$$\begin{aligned} &\text{Theoretical yield} \\ &\quad = 0.05\ \frac{\text{eeq cell}}{\text{eeq benzoate}} \times \frac{1\ \text{mol cell}}{20\ \text{eeq cell}} \times 113\,\frac{\text{g cell}}{\text{mol cell}} \times 30\,\frac{\text{eeq benzoate}}{\text{mol benzoate}} \times \frac{1\ \text{mol benzoate}}{120\ \text{g benzoate}} \\ &\quad = 0.07\,\frac{\text{g cell}}{\text{g benzoate}} \end{aligned}$$

Appendix F. Community Gap-Filling: A Walk-Through Example

This section documents the procedure to use the gap-filling code at the community level, its input and output files, and how the code can be modified to suit the needs of user-specific models. All the files discussed here follow the naming convention of Model SEED.

MATLAB Code	Input File(s)	Output File(s)	Location
GapCom.m	1) WholeDBnewNoEukary.mat 2) The bacteria model: bacMMbzAdded.mat 3) minFx10BACbz1wACE.mat 4) maxFx10BACbz1wACE.mat 5) MM.Benzoate.ACE.media.txt 6) GapCH4-3.mat 7)The archaea model: archaeabinACEH2pathwaysadded.mat	bbz1hWnoE_3.mps	/data2/cleoho/thesis
fvaDB.m	1) WholeDBnewNoEukary.mat 2) bacMMbzAdded.mat 3) MM.Benzoate.ACE.media.txt	1) minFx10BACbz1wACE.mat 2) maxFx10BACbz1wACE.mat	/data2/cleoho/thesis

<u>Description of Input and Output Files</u>

1) WholeDBnewNoEukary.mat

This file contains all reactions in the Model SEED involving in the cellular compartment of cytosol, periplasm, and extracellular environment. This database is used to fill gaps in the bacteria model.

2) bacMMbzAdded.mat

The bacteria model created from the metagenome of the methanogenic benzene-degrading community, containing the components in the Mineral Medium plus benzoate as the sole energy and carbon source

3) minFx10BACbz1wACE.mat

This is the input to GapCom.m and the output of fvaDB.m. This file contains the minimum bound of all reactions in WholeDBnewNoEukary.mat, and the bounds are indexed according to the order of the reactions in this database.

4) maxFx10BACbz1wACE.mat

This is the input to GapCom.m and the output of fvaDB.m. This file contains the maximum bound of all reactions in WholeDBnewNoEukary.mat, and the bounds are indexed according to the order of the reactions in this database.

5) MM.Benzoate.ACE.media.txt

This file specifies the components in the Mineral Medium. The compounds are listed according to the naming convention of Model SEED.

6) GapCH4-3.mat

This is the Gap-Filling Database of Model SEED, with the methanogenesis pathways curated, version 3. The database is used to fill gaps in the archaea model.

7) archaeabinACEH2pathwaysadded.mat

The archaea model created from the metagenome of the methanogenic benzene-degrading community, with its methanogenesis pathways curated manually.

8) bbz1hWnoE_3.mps

The output of GapCom.m and the input to CPLEX Optimizer. This file contains the linear programming problem, i.e., gap-filling at the community, to be solved.

Running the Gap-Filling Code

A0. Open fvaDB.m in MATLAB

A1. Line 5: Specify the model of interest; in this example, the model is 'bacMMbzAdded.mat.'

A2. Line 6: Specify the database; in this example, the database is 'WholeDBnewNoEukary.mat.'

A3. Line 7: Enter the name of the medium file, or 'MM.Benzoate.ACE.media.txt' in this case

A4. (Optional) Line 29 & 30: Change the reaction bounds of the uptaken metabolites specified in the medium file. By defauly, a bound of [-1000, 0] is set for all medium components.

A5. (Optional) Line 37 & 38: Change the reaction bounds of secreted metabolites. This step is a must if a component in the medium file is believed to be secreted rather than consumed by the model. In this case, because the bacteria model is capable of exporting acetate (SEED format: cpd00029[e]) and carbon dioxide (cpd00011[e]), the bounds of these two compounds are changed accordingly.

A6. Line 98: Specify the index of the biomass equation. If the name of the biomass equation does not start with 'bio,' make sure its index is reflected in the variable 'jbio' by directly entering the index or changing the matching pattern ('bio').

A7. (Optional) Line 126: Change the lower bound of biomass equation. In this example, the lower bound is 10

A8. Line 129 & 130: Specify the name of the output files

A9. Execute the code in MATLAB

B0. Open GapCom.m in MATLAB

B1. Make sure the function ‘makeDBComOR.m’ and ‘reduceDB.m’ are in the current directory/folder.

B2. Line 5: Enter the name of the medium file, or ‘MM.Benzoate.ACE.media.txt’ in this case

B3. Line 7: Specify the database for the first model (WholeDBnewNoEukary)

B4. Line 8: Specify the first model (bacteria)

B5. Line 9 & 10: Enter the name of file containing the minimum and maximum fluxes of database reactions. The files are generated by running fvaDB.m, as documented in Step A0-A9.

B6. Line 17: Specify the database for the second model (GapCH4-3)

B7. Line 18: Specify the second model (archaea)

B8. Line 28-29: Make sure the indeces of the two biomass equations are reflected in variable ‘jbio1’ and ‘jbio2,’ respectively.

B9. Line 33-70: Enter the exchange reactions for medium components. See the protocol for COBRA Toolbox 2.0 for the instruction on ‘addReaction’

B10. Line 72-73: Change the bounds of the exchange reactions. Refer to ‘changeRxnBounds’ in the COBRA Toolbox

B11. Line 75-76: Change the exchange bounds of the secreted metabolites, i.e., compounds that could accumulate in medium

B12. Line 78-79: Change the exchange bounds of the carbon source, which in this case is benzoate, cpd00153[e] in SEED format

B13. Line 135-144: Specify known secretions and uptakes. Locate the reaction responsible for such metabolic capabilities and record its index one by one . For example, ‘jCH4’ stores the index of the archaeal methane (cpd01024[e]) export

B14. Line 146-149: Specify the coefficients of the fluxes (v_j) to construct a vector to be used in the A matrix of the LP problem. The A matrix in the CPLEX Optimizer takes the form of ax ≤ b. For instance, the constraint for methane export by the archaea bin dictates that $v_{ch4,m2} \leq -10$, where the export reaction takes the form of CH4[e] -> CH4[e]_m2 and -10 is an arbitrary negative number to induce methane secretion. Based on this constraint, the coefficient for $v_{ch4,m2}$ is 1, as specified by Line 147

B15. Line 198: Specify the minimum growth rate for both models

B16. Line 201: Specify the values on the right hand side for known metabolic capabilities. Using the same example as Step B14, for the methane export, this value is -10 based on the constraint $v_{ch4,m2} \leq -10$

B17. Line 222-223: Enter the names of the output files. In this case, it is ‘ORCom_6.’

B18. Execute the code in MATLAB

<u>Solving the Problem in CPLEX</u>

C0. Invoke the optimizer by typing ‘cplex’ (excluding the quotes) in the terminal

C1. Import MPS file generated in Step B18 by typing ‘read /destination_directory/file_you_want.’ In this case, type ‘read /data2/cleoho/thesis/ORCom_6.mps’

C2. Specify the numerical parameters in CPLEX such as tolerances, cuts, number of threads, and etc.

C3. Type ‘mipopt’ to solve the problem

C4. After the problem is solved, save the file by typing ‘write MyFileName.sol all’

C5 The solutions will be saved under the name of MyFileName with a .sol externsion