Chapter 2

Automated Genome Annotation and Metabolic Model Reconstruction in the SEED and Model SEED

Scott Devoid, Ross Overbeek, Matthew DeJongh, Veronika Vonstein, Aaron A. Best, and Christopher Henry

Abstract

Over the past decade, genome-scale metabolic models have proven to be a crucial resource for predicting organism phenotypes from genotypes. These models provide a means of rapidly translating detailed knowledge of thousands of enzymatic processes into quantitative predictions of whole-cell behavior. Until recently, the pace of new metabolic model development was eclipsed by the pace at which new genomes were being sequenced. To address this problem, the RAST and the Model SEED framework were developed as a means of automatically producing annotations and draft genome-scale metabolic models. In this chapter, we describe the automated model reconstruction process in detail, starting from a new genome sequence and finishing on a functioning genome-scale metabolic model. We break down the model reconstruction process into eight steps: submitting a genome sequence to RAST, annotating the genome, curating the annotation, submitting the annotation to Model SEED, reconstructing the core model, generating the draft biomass reaction, auto-completing the model, and curating the model. Each of these eight steps is documented in detail.

Key words: Model SEED, RAST, Automated metabolic model reconstruction, Flux balance analysis, Gap filling, Microbial metabolism, Systems metabolic engineering

1. Introduction

Over the past decade, genome-scale metabolic models have proven to be a crucial resource for predicting organism phenotypes from genotypes. These models provide a means of rapidly translating detailed knowledge of thousands of enzymatic processes into quantitative predictions of whole-cell behavior. They have been applied extensively to identify essential genes and genes sets, predict organism phenotypes and growth conditions, design metabolic engineering strategies, and simulate the effects of transcriptional regulation on organism behavior (1). Yet until recently, the pace of new metabolic model development was eclipsed by the pace at which new genomes were being sequenced. To address this problem, the Model SEED

Hal S. Alper (ed.), *Systems Metabolic Engineering: Methods and Protocols*, Methods in Molecular Biology, vol. 985,
DOI 10.1007/978-1-62703-299-5_2,

framework (http://www.theseed.org/models/) (2) was developed as a means of automatically producing draft genome-scale metabolic models to increase the pace of new model development and close the gap between the number of metabolic models and the number of sequenced genomes. The Model SEED integrates existing methodologies (3–9) and introduces new techniques to automate nearly every step of the metabolic reconstruction process (10), enabling generation of functioning draft models from assembled genome sequences in approximately 48 h. Today, the Model SEED has been applied to generate over 15,000 draft metabolic models, including a model of the over 3,500 complete prokaryotic genome sequences currently available in GenBank (11).

A genome-scale metabolic model consists of three primary components: (1) a list of reactions that take part in the metabolic pathways of the organism including reaction stoichiometry and reversibility, (2) a set of gene–protein-reaction (GPR) associations that capture how gene activity is related to the activity of metabolic reactions, and (3) a biomass composition reaction that indicates which small molecules must be produced for an organism to grow and divide (12). All of these components are used in a method called flux balance analysis (FBA) to simulate microbial metabolism in a specified environmental condition.

FBA involves the use of linear optimization to define the limits on the metabolic capabilities of a model organism by assuming that the interior of the cell exists in a quasi-steady-state (13–16). This quasi-steady-state assumption is enforced by a set of linear mass balance constraints written for each metabolite included in the model. These mass balance constraints and reaction flux bounds form a set of underdetermined linear equations with many possible solutions. Because these equations are underdetermined, an optimization criterion is used to capture the most physiologically relevant region of the solution space. The optimization criteria vary depending on the application, but the most common criterion is the maximization of growth yield (16, 17). Maximum growth yield is simulated by maximizing the flux through the biomass reaction in the model, while the uptake of nutrients is fixed at a specific ratio. This is a meaningful optimization criterion because organisms have been observed to grow at the maximum predicted yield when nutrients are plentiful (18).

In this chapter, we provide detailed descriptions of how to construct a new draft genome-scale metabolic model starting from a new unannotated genome sequence. We describe in detail how to use the SEED (4) and RAST (3) tools to annotate a genome and review the genome annotations. We then describe how to use the Model SEED and other tools to construct, review, and analyze a metabolic model from the RAST annotation. We also move beyond simple descriptions of how to use these tools to also include details on how these tools work "under the hood." These details are useful

for developing a complete understanding of the data and assumptions that enter into the automated model reconstruction process. Overall, we break down the automated model reconstruction process into eight sequential steps including (1) submitting a genome sequence to RAST, (2) annotation of the genome, (3) review and curation of the annotation, (4) submitting a RAST annotation to Model SEED, (5) reconstruction of a core metabolic model, (6) generation of a draft biomass composition reaction, (7) autocompletion of the metabolic model, and (8) review and curation of the metabolic model.

