Chapter 5

Reconstruction and Analysis of Central Metabolism in Microbes

Janaka N. Edirisinghe, José P. Faria, Nomi L. Harris, Benjamin H. Allen, and Christopher S. Henry

Abstract

Genome-scale metabolic models (GEMs) generated from automated reconstruction pipelines often lack accuracy due to the need for extensive gapfilling and the inference of periphery metabolic pathways based on lower-confidence annotations. The central carbon pathways and electron transport chains are among the most well-understood regions of microbial metabolism, and these pathways contribute significantly toward defining cellular behavior and growth conditions. Thus, it is often useful to construct a simplified core metabolic model (CMM) that is comprised of only the high-confidence central pathways. In this chapter, we discuss methods for producing core metabolic models (CMM) based on genome annotations. With its reduced scope compared to GEMs, CMM reconstruction focuses on accurate representation of the central metabolic pathways related to energy biosynthesis and accurate energy yield predictions. We demonstrate the reconstruction and analysis of CMMs using the DOE Systems Biology Knowledgebase (KBase). The complete workflow is available at http://kbase.us/core-models/.

Key words Central metabolism, Core metabolic models, Metabolic model reconstruction, Flux balance analysis, Biochemical pathways, Model comparison

1 Introduction

Central carbon metabolism is a key component in the metabolic network of living organisms as these pathways harbor many of the most important mechanisms for energy biosynthesis, as well as producing the precursor compounds for most essential biomass building blocks. The energy production strategies defined in the central metabolic pathways have a significant impact on the behavior and growth conditions of microorganisms, thus playing a crucial role in the quantitative prediction of biomass and energy yields [1, 2]. Energy production strategies in microbes are highly diversified, unlike those in higher eukaryotes. These strategies primarily depend on environmental factors such as: (1) carbon source

Marco Fondi (ed.), *Metabolic Network Reconstruction and Modeling: Methods and Protocols*, Methods in Molecular Biology, vol. 1716, https://doi.org/10.1007/978-1-4939-7528-0_5,

utilization; (2) ability to respire by reducing numerous electron acceptors; and (3) fermentation capabilities.

It continues to be challenging to make accurate computational predictions based on metabolic models and in silico simulations interpreting complex microbial behavior. Tools for automated metabolic model reconstruction such as ModelSEED [3–5] can rapidly generate draft genome-scale metabolic models from annotated genome sequences [6]. However, these draft models, and in some cases even curated published models, can lack accuracy in predicting growth yields, ATP production yields, and central carbon flux profiles. This poor accuracy stems primarily from three common problems: (1) poor representation of energy biosynthesis pathways; (2) a lack of diverse electron transport chain (ETC) variations; and (3) addition of extensive gapfilling reactions that can sometimes misrepresent an organism's behavior [7].

Many of these problems can be avoided by using a simplified model comprised of only the most confidently annotated and biologically critical pathways for energy biosynthesis [8] (Fig. 1). We define these models as Core Metabolic Models (CMM), and they consist primarily of the sugar oxidation pathways, the fermentation pathways (Fig. 2), and the ETC variations. We previously developed an approach for the reconstruction and analysis of CMMs based on annotated genome sequences [9], which we implemented as a pipeline in the DOE Systems Biology Knowledgebase (KBase). In this chapter, we demonstrate how this analysis workflow can be run in KBase. The complete workflow, including example data and commentary, can be accessed from http://kbase.us/core-models. The pipeline is comprised of four main steps: (1) genome annotation by RAST [10]; (2) CMM reconstruction [9]; (3) gapfilling [7]; and (4) flux balance analysis (FBA) [11]. We also discuss methods for exploring metabolic diversity by studying the variations in central metabolic pathways in a phylogenetic context.

2 Materials

In this section, we describe the data and tools required to build CMMs using the KBase Narrative Interface (https://narrative.kbase.us). Methods that use the data and tools listed in this section are described in detail in Subheading 3.

2.1 KBase Narrative Interface

In KBase, reproducible workflows called *Narratives* can be created and shared. Narratives can include data, analysis steps, results, visualizations, and commentary. Narratives can be shared with collaborators as "active papers" that let others repeat the analysis workflows and even alter parameters or input data to achieve different or improved results. We encourage readers to view and copy the Core Model Construction Narrative (see http://kbase.us/core-