2. Materials

2.1. Requirements for Submitting a Genome Sequence to Automated Annotation

1. A FASTA file containing the sequence of all chromosomes and plasmids for the microbial genome to be annotated, typically obtained from the submitters' own sequencing project or GenBank (http://www.ncbi.nlm.nih.gov/genbank/).
2. Web access to the RAST home page (http://rast.nmpdr.org/) or an installation of the myRAST desktop application (http://blog.theseed.org/servers/).
3. A user account in the SEED, which can be freely obtained via the SEED registration page (http://rast.nmpdr.org/?page=Register).

2.2. Data Supporting the RAST Approach to Automated Genome Annotation

1. A curated, controlled vocabulary of functional roles that define the specific biological functions that will be mapped onto genes in the annotation process (e.g., pyruvate kinase (EC 2.7.1.40)).
2. An organization of related functional roles into a set of well-curated subsystems (e.g., glycolysis and gluconeogenesis).
3. A large number of diverse microbial genome sequences to serve as the reference genomes to which all other genome sequences will be mapped for initial annotation.
4. A database of protein families, called FIGfams (19) in the SEED, that represent isofunctional homologues.

2.3. Data Supporting the Curation of RAST Genome Annotations

1. Web access to the RAST (http://rast.nmpdr.org/) and PubSEED (http://pubseed.theseed.org) home pages.
2. At least some genome sequences of organisms that are phylogenetically close to your organism to support comparative genomics approaches.

3. Some expertise in the subsystems you are curating, including knowledge of how biological functions in the subsystem interact (e.g., neighboring steps in a metabolic pathway).

2.4. Requirements for Submitting a Genome for Automated Model Reconstruction

1. A genome sequence annotated by RAST (see Subheading 3.1) or a genome currently available in the PubSEED (http://pubseed.theseed.org).
2. Web access to the Model SEED home page (http://www.theseed.org/models/).
3. A user account in the SEED, which can be freely obtained via the SEED registration page (http://rast.nmpdr.org/?page=Register).

2.5. Data Supporting Reconstruction of Metabolic Models in Model SEED

1. A comprehensive database of the biochemical reactions that comprise the known metabolic pathways that will be included in the models.
2. An annotation ontology with a strict controlled vocabulary.
3. A curated mapping from the reactions in the biochemistry database to protein complexes and from protein complexes to functional roles in the annotation ontology.
4. A genome with genes consistently annotated with the annotation ontology.
5. A list of spontaneous and universal reactions that should be added to all models.

2.6. Data Supporting Generation of Biomass Composition Reactions in Model SEED

1. An estimation of the fraction of biomass that consists of DNA, RNA, protein, lipids, cell wall, and cofactors and an estimation of growth-associated ATP consumption.
2. An approximate estimation of the amino acid composition of protein, the nucleotide composition of RNA, and the GC content of the genome (see Notes 3–4).
3. An annotation ontology with a strict controlled vocabulary.
4. A list of potential lipid, cell wall, and cofactor with conditions on what metabolic functions and subsystems are indicative of a dependence on these metabolites (stated in terms of the annotation ontology).
5. A genome with genes consistently annotated with the annotation ontology.

2.7. Data Supporting Model Auto-completion in the Model SEED

1. A biochemistry database for which generic reactions, unbalanced reactions, nonmicrobial reactions, and lumped reactions have been removed.
2. A media condition in which the auto-completion will be performed.

3. A mapping between reactions and compounds in the model and reactions and compound in the biochemistry database.
4. Optimization software capable of solving a large-scale mixed-integer optimization problem.

2.8. Requirements for Reviewing and Curating a Model SEED Model

1. Access to the Model SEED website (http://www.theseed.org/models/).
2. Cytoscape (http://www.cytoscape.org/) and CytoSEED plugin (http://sourceforge.net/projects/cytoseed/) for viewing metabolic models.
3. Software for running flux balance analysis on metabolic models using SBML files. For example, the COBRA Toolbox (opencobra.sourceforge.net/) or OptFlux (www.optflux.org/).

3. Methods

As described above, the automated genome annotation and metabolic model reconstruction process using the SEED and Model SEED approach can be broken down into eight sequential steps, which we describe in detail here: (1) submitting a genome sequence to RAST, (2) annotation of the genome, (3) review and curation of the annotation, (4) submitting a RAST annotation to Model SEED, (5) reconstruction of a core metabolic model, (6) generation of a draft biomass composition reaction, (7) auto-completion of the metabolic model, and (8) review and curation of the metabolic model.

3.1. Submitting a Genome Sequence to RAST for Automated Annotation

One of the simplest methods for obtaining a consistent annotation for a new genome sequence is to submit the sequence to the RAST server for genome annotation. The RAST server has essentially automated the genome annotation process, requiring the user to supply only a genome sequence and a few simple parameters to get the process started. Once started, the entire annotation process is typically complete in less than 48 h. Here, we provide step-by-step instructions on how to submit a genome sequence for annotation in RAST using our genome submission website: http://rast.nmpdr.org/.

1. All users desiring to submit a genome to RAST for annotation must first register for a SEED user account. Registration is completely open and free of charge and simply provides a mechanism by which we can assign job ownership and ensure private access to submitted genomes and subsequent genome annotations. New users can register for accounts using the account registration webpage on RAST: http://rast.nmpdr.org/?page=Register (Fig. 1a).

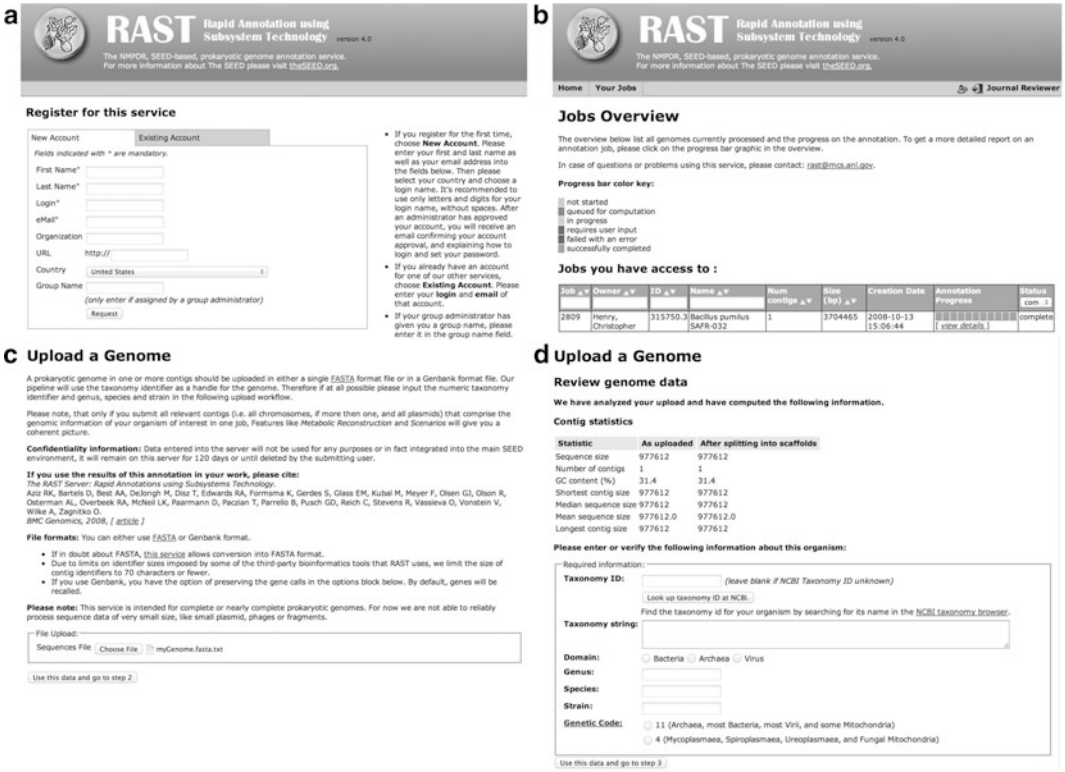

Fig. 1. Web interface for submitting genome sequences to RAST for automated annotation.

2. Once a user has registered for an account, the next step is to log in to RAST using the registered account credentials. Once logged in, the user will be taken to the "Job Overview" page (Fig. 1b), which summarizes all currently running and complete genome annotation jobs submitted by the user to RAST. To upload a new job, simply hover over the "Your Jobs" entry in the menu bar at the top of the page and select the "Upload new job" item.
3. Now you will arrive on the "Upload a genome" page of RAST (Fig. 1c). Simply click on the box labeled "Sequences file," select the FASTA file on your computer, and click the "Use this data and go to step 2" button.
4. If the upload of your FASTA file goes well, you will arrive on the "Review genome data" page (Fig. 1d). This page includes a preliminary analysis of your genome that includes the length of the chromosomes found in your FASTA file. This is useful to ensure the upload and parsing of your file went well. This page also requests an NCBI taxonomy ID (e.g., 83333), an NCBI taxonomy (e.g., Bacteria; . . .*Escherichia coli*), domain (e.g., Bacteria), genus (e.g., *Escherichia*), species (e.g., *coli*), strain (e.g., K-12), and genetic code (e.g., 11) for your genome. Simply fill in the data and click on the "Use this data and go to step 3" button.