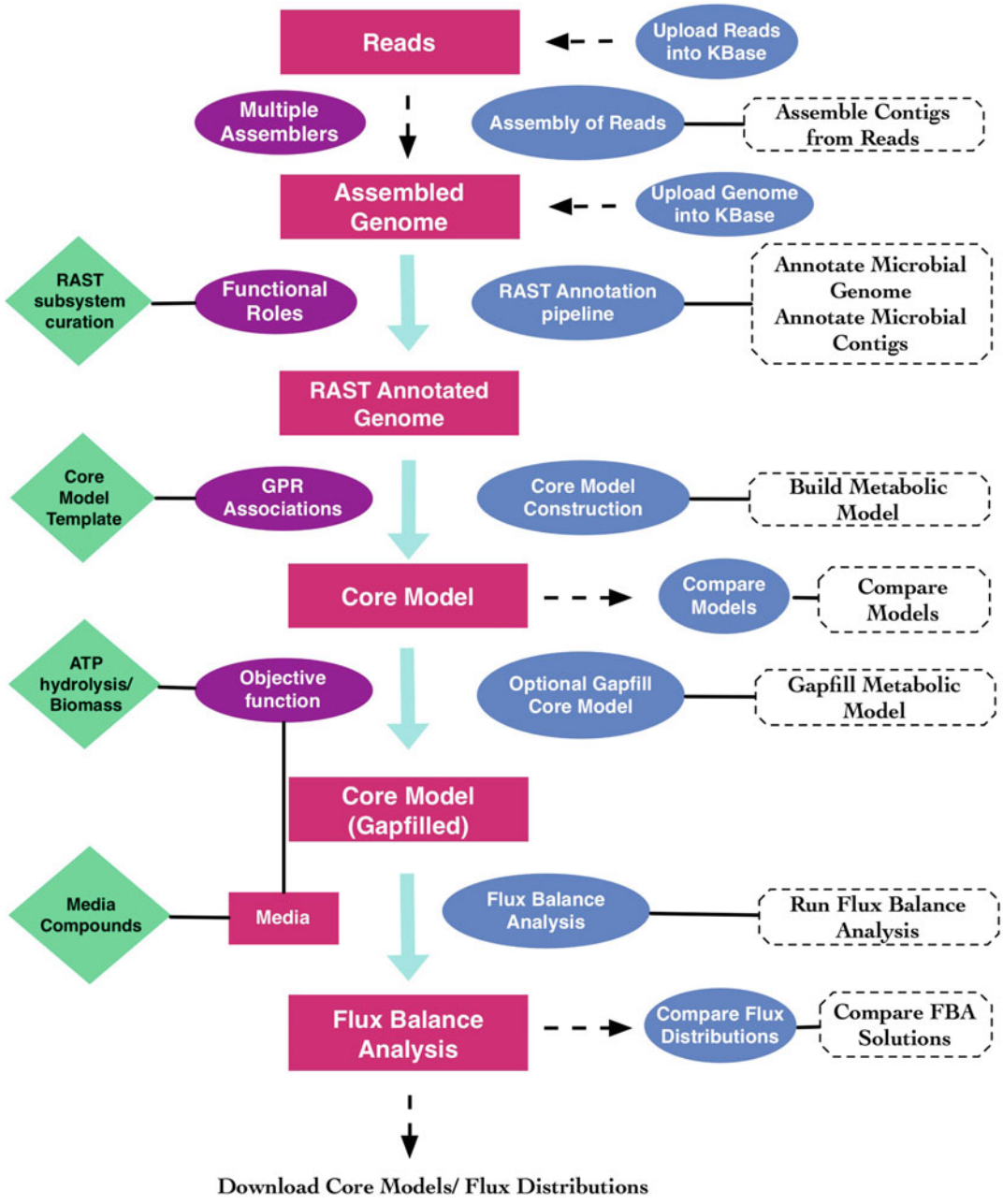

Fig. 1 Seven-step pipeline used to construct and analyze core metabolic models in KBase. The core model reconstruction pipeline is comprised of seven apps (rounded rectangles with dashed borders), which operate on specific data types (magenta rectangles). These apps are driven by several curated reference data sources (green diamonds), including RAST subsystems, the template model pathways, the template model objective functions, and the compounds that make media formulations used for gapfilling and FBA. The purple ovals identify the essential components/data (explained in the text) required for apps; the blue ovals show the steps performed by the apps. Dashed arrows show optional steps while turquoise arrows show major steps of the pipeline. The resulting data can be downloaded as explained in **Note 1**

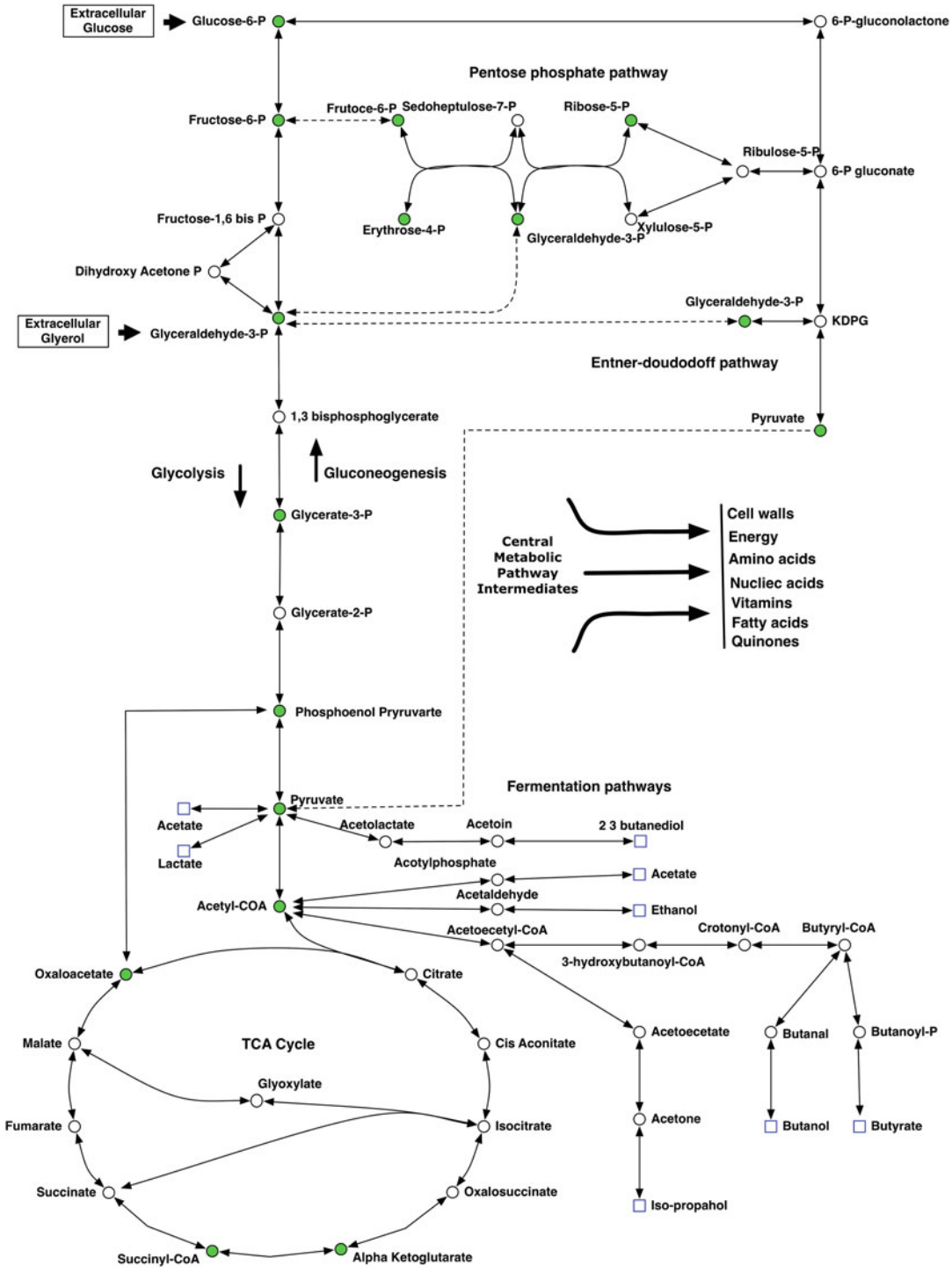

Fig. 2 Metabolic pathways comprising the core metabolic model. The core model template encompasses 12 central metabolic pathways including sugar oxidation (glycolysis, gluconeogenesis, Entner-Doudoroff, pentose phosphate), TCA cycle and fermentation pathways (fermentation end products displayed as squares with blue borders). These pathways produce 16 biomass precursor molecules (green circles) (Table 1)

models) and try running the steps on the example data or using their own data.

2.2 Core Metabolic Model Reconstruction Pipeline (Apps)

The CMM reconstruction pipeline in KBase permits a user to progress from raw genome sequencing reads through assembly and annotation to core model reconstruction and then perform model analysis and comparison. The pipeline is comprised of seven apps (centered on the four main steps previously mentioned), which are described in this section (Fig. 1).

The first step of the pipeline is the *Assemble Contigs from Reads* app, which accepts short reads from Next-Generation Sequencing (NGS) as input and produces assembled contigs as output. KBase includes numerous apps for genome assembly, but the *Assemble Contigs from Reads* app is the most sophisticated, as it enables users to run multiple assemblers at once, then aids in selecting the best set of contigs produced by all the assemblers.