5. This will take you to the final input screen, which requests optional information including the sequencing technique (e.g., nearly always pyrosequencing today), coverage (e.g., typically over 10×), the number of contiguous strings of DNA (e.g., the number of chromosomes for completely assembled genomes), and the average read length (e.g., 100–200 on modern pyrosequencing machines). This page also enables the user to select additional optional parameters to be used during the RAST annotation process (3). These include selecting the gene calling methodology (RAST or Glimmer3 (20, 21)), selecting the release of FIGfams (19) to be used (FIGfams are the protein families used as references against which all RAST annotations are made), and binary parameters for fixing errors, fixing frameshifts, building a metabolic model (see Subheading 3.4), backfilling gaps, and disabling replication. We recommend reviewing RAST documentation for details on these parameters (http://www.nmpdr.org/FIG/wiki/view.cgi/Main/RAST); otherwise, the default settings work well. Once the desired settings have been selected, simply click on the "Finish the upload" button to submit the genome for annotation to RAST. The job is scheduled and run on computer servers maintained at Argonne National Laboratory.
6. Users can check on the status of their genome annotation job at any time by returning to the "Job Overview" page on RAST, which will provide regular updates on job progress. RAST jobs are accessible only by the user and RAST administrators, unless a user wishes to make the annotated genome accessible by a wider audience. In that case, the user must explicitly grant access privileges to users as desired.

It is worth noting that RAST itself can run on relatively cheap equipment (under $10,000) and be used to annotate tens of genomes per day. The existing servers support much higher loads (occasionally reaching several hundred genomes per day). RAST scales, so it would be straightforward to support thousands of jobs per day. Over its lifetime thus far, over 50,000 jobs have been submitted to the RAST annotation service.

3.2. The RAST Approach to Automated Genome Annotation

While a user who has submitted a genome sequence to RAST for annotation does not need to know how the RAST annotation process works, it is still useful to know the details of the process to better understand the assumptions and caveats that go into a RAST annotation. When a newly sequenced prokaryotic genome is submitted to RAST for annotation, the job goes through roughly five steps:

1. First, there is a targeted search for a small, well-defined set of elements. Currently, these include rRNAs, tRNAs, genes relating to synthesis and use of selenocysteine, and genes relating to the synthesis and use of pyrrolysine. The set of elements sought will undoubtedly expand as new tools to recognize elements like microRNAs and CRISPRs will be added.
2. Then an iterative step to identify protein-encoding genes is initiated. The bulk of the effort normally is based on use of Glimmer (20, 21). The search usually involves an attempt to recognize common genes, use these as a training set, recall using the training set, and then attempt to remove lengthy overlaps and to fill unusually large gaps.
3. Once estimates of protein-encoding genes (PEGs) have been derived, an initial pass is made to assign functions to the subset of PEGs that can be reliably assigned functions based on the FIGfams/kmers. In genera with numerous existing well-annotated reference genomes, this step often assigns functions to over 90% of the PEGs. In diverse genomes, the percentage can be far less. A second pass is made using BLAST (22) to estimate similarities, and then these are used to assign functions. It is important to note that the reliable assignments in a *controlled vocabulary* are largely based on the first pass and that this second pass is thought of as assigning functions that are clues in an *uncontrolled vocabulary*.
4. PEGs with assigned functions in the *controlled vocabulary* can be gathered into subsystems, when the annotation tools identify *all* of the roles needed to form an active variant of this subsystem. It is precisely the cases in which most, but not all, of the needed roles are identified that require manual curation to clarify what needs to be done to achieve more accuracy and consistency. PEGs that were annotated with *uncontrolled vocabulary* or uncalled ORFs resulting from low-quality DNA sequence are usually responsible for that. For details on this process, see step 1 in Subheading 3.3.

This is an abbreviated description of the process RAST uses to annotate microbial genome sequences. It is important to note that the reference set of genomes that serve as the basis for most of the RAST annotations undergo continuous manual curation, ensuring that annotations remain up-to-date with the latest biological data, ensuring that annotations of genes and gene clusters are consistently propagated to new genomes, and supporting the construction and maintenance of the SEED subsystems and functional roles that form the foundation of the RAST annotation ontology. This manual curation is an essential ingredient contributing to the accuracy and consistency of RAST annotations. Although there is manual annotation of reference genome annotations continuously occurring in

RAST, it is still important for users to review and curate the annotations of their own genomes once the RAST automated annotation process is complete.