The second step of the pipeline is the *Annotate Microbial Contigs* app. This app is based on the RAST (Rapid Annotations using Subsystems Technology) pipeline for microbial genome annotation [12]. The app accepts assembled contigs as input, performs gene calling using a combination of Glimmer [13] and Prodigal [14], then functionally annotates genes from the SEED subsystems ontology [15] using a kmer-based approach [16]. Alternatively, if one already has a genome with existing gene calls in GenBank format, the *Annotate Microbial Genome* app can be used to simply re-annotate the existing genome while keeping the gene calls intact. This app also uses the RAST approach for functional annotation. Note that when an existing genome is imported into KBase, its original annotations are kept intact. Unless the imported genome was generated by RAST or PATRIC [17], it is likely that the annotations do not conform to the SEED subsystems ontology. As a result, it is currently necessary to re-annotate these genomes using the *Annotate Microbial Genome* app prior to building a metabolic model. Both annotation apps produce an annotated genome as output, which includes data on all genome contigs, genes, proteins, and functional annotations.

The third step of the pipeline is the *Build Metabolic Model* app. In this app, the functional annotations generated by the RAST-based genome annotation apps are used to generate a draft metabolic model. A draft model consists of three parts: (1) a network of metabolic reactions (including both gene-associated reactions and spontaneous reactions); (2) a set of gene-protein-reaction (GPR) associations that dictate how each reaction activity depends on associated gene activity; and (3) a biomass composition reaction that defines the small molecule building blocks that comprise 1 g of biomass (e.g., amino acids, nucleotides, lipids, cofactors, cell-wall components, and energy). This app produces genome-scale metabolic models by default, but it is possible to select the core template

to build a core model instead. Core models contain far fewer reactions, and their biomass composition reaction uses only central-carbon precursor molecules. The reactions included in the models produced by the *Build Metabolic Model* app are selected from the ModelSEED [18] biochemistry database. This curated database contains mass and charge balanced reactions, standardized to aqueous conditions at neutral pH. The Model SEED reaction database integrates biochemistry from KEGG [19, 20], MetaCyc [21], EcoCyc [22], Plant BioCyc, Plant Metabolic Networks, and Gramene [23]. The database is available for download from GitHub (https://github.com/ModelSEED/ModelSEEDDatabase/blob/master/Biochemistry/).

The fourth step of the pipeline is the *Gapfill Metabolic Model* app. Draft metabolic models (built using the *Build Metabolic Model* app) usually have missing reactions (gaps) due to incomplete or incorrect functional genome annotations. As a result, these models are unable to produce biomass using media on which the organism typically is capable of growing. Gapfilling algorithms can overcome this problem by identifying the minimum number of new reactions that must be added to the model, or existing reactions that must be made reversible to enable the production of biomass. The gapfilling app in KBase uses Model SEED reaction database for gapfilling. It works equally well on core or genome-scale metabolic models. When gapfilling a core model, only the reactions present in the core model template (*see* Subheading 2.3) are considered for gapfilling. When gapfilling a genome-scale model, all 13,000 reactions from the ModelSEED [18] biochemistry database are considered for gapfilling.

The fifth step of the pipeline is the *Run Flux Balance Analysis* app. This app predicts the flow of metabolites through the metabolic network of an organism by optimizing for the selected cellular objective function, which is typically the production of biomass. Flux Balance Analysis (FBA) is a constraint-based approach that estimates growth-optimal fluxes through all the reactions in the metabolic network, thereby making it possible to estimate the growth rate of an organism (the rate of biomass production) or the rate of production of a given metabolic output on a specified media. This app makes it possible to analyze an organism's growth on different substrates and to evaluate the reactions and metabolites that carry fluxes in each growth condition. In addition to optimizing the biomass, one can choose to optimize a certain reaction (e.g., transporter reaction) so that the model optimizes to produce flux through that reaction. The *Run Flux Balance Analysis* app requires the user to specify a media formulation in which the growth will be simulated. In KBase, the media contains a list of the chemical compounds that are available for consumption in the flux simulation. KBase currently maintain more than 500 commonly used media conditions. In addition, users are able

to build and upload their own custom media formulations. The *Run Flux Balance Analysis* app includes a range of FBA algorithms, including flux variability analysis, gene essentiality prediction, and expression data analysis.

The sixth step of the pipeline is the *Compare Models* app. This app provides comparative analysis of two or more models based on reactions, compounds, biomass, and proteins families. The app provides overall statistics of conserved reactions, conserved compounds, and conserved biomass precursors across metabolic models.

The seventh and final step of the pipeline is the *Compare FBA Solutions* app. KBase permits the use of flux balance analysis to predict how an organism will behave metabolically in a wide range of growth conditions. With this capability, it quickly becomes important to be able to compare the flux profiles predicted by FBA side-by-side in order to understand how an organism's behavior changes from one condition to the next, or how the behavior of two different organisms differs within a single condition. The *Compare FBA Solutions* app enables this comparison. FBA solutions are compared on three levels: (1) the objective value for each FBA solution; (2) the flux through each reaction in each FBA solution; and (3) the uptake and excretion of metabolites in each FBA solution. For the flux comparison, reaction fluxes are categorized into four possible states: not in model; no flux; forward flux; and reverse flux. Metabolite fluxes are categorized into similar states: not in model; no flux; uptake; and excretion. FBA solutions are compared based on these states, and solutions with similar states are compared based on magnitude of flux.

2.3 Metabolic Pathways in the Core Model Template

All metabolic model reconstruction in KBase is built upon a set of model templates, each of which integrates the three types of data needed to build a model from an annotated genome: (1) the full set of reactions that comprise the metabolic pathways across a wide range of organisms; (2) the SEED functional roles associated with the enzymes that perform all metabolic reactions; and (3) the default objective functions to be used in the reconstructed models (*see* Subheading 2.5). Different model templates are used to construct different types of models (e.g., plants, gram negative genomes, gram positive genomes), and for core models, a specific core model template (CMT) is applied.

The CMT integrates 200 highly curated reactions (https://github.com/ModelSEED/ModelSEEDDatabase/blob/master/Templates/Core/Reactions.tsv) encompassing 12 key energy biosynthesis pathways (Fig. 2) linked to central metabolism including: sugar degradation pathways (Glycolysis, Entner-Doudoroff, Citric acid cycle and Pentose phosphate), fermentation pathways (producing end products: lactate, acetate, formate, ethanol, 2,3-butanediol, butyrate, butanol, and acetone) that are derived

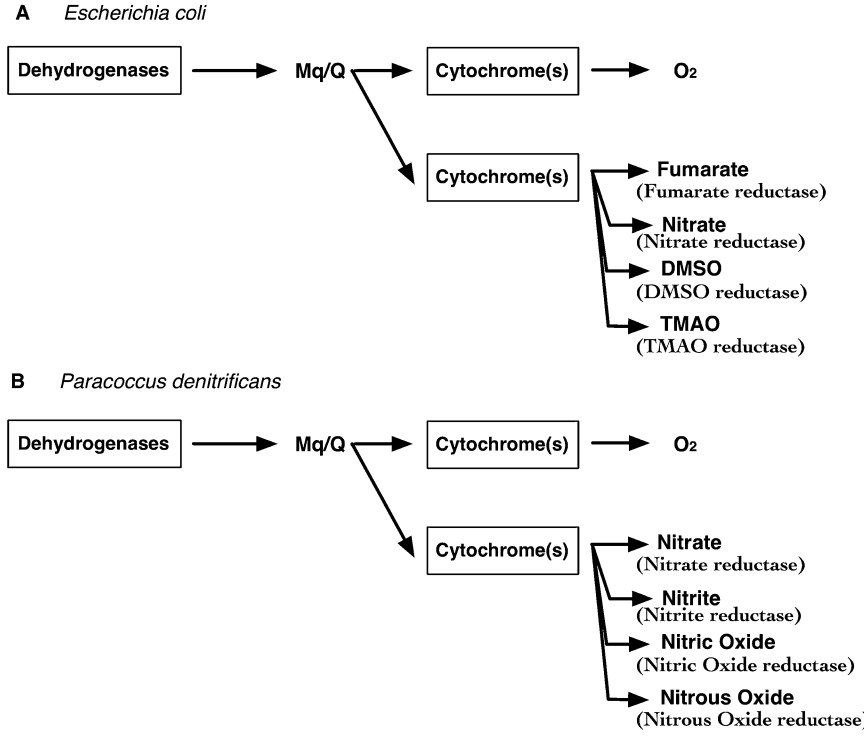

Fig. 3 Diverse electron transport chains in bacteria. Escherichia coli (**a**) and *Paracoccus denitrificans* (**b**) are able to respire aerobically and anaerobically by reducing nitrate. *E. coli* (**a**) is able to reduce organic electron acceptors fumarate, DMSO, and TMAO. *P. denitrificans* (**b**) is able to reduce more inorganic electron acceptors including nitrite, nitric oxide, and nitrous oxide

from central metabolism as well as various aerobic and anaerobic Electron Transport Chains (*see* Subheading 2.4). These pathways are considered the building blocks of the central metabolism that is represented by our CMT, and they were derived from an analysis of a phylogenetically diverse set of well-studied model organisms, including *Escherichia coli*, *Bacillus subtilis*, *Pseudomonas aeroginosa*, *Clostridium acetobutylicum*, and *Paracococcus denitrificans* [9]. We also added a number of manually curated ETC reactions to the CMT. These reactions reflect the diverse ETC variations in aerobic respiration as well as facilitating the reduction of number of anaerobic electron acceptors (*see* Subheading 3.4). The CMT maps its 200 reactions to over 400 SEED functional roles through complexes (*see* Subheading 3.3, Fig. 3). These mappings are used to associated genes to the CMT reactions when building a CMM from an annotated genome.

2.4 Encoding of ETC Diversity in the Core Model Template

Unlike the electron transport chains of higher eukaryotes, bacterial ETCs are highly diversified. As a result, they are able to grow in a variety of aerobic and anaerobic environments reducing anaerobic electron acceptors such as nitrate, nitrite, fumarate, dimethyl

sulfoxide (DMSO), and trimethylamine N-oxide (TMAO). Given the importance of the ETCs in governing cell behavior and growth conditions, significant curation was invested to encode various ETCs as a part of the CMT. For instance, *Escherichia coli* (Fig. 3a) is known to respire aerobically and anaerobically reducing nitrate, fumarate, TMAO, and DMSO. *Paracoccus denitrificans* (Fig. 3b) is also able grow aerobically and anaerobically reducing a variety of nitrogen-based compounds including nitrate, nitrite, nitrous oxide, and nitric oxide. Better annotation of ETCs aids identification of complex respiration types and makes energy yield predictions more accurate.

2.5 Default Objective Functions in the Core Model Template

The CMT integrates two default objective functions for the CMMs: (1) a biomass production objective function, modeled by maximizing the simultaneous production of 16 central carbon precursors needed to produce 1 g of biomass (green circles in Fig. 2); and (2) an energy production objective function modeled by maximizing flux through an ATP hydrolysis reaction. The biomass biosynthesis objective function in our CMT was constructed based on the biomass precursor stoichiometry that was derived by Varma and Palsson [24] and used in one of the earliest models of *E. coli* (Table 1) [25]. In our analysis of the biomass objective function, we found that gapfilling was occasionally required to enable synthesis of all essential biomass precursors in our biomass object function [9]. For this reason, we also include the energy object function in the CMT, which permits a focused study of energy biosynthesis in our core models without any gapfilling. Using this objective function, we computed ATP production yields in all models without any gapfilling; hence, these computations were based solely on reactions derived from existing RAST annotations.

3 Methods

3.1 Construction of a Draft Core Metabolic Model from an Annotated Genome

Here, we apply our core model reconstruction pipeline (*see* Subheading 2.2 and Fig. 1) in KBase to build and analyze a core model for the genome *Escherichia coli* K12 (see http://kbase.us/core-models/). Because we are starting with an imported genome, we skip the genome assembly step of our pipeline and apply the *Annotate Microbial Genome* app to re-annotate our genome with functions from the SEED subsystems ontology [15]. In this re-annotation step, RAST assigns 3889 genes with 3797 distinct functions. 1804 of these functions appear in the SEED subsystem ontology.

Now that the genome is annotated with SEED functions, the *Build Metabolic Model* app can be used with the CMT (*see* Subheading 2.3) selected to build a draft CMM. In addition to constructing

Table 1
Central carbon precursors of small-molecule building blocks of biomass

Biomass compound	Coefficient
NADPH	−1.8225
D-Erythrose4-phosphate	−0.8977
NADH	3.547
Phosphoenolpyruvate	−0.5191
NADP	1.8225
NAD	−3.547
H_2O	−41.257
Acetyl-CoA	−3.7478
ADP	41.257
CoA	3.7478
ATP	−41.257
Pyruvate	−2.8328
3-Phosphoglycerate	−1.496
Oxaloacetate	−1.7867
Phosphate	41.257
D-fructose-6-phosphate	−0.0709
ribose-5-phosphate	−0.8977
H+	46.6265
Glyceraldehyde3-phosphate	−0.129
2-Oxoglutarate	−1.0789
D-glucose-6-phosphate	−0.205

Compound names and associated coefficients of the biomass biosynthesis objective function used in CMMs. This biomass stoichiometry originally derived by Varma and Palsson [8] and used in CMMs with modifications [9]

a core model, this app has an optional "Gapfill metabolic model" checkbox. If this option is selected, then when the *Build Metabolic Model* step finishes, the *Gapfill Metabolic Model* step will start automatically. For the sake of our example, this checkbox is left "unchecked" so these steps can be run separately. When the *Build Metabolic Model* step completes, a draft model is created, which is comprised of 158 reactions, 168 compounds, and 478 genes. Because gapfilling was not run automatically, this draft model only includes reactions that are associated with genes.