3.3. Reviewing and Curating a RAST Annotation

The majority of genomes submitted to RAST are "phylogenetically close" to already annotated genomes. Indeed, it is becoming common for users to submit hundreds of genomes from a single genus or even species. While close genomes do often contain quite different collections of genes, the subsystems identified in the reference genomes become good candidates for the new genomes. Thus, if we computed the "closest 30 genomes" based on rRNA comparisons or upon analysis of a collection of universal (or near universal) genes, then we could reasonably use the subsystems in these close genomes as collections of genes to be sought for in the new genome. A common error in the automated RAST annotation of a genome is the case in which all roles but one within a subsystem were recognized, but the missing role was not detected due to frameshift errors or truncations in the coding gene (i.e., to poor quality sequence data). Implementing a simple, effective way to spot these cases allows us to take advantage of the existing well-annotated close reference genomes. This is the reason why we suggest choosing the "Fix frameshifts" option when submitting the genome to RAST. Quickly identifying the subsystems present in a new genome will allow an understanding of what roles can be identified, and using these roles to characterize which complexes are present creates the needed bridge to the enzymatic ontology.

1. To compare the annotations in a new genome against those in a reference genome, RAST offers the ability to ask for a list of all subsystems present in both genomes or present in one but not the other. From the Job Overview page choose "view details" for the genome of interest. This leads to the Job Details page from which you can choose to "Browse annotated genome in SEED Viewer." The SEED Viewer is the browsing and annotation environment provided by RAST. Under "Comparative Tools," select to run the "Function based Comparison." You will be prompted to select a reference genome and run the comparison. The output is a table summarizing the similarities and differences for all genes, which were associated with subsystems. From this table, one can invoke search tools ("find" button) that will attempt to find functional role assignments that are present in the reference genome, but not in the newly annotated one. Learning how to gather and use this data is the first step towards correcting the initial RAST annotations.
2. Most manual refinements will come from comparison with reference genomes, detection of frameshifts and truncations

by walking the genome in the SEED Viewer environment, and searches for conjectured functions suggested by gap-filling algorithms. The RAST environment offers the registered user the ability to browse the genes, annotations, roles, and subsystems identified for the new genome. The capability exists to add new features, delete features, or to change the annotations associated with features. Changes made by the user can be used to recompute the subsystems ("recompute subsystems" button on the "Organism Overview" page). With experience, a user can learn to compare annotations between a set of genomes to locate potentially unidentified genes, to spot inconsistent annotations, and to locate genes unique to specific genomes (please refer to the "SEED Viewer Tutorial" under the Help menu).

3. Inevitably, there will be an ongoing need to extend the controlled vocabulary. This is achieved by adding subsystems containing the needed roles. Within the SEED Project, there is active encoding of new subsystems by experienced annotators. However, there is a growing need to make it possible for expert users to define and curate their own subsystems. To support this, we have made available within the PubSEED the possibility for users to create their own subsystems. These subsystems will get integrated into the annotation cycle. That is, they will lead to the creation of new FIGfams and then will directly impact future annotations produced by RAST.
4. Users may well have exceptional expertise but no desire to spend the effort needed to construct and maintain a subsystem. Such users can request a new subsystem be built by supplying a definition of the roles and one or more carefully annotated rows in the subsystem. To do so, choose "Request a Subsystem" under the "Navigate" menu. The request will be considered by the SEED annotation team, and if it does represent new functionality with experimental characterization, the new subsystem will get constructed and added to the collection.

Once the RAST genome annotation has been reviewed and curated, the annotated genome is ready to be submitted to the Model SEED for reconstruction of a draft genome-scale metabolic model.