The gene associations were generated based on a two-step process. In the *Annotate Microbial Genome* app, the genes in our genome were assigned biological functions (e.g., Pyruvate kinase

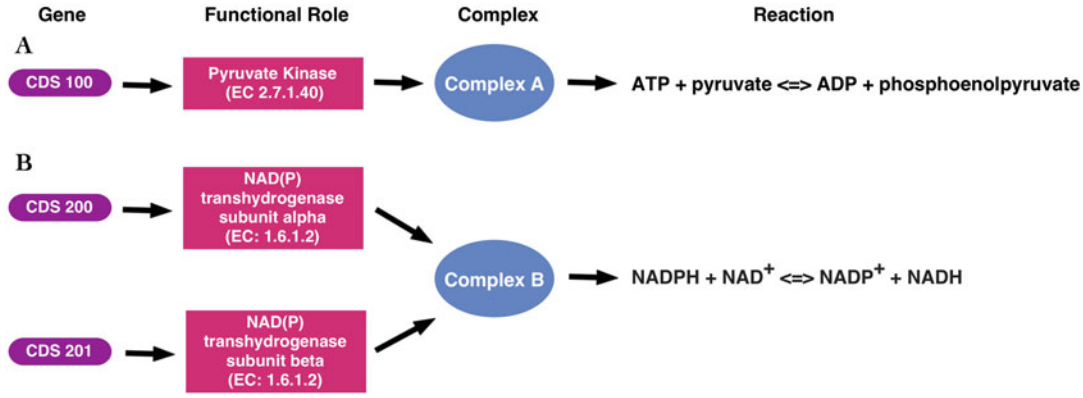

Fig. 4 Organization of genes, gene annotations, complexes, and the biochemical reactions in gene protein reaction (GPR) mappings. Panel (**a**) shows a gene assigned Pyruvate Kinase (EC 2.7.1.40) as a function. This gene is mapped first to a complex (Complex A), then to a biochemical reaction. Panel (**b**) shows two genes that were assigned the NAD(P) transhydrogenase (EC 1.6.1.2) alpha and beta subunits as functions. These genes are mapped first to a single complex (Complex B), then to the appropriate reaction

(EC 2.7.1.40)), and in our CMT, these functions are mapped to the appropriate biochemical reactions. Thus, the *Build Metabolic Model* app maps the reactions associated with function *A* in the CMT to the gene(s) associated with function *A* in the genome (Fig. 4a). As some metabolic enzymes have multiple functional subunits encoded by separate genes, this mapping process also integrates information about such complexes, so the genes encoding separate subunits are mapped to the appropriate reaction as a group (Fig. 4b). If only one subunit is annotated in the genome, the reaction is still added to the model, although a note is made that the other subunits appear to be missing.

The draft model also has two different objective functions, as defined by our CMT (*see* Subheading 2.5): an energy production function (called bio2) and a biomass production function (called bio1). These objective functions play a role in the gapfilling of the model performed in the next step of our model reconstruction pipeline, as well as in how the model is analyzed during flux balance analysis.

3.2 Gapfilling Core Metabolic Model for Energy and Biomass Production

The next step of our pipeline is to gapfill the CMM to enable the production of energy and biomass in a specified growth condition. We must specify a growth condition when gapfilling because the nutrients present in the growth condition have a major impact on the reactions required to permit growth and energy production. In KBase, growth conditions are specified as media formulations, which specify the concentration and uptake ranges of all metabolites known to be available in the growth condition. By default, gapfilling will use a special growth condition called *Complete* media, which includes all metabolites for which there is a transport

reaction in KBase (*see* Subheading 2.2). In the case of our *E. coli* K12 CMM, we will gapfill in glucose minimal media as *E. coli* K12 is known to grow in this condition.

Gapfilling also requires that a specific objective function be specified for the gapfilling operation. The output of this specified objective function (e.g., biomass reaction or a transporter reaction) is constrained to a nonzero value, while linear programming algorithms are applied to identify a minimal set of additional reactions that must be added to the model to permit the function to achieve a nonzero value. In the case of our *E. coli* K12 CMM, we have specified the biomass reaction, which produces all the compounds involved in the biomass biosynthesis (Table 1), as the objective function resulting in an objective value greater than zero.

Because we have two separate objective functions in our CMM, we run the gapfilling with each function. As it turns out, our *E. coli* K12 CMM requires no additional reactions to reach a nonzero value with either of our objective functions. This result was expected for our energy production objective function, which was specifically designed to require minimal or no gapfilling. However, some genomes do require at least some gapfilling to permit standard biomass production. This lack of significant required gapfilling highlights one of the major strengths of using CMMs: CMMs generally require far less (or no) gapfilling compared to genome-scale models, meaning the predictions they make will be based primarily if not entirely on the genome annotations. Now that our *E. coli* K12 CMM has been demonstrated to be capable of producing both energy and biomass, we can use flux balance analysis to predict the flux profile in *E. coli* that optimizes each of these objective functions.

3.3 Analysis of Core Metabolic Model with Flux Balance Analysis

In the fifth step of our pipeline, we use the *Run Flux Balance Analysis* app in KBase to optimize the production of biomass and energy (*see* Subheading 2.5) in our CMM, while also predicting metabolite uptake, intracellular flux profile, and growth/ATP production yields. As with gapfilling, FBA requires that a growth condition (media) be specified for the analysis, and as before, we select glucose minimal media under aerobic conditions as our desired growth condition for analysis. Our FBA reveals a biomass yield of 0.12 g biomass/mmol glucose uptake and an energy yield of 26.5 mmol ATP per mmol glucose uptake. In KBase, the *Run Flux Balance Analysis* app automatically also runs flux variability analysis (FVA), which enables the classification of model reactions during predicted growth or energy production. FVA reveals that 37 (27%) of reactions in *E. coli* are essential for biomass production when growing in glucose minimal media, while 31 (19%) are essential for energy production. As expected, simple energy production requires fewer pathways than biomass production.

3.4 Comparative Analysis of CMMs and Flux Distributions

Now that our *E. coli* CMM has been built and analyzed, it is interesting to apply this same pipeline to examine other genomes and/or growth conditions. KBase includes tools that support the comparison of models and/or FBA solutions when studying multiple genomes in multiple growth conditions. Comparative analysis of the models helps to reveal metabolic pathways that are common to all models compared and also to identify unique parts of metabolism for each individual model. Comparing flux distributions allows the identification of high flux pathways and pathways that are not being utilized under certain environmental conditions.