3.4. Submitting a RAST Annotation for Model Construction in the Model SEED

There are two places where a RAST-annotated genome can be submitted to the Model SEED for automated reconstruction of a draft genome-scale metabolic model. First, the web interface for submitting a genome sequence for annotation in RAST includes an option to automatically submit the annotated genome for model reconstruction (see Fig. 2d and step 5 of Subheading 3.1). Here we

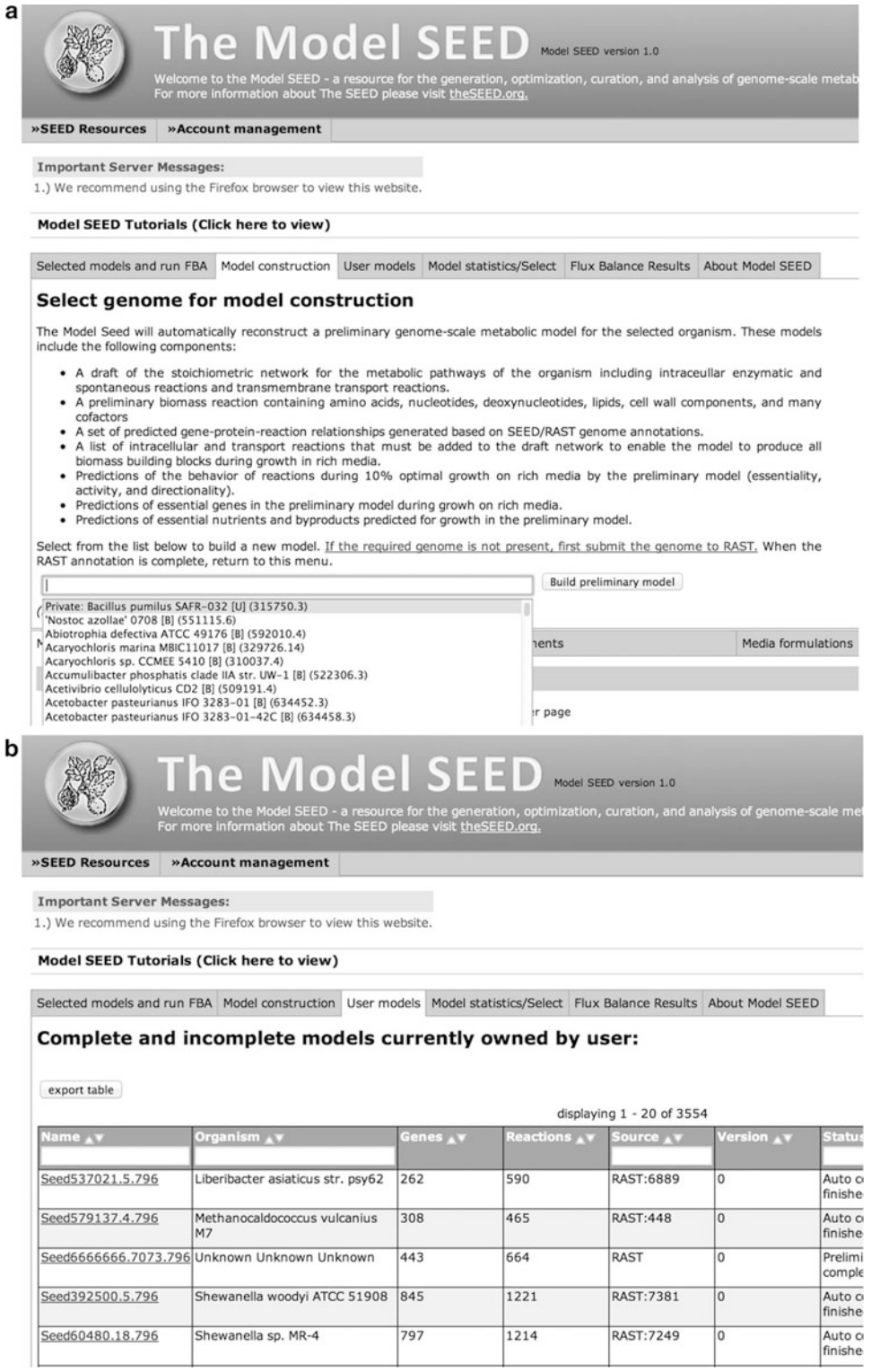

Fig. 2. Model reconstruction and user models.

describe the alternative approach of using menus available on the Model SEED website itself.

1. As with RAST, all users must first register a SEED user account before that can submit genomes for model reconstruction in Model SEED. Note that RAST, SEED, and Model SEED all share the same user registration system, meaning the username and password used to log in to RAST can also be used to log in to the Model SEED and the PubSEED. Registration is completely open and free of charge and simply provides a mechanism by which we can assign model ownership and ensure private access to private models. New users can register for accounts using the account registration webpage on Model SEED: http://www.theseed.org/seedviewer.cgi?page=Register (Fig. 2a).
2. Once an account has been registered, simply visit the Model SEED home page (http://www.theseed.org/models) and log in. Once logged in, click on the "Model Construction" tab in the upper frame of the Model SEED home page.
3. Once the "Model Construction" tab loads, a description of the model construction process will be displayed with a filter select for genomes beneath the description (Fig. 2a). Included in this filter select are all SEED genomes that are publically available for all SEED users as well as any private genome that the logged user has submitted to RAST for annotation. Note that all private genome entries in this filter select begin with the word "PRIVATE." Simply select the genome for which you want to build a model and click on the "Build preliminary model" button. Clicking this button will immediately take you to the "User models" tab of the Model SEED website, which after a moment will now display a table that includes your newly submitted model (Fig. 2b). Note that while you can select the model for viewing immediately, no data will be available until the model reconstruction process is complete.