We demonstrate this capability by applying our CMM reconstruction pipeline to build a model of *Paracoccous denitrificans* PD1222. We then compare the models (*E. coli* and *P. denitrificans*) and their predicted flux profiles using the *Compare Models* and *Compare FBA Solutions* apps respectively. This analysis reveals that *P. denitrificans* and *E. coli* have 50 reactions in common, with only 4 reactions unique to *P. denitrificans* and 112 reactions unique to *E. coli*. Comparing the FBA predictions, we find that the reactions common to both models were largely essential for biomass production, while the reactions unique to each model were active but not essential. We also compared the flux profiles of *E. coli* that optimize the energy yield in glucose minimal media under aerobic and anaerobic conditions. This comparison reveals that ATP yield is much higher (26.5 ATPmmol/mmol of glucose) in the aerobic condition where 31 reactions have nonzero fluxes compared to in the anaerobic condition (2.75 ATPmmol/mmol of glucose) where only 21 reactions have nonzero fluxes. The difference in energy yield is due to the fact that under aerobic conditions, *E. coli* is able to fully oxidize glucose-utilizing aerobic ETCs, yielding more energy, whereas in anaerobic conditions with no anaerobic electron acceptors present, energy is produced solely through fermentation by substrate level phosphorylation.

3.5 Determining Metabolic Pathways in CMMs and Phylogenetic Distribution

To study and evaluate the metabolic potential of an organism, it is useful to ascertain the existence of classical metabolic pathways. We have used CMMs to determine the presence or absence of key energy biosynthesis-related pathways (Fig. 2). We have developed a set of Boolean rules to determine the presence and absence of each pathway based on reactions present in each of the CMMs [9]. This methodology allows for alternative reactions within an individual step of each pathway, but every step of each defined pathway must be annotated in order for the pathway to be classified as present. Boolean rules that were used to determine the existence for Glycolysis and Gluconeogenesis are listed in Table 2.

Once the presence and absence of pathways have been determined, there are multiple ways to analyze the pathway data. In Fig. 5, we have painted pathway presence and absence data for Glycolysis, Gluconeogenesis, and Entner-Doudoroff on a

Table 2
Booleans rules used to govern pathway presence/absence in CMMs (rules that are displayed in this table were used to determine presence/absence of glycolysis and gluconeogenesis)

Section 1	
Enzyme names	*Reaction ID*
phospho_glucose_isomerase	rxn00558
ATP-dependent-pfk	rxn00545
ADP-dependent-pfk	rxn04043
ppi-dependent-pfk	rxn00551
NAD-dependent_phosphoglycerate_dehydrogenase	rxn00781
NADP-dependent_phosphoglycerate_dehydrogenase	rxn00782
ATP-dependent_phosphoglycerate_kinase	rxn01100
GTP-dependent_phosphoglycerate_kinase	rxn01105
phosphoglycerate_mutase	rxn01106
Enolase	rxn00459
pyruvate_kinase	rxn00148
fructose_bis_phosphotase	rxn00549
f1,6_bisphosphate_aldolase	rxn00786
ATP_pyruvate_water_phosphotransferase	rxn00147
Section 2	
Enzyme names/pathway segments	*Rule*
gdh	means 1 of {NAD-dependent_phosphoglycerate_dehydrogenase,NAD-dependent_phosphoglycerate_dehydrogenase}
pgk	means 1 of {ATP-dependent_phosphoglycerate_kinase, ADP-dependent_phosphoglycerate_kinase}
pgm	means phosphoglycerate_mutase
pyrk	means pyruvate_kinase
pfk	means 1 of {ATP-dependent-pfk,ADP-dependent-pfk,ppi-dependent-pfk}
G3P-PYR	means gdh and pgk and pgm and enolase and pyrk
G3P-PEP	means gdh and pgk and pgm and enolase
F6P-PYR	means pfk and f1,6_bisphosphate_aldolase and G3P-PYR
G6P-PYR	means phospho_glucose_isomerase and F6P-PYR

(continued)

Table 2
(continued)

glycolysis_t1	means G6P-PYR
glycolysis_t2	means F6P-PYR and not G6P-PYR
glycolysis	means glycolysis_t1 or glycolysis_t2
gluconeogeneis	means fructose_bis_phosphotase and G3P-PEP or (fructose_bis_phosphotase and G3P-PEP and ATP_pyruvate_water_phosphotransferase)
glycolysis_is_supported	means glycolysis
glycolysis_is_not_supported	means not glycolysis
glycolysis_is_ADP-dependent	means glycolysis and (ADP-dependent-pfk or ADP-dependent_phosphoglycerate_kinase)
glycolysis_is_ppi-dependent	means glycolysis and ppi-dependent-pfk

Section 1 of the table displays the reaction names and the corresponding reaction ids for the pathways that are considered for establishing Boolean rules. Section 2 displays assigned rules for reactions that are mentioned in Section 1. Rules able to facilitate: (1) pathway steps that may have more than one enzymatic reaction (e.g., *gdh* means 1 of {NAD dependent_phosphoglycerate_dehydrogenase,NAD-dependent_phosphoglycerate_dehydrogenase}), (2) partial pathway segments (e.g., G3P-PEP means *gdh* and *pgk* and *pgm* and enolase), and (3) complete pathways (e.g., glycolysis means glycolysis_t1 or glycolysis_t2)
Boolean rules were established for 12 central metabolic pathways including pathways that are mentioned in this table and were originally published in Edirisinghe et al. [9])

phylogenetic tree where it depicts phylogenetic distribution of a given biochemical pathway. In Fig. 6, we have organized all CMMs by their taxonomic groups against pathway presence and absence data. Taxonomic groups that are displayed along the horizontal axis of Fig. 6 were sorted sequentially as they appear in a 16S rRNA-based phylogenetic tree [9].

3.6 Overview and Discussion

In this chapter, we present a detailed protocol for the reconstruction and analysis of core metabolic models (CMMs) in KBase. In comparison to genome-scale models, CMMs are simpler and can accurately determine: (1) ATP yields based on different growth/environmental conditions, (2) ETC variations and respiration types, (3) ability to produce fermentation products, (4) presence and absence of classical biochemical pathways in central metabolism, and (5) ability to produce key metabolic pathway intermediates in central metabolism which are precursors of essential biomass components of the cell.

We have implemented the CMM construction and analysis pipeline using KBase apps (*see* Subheading 2.2) with commentary (*see* Subheading 2.1), where the following major steps are demonstrated: (1) annotation of microbial genomes, (2) reconstruction of CMM, (3) gapfilling of CMM, and (4) perform flux balance analysis (Fig. 1). Comparative analysis of CMMs and flux distributions based on different media conditions is also demonstrated.