Once a genome has been submitted for reconstruction in the Model SEED, it enters the automated model reconstruction pipeline of the Model SEED, which we describe in Subheadings 3.5–3.7. This process is entirely automated, so it is not necessary for the user to intervene at any point. This description exists to inform about what goes on under the hood of the Model SEED pipeline. The initial reconstruction of a core model (see Subheadings 3.5–3.6) requires approximately 5–10 min to complete, at which point the core model may be viewed in the Model SEED site (see Subheading 3.8). Initial reconstructions are automatically submitted for auto-completion (see Subheading 3.7), which can require up to an additional 24 h to complete.

3.5. Automated Metabolic Model Reconstruction in the Model SEED

While a user that has submitted an annotated genome for model reconstruction does not need to know how the Model SEED reconstruction works, it is still useful to know the details of the process to better understand the assumptions and caveats that go into the resulting model. To construct a preliminary model, four steps are taken:

1. A comprehensive database of biochemistry is constructed, representing all possible reactions and compounds that a model could use (see Note 1). This database consists of all of the reactions and compounds from the KEGG (23) and many published genome-scale metabolic models (24), combined into a nonredundant set. In total, this database contains over 13,000 reactions and over 16,000 reactants. All reactions that include generic reactants (e.g., alcohol) and all mass and charge imbalanced reactions are disallowed from inclusion in metabolic models (Fig. 3a).
2. All compounds in the database are adjusted to their predominant charged for at pH 7 (see Note 2), and all reactions are proton balanced using the charged forms of the reactants. Gibbs free energy change is then calculated at pH 7 for all reactions in the database using the group contribution method (6), and this data is used to predict reaction reversibility and directionality (7, 25).
3. A mapping is prepared between the SEED annotation ontology and the reactions in the biochemistry database through protein complexes as an intermediate (Fig. 3b). This mapping

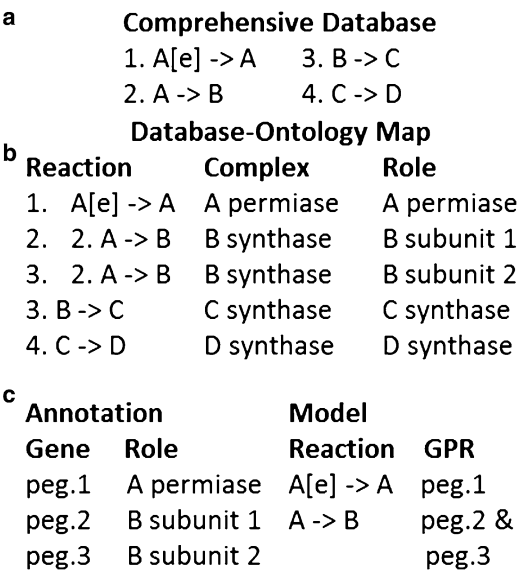

Fig. 3. Model reconstruction process.

is continuously maintained as the SEED annotation ontology and reaction database evolve over time.

4. The annotation ontology map is used to translate the gene annotations in the RAST-annotated genome into a list of metabolic reactions with associated gene–protein-reaction (GPR) rules that govern how gene activity impacts reaction activity (Fig. 3c). A list of spontaneous reactions (e.g., $CO_2 + H_2O \geq HCO_3$) is added to every core model constructed by the Model SEED, as these reactions occur regardless of the genes present in the genome. These reactions form the metabolic pathways of the core model.

This completes the reconstruction of the metabolic pathways and gene–protein-reaction associations for the metabolic model. The next step in the reconstruction process is to build a draft biomass composition reaction for the model. This step is described in detail in the next portion of this chapter (Subheading 3.6).