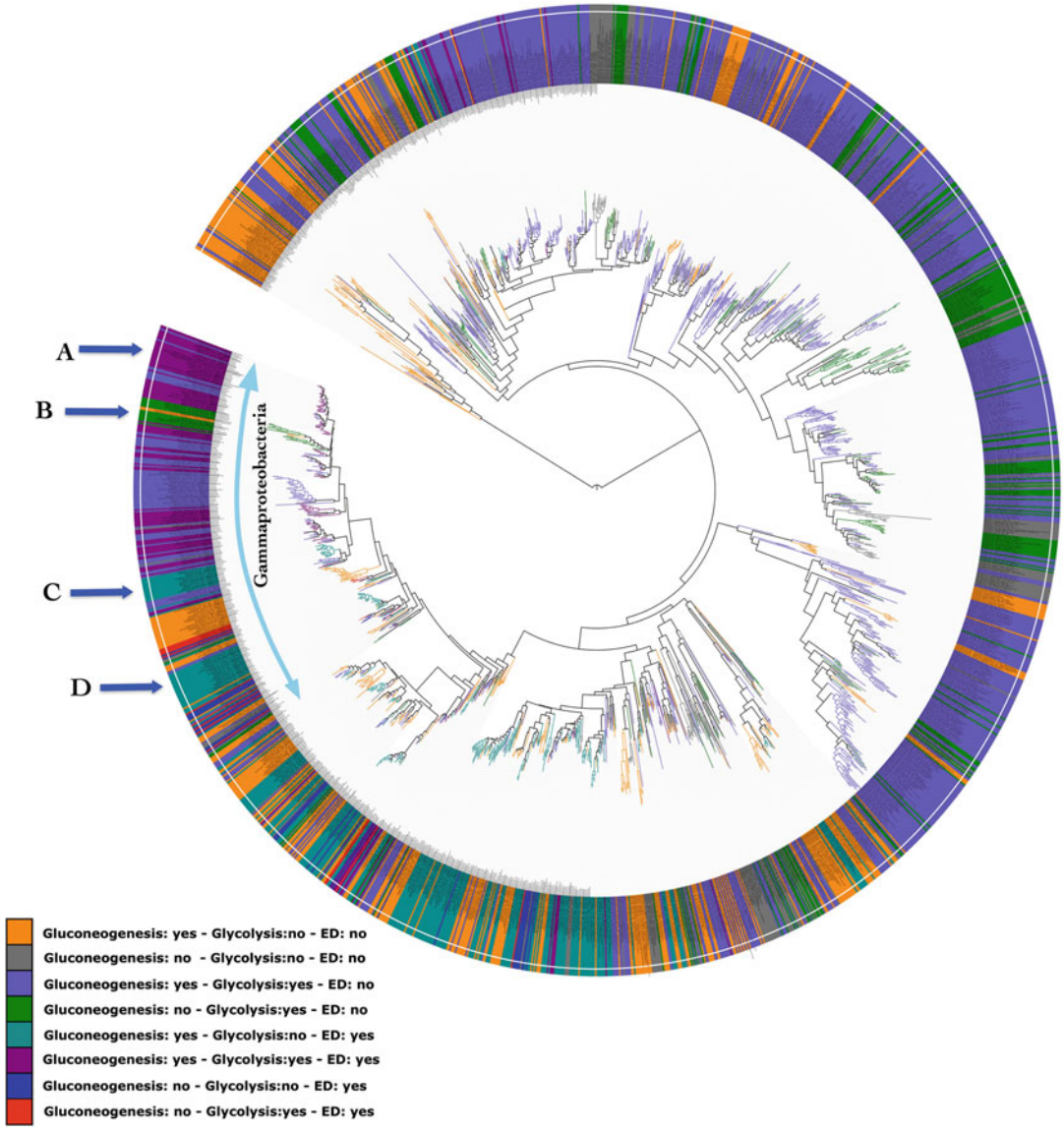

Fig. 5 Phylogenetic distribution of central metabolic pathways (originally published in Edirisinghe et al. [9]). Microbial life tree (16S $OTU_{98.5}$) depicting the presence and absence of sugar degradation pathways glycolysis, gluconeogenesis, and Entner-Doudoroff. The name of the organism and the phylum can be found at the leaf of the tree in the high-resolution image. The colored branches depict which clades gained or lost certain metabolic pathways. The curved arrow shows the range of the group Gammaproteobacteria, and the straight arrows indicate the regions where species belongs to several different genera have different phenotypes with the same taxonomic group: Escherichia and Salmonella (purple) (A), (B) *Buchnera* (green), (C) *Shewanella* (light blue) and (D) *Pseudomonas* (light blue). A high-resolution image of this figure can be accessed at http://bioseed.mcs.anl.gov/~janakae/coremodel/springer/fig5.pdf

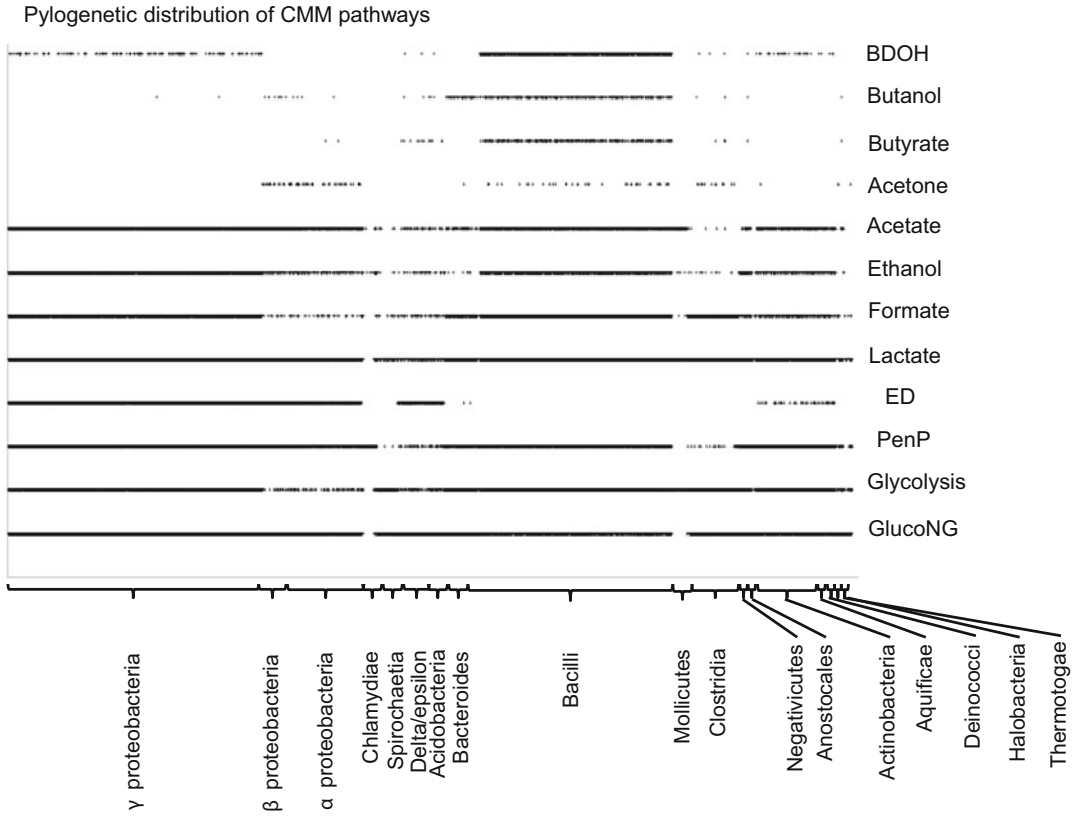

Fig. 6 Presence and absence of key central metabolic pathways of about 8100 organisms sorted by major phylogenetic groups originally published in Edirisinghe et al. [9]). Taxonomic groups that are displayed in the horizontal axis of the graph were sorted sequentially as they appear in a 16SrRNA-based phylogenetic tree (*GlucoNG* gluconeogenesis, *ED* Entner-Doudoroff, *PenP* Pentose Phosphate)

Along with the described reconstruction and analysis tools, KBase also offers a large amount of public data including microbial genomes and media formulations that aid in the CMM reconstruction process across the microbial tree of life (*see* Subheadings 2.2 and 3.5).

In our specific example where we used the *Escherichia coli* K12 annotated genome as the starting point, we demonstrated the CMM's ability to predict energy yields and biomass without requiring any gapfilling reactions, thus the CMM predictions are solely based on genome annotations. We performed flux balance analysis (FBA) coupled with flux variability analysis (FVA) (*see* Subheading 2.2) using the *E. coli* K12 CMM in glucose minimal media to predict metabolite uptake and excretion, intracellular flux profiles, and growth/ATP production yields. These analyses reveal the essential reactions required for *E. coli* K12 to predict energy yields or biomass/growth under a specific media/environmental condition. A comparative analysis of the CMM of *E. coli* and

P. denitrificans showed conservation of essential metabolic reactions across both organisms, while reactions reflecting each organism's unique biology were deemed mainly nonessential. Comparative analysis of *E. coli* flux profiles under aerobic and anaerobic conditions has revealed the differences in energy yield predictions due to the presence of ETCs. We conclude our analysis by showing the presence and absence of key energy biosynthesis pathways in CMMs, and we present the pathway conservation data in phylogenetic context. In addition to the CMM reconstruction and analysis tools that are discussed in this chapter, KBase offers an extensive catalog of apps (*see* **Note 2**) that provide analysis and comparison capabilities that allow researchers to investigate important biological questions related to microbial metabolism and other topics in systems biology.

4 Notes

1. Genomes, CMMs, Flux distributions, comparative analysis of the models and flux distributions data can be downloaded from the KBase Narrative interface (see the instructions at http://kbase.us/data-upload-download-guide/downloading-data/).
2. KBase offers an extensive catalog of apps for metabolic model construction and for comparative analysis genomes. The list of apps can be found at https://narrative.kbase.us/#catalog/apps/

References

1. Gottschalk G (1988) Bacterial metabolism. Springer, New York, NY
2. Gottschalk G (1989) How *Escherichia coli* synthesizes ATP during aerobic growth of glucose. In: Bacterial metabolism. Springer, New York, NY, pp 13–35
3. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol 28(9):977–982. https://doi.org/10.1038/nbt.1672
4. Karp PD, Paley S, Romero P (2002) The Pathway Tools software. Bioinformatics 18(Suppl 1):S225–S232
5. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nat Protoc 2(3):727–738
6. Monk J, Palsson BO (2014) Genetics. Predicting microbial growth. Science (New York, NY) 344(6191):1448–1449. https://doi.org/10.1126/science.1253388
7. Kumar VS, Dasika MS, Maranas CD (2007) Optimization based automated curation of metabolic reconstructions. BMC Bioinformatics 8:212
8. Varma A, Palsson BO (1994) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia-coli W3110. Appl Environ Microbiol 60(10):3724–3731
9. Edirisinghe JN, Weisenhorn P, Conrad N, Xia F, Overbeek R, Stevens RL, Henry CS (2016) Modeling central metabolism and energy biosynthesis across microbial life. BMC Genomics 17:568. https://doi.org/10.1186/s12864-016-2887-8
10. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch

GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9:75. https://doi.org/10.1186/1471-2164-9-75

11. Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? Nat Biotechnol 28 (3):245–248. https://doi.org/10.1038/nbt.1614. nbt.1614 [pii]
12. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O (2008) The RAST server: rapid annotations using subsystems technology. BMC Genomics 9:15. https://doi.org/10.1186/1471-2164-9-75. 75 [pii]
13. Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23(6):673–679. https://doi.org/10.1093/bioinformatics/btm009, btm009 [pii]
14. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119
15. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res 42(Database issue):D206–D214. https://doi.org/10.1093/nar/gkt1226
16. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, Shukla M, Thomason JA III, Stevens R, Vonstein V, Wattam AR, Xia F (2015) RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Sci Rep 5:8365. https://doi.org/10.1038/srep08365
17. Snyder EE, Kampanya N, Lu J, Nordberg EK, Karur HR, Shukla M, Soneja J, Tian Y, Xue T, Yoo H, Zhang F, Dharmanolla C, Dongre NV, Gillespie JJ, Hamelius J, Hance M, Huntington KI, Jukneliene D, Koziski J, Mackasmiel L, Mane SP, Nguyen V, Purkayastha A, Shallom J, Yu G, Guo Y, Gabbard J, Hix D, Azad AF, Baker SC, Boyle SM, Khudyakov Y, Meng XJ, Rupprecht C, Vinje J, Crasta OR, Czar MJ, Dickerman A, Eckart JD, Kenyon R, Will R, Setubal JC, Sobral BW (2007) PATRIC: the VBI PathoSystems Resource Integration Center. Nucleic Acids Res 35(Database issue): D401–D406. https://doi.org/10.1093/nar/gkl858. gkl858 [pii]
18. Tran TT, Dam P, Su Z, Poole FL II, Adams MW, Zhou GT, Xu Y (2007) Operon prediction in Pyrococcus furiosus. Nucleic Acids Res 35(1):11–20. https://doi.org/10.1093/nar/gkl974, gkl974 [pii]
19. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res 36 (Database issue):D480–D484
20. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30
21. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Paley S, Popescu L, Pujar A, Shearer AG, Zhang P, Karp PD The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 38(Database issue):D473–D479. https://doi.org/10.1093/nar/gkp875. gkp875 [pii]
22. Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bonavides C, Gama-Castro S (2002) The EcoCyc database. Nucleic Acids Res 30(1):56–58
23. Ware D, Jaiswal P, Ni J, Pan X, Chang K, Clark K, Teytelman L, Schmidt S, Zhao W, Cartinhour S, McCouch S, Stein L (2002) Gramene: a resource for comparative grass genomics. Nucleic Acids Res 30(1):103–105
24. Varma A, Palsson BO (1993) Metabolic capabilities of Escherichia-coli. 2. Optimal-growth patterns. J Theor Biol 165(4):503–522
25. Varma A, Palsson BO (1993) Metabolic capabilities of Escherichia-coli.1. Synthesis of biosynthetic precursors and cofactors. J Theor Biol 165(4):477–502