3.6. Construction of Draft Biomass Objective Function in the Model SEED

The biomass composition reaction (BCR) describes the relative quantity of all small molecule metabolites that must be produced in order to generate 1 g of biomass. BCRs account for proteins, lipids, DNA, RNA, cell walls, and cofactors. The small molecule building blocks of microbial biomass vary significantly depending on the metabolic pathways utilized, the electron transport chain, and the cell wall type. For this reason, BCRs are automatically assembled in the Model SEED based on the genome annotation and a set of template BCR, generated manually for four classes of cell wall: gram negative, gram positive, Mycoplasma, and Archaea. As with the core model reconstruction, it is not necessary for the user to intervene at any point during the BCR generation process in the Model SEED. We describe the process here just to improve understanding of how this process works. It is important to note that an exact BCR cannot be automatically generated from genome sequence alone, because the relative quantity of biomass components cannot be predicted exactly from genome sequence. Thus, the coefficients for metabolites in automatically generated BCRs are only approximations, and experiments must be performed to obtain exact values for these coefficients (see Note 3 for more information on the BCR coefficients).

1. The first step in the BCR generation process is to identify which BCR template to use, which requires that the input genome be classified as gram negative, gram positive, Mycoplasma, and Archaea. This classification can be done manually via experimental characterization or literature search; it can be done automatically via phylogenetic characterization of the organisms based on 16S RNA; or it can be done based on the genome annotations, as the various cell walls are associated with different

cell wall biosynthesis subsystems. The third approach is the method used by the Model SEED; a list of functional roles specific to each cell wall type has been assembled, and we predict cell wall type by assessing which list of roles has more associated genes in the input genome annotation.

2. Once the organism cell wall type has been determined, a template BCR is selected based this typing. Template BCRs typically contain approximately 100 candidate metabolites for inclusion in the model BCR. Each candidate metabolite is associated with a set of conditions that must be satisfied in order for the metabolite to be included in the model BCR. Half of the metabolites in the template BCRs are typically universal, meaning they will be included in the BCR of every model. These include all amino acids, all nucleotides, all deoxynucleotides, and many common cofactors (e.g., NAD). The remaining half of metabolites (e.g., lipids, cofactors, cell wall components) are included in the model BCR only if the genome annotation includes evidence for functional roles associated with the biosynthesis or utilization of the metabolites.
3. Once all the metabolites in the model BCR have been determined, the stoichiometric coefficients for the metabolites must be computed. The model BCR should represent the metabolites consumed to produce 1 g of biomass, so coefficients are computed such that the net mass of metabolites consumed in the BCR add up to 1 g. The mass of metabolites produced as products in the BCR (e.g., ADP, phosphate, and H^+ generated by the ATP hydrolysis) is subtracted from this net mass when adding up BCR metabolite mass (see Note 4). To support this computation, each candidate metabolite in the template BCR is associated with a category: DNA, RNA, protein, lipid, cell wall, cofactor, or energy. The template BCR includes estimations of the fraction of biomass associated of each category and the mole fraction of each small molecule to each category. These estimations are based on the values reported for the representative organisms on which these templates are based (*E. coli*, *B. subtilis*, *M. barkeri*, and *M. genitalium*). As such, it is important to note that these values are approximations for the organism being modeled, and they need to be adjusted based on experimental data before precise quantitative predictions can be produced by the model.
4. The template biomass reactions all include an "energy" category that contains ATP and water as reactants and ADP, phosphate, and H^+ as products. These metabolites represent the growth-associated ATP consumption for the organism, and the coefficient on all of these terms in the biomass reaction represents one of the most important

adjustable parameters for fitting growth–yield predictions to experimental data in metabolic models. Growth-associated ATP consumption varies widely among published metabolic models, with values typically falling between 30 and 100 mol/g cell dry weight hour (24). As with all other coefficients in the template BCR, the growth-associated ATP consumptions included in the template BCRs are approximations garnered from representative organisms. We make special mention of the growth-associated ATP consumption coefficient, because it is the largest BCR coefficient by far and as such has the greatest influence on yield computations. When building a model to compute growth yield, one of the first steps should be to adjust the growth-associated ATP consumption coefficient to fit experimentally measured growth yields.

5. DNA is the one category of BCR components for which stoichiometry can be calculated directly from the DNA sequence, as this portion of the biomass represents the replication of the chromosome itself (see Note 5). DNA coefficients are calculated by first computing the GC content of the chromosome. Then the molar fractions of the deoxynucleotides are set according to the GC content (e.g., deoxyguanine = deoxycytosine = GC and deoxyadenine and deoxythiamine = 1−GC).

At the end of this process, we will have a draft BCR, which is the final piece required to have a complete genome-scale metabolic model. The BCR is a critical element to the predictive capacity of a metabolic model, and both the coefficients and the small molecular metabolite content of the BCR require curation and adjustment. The coefficients in the BCR directly impact the ability of the model to quantitatively predict growth yield; the small molecule metabolites included in the BCR directly impact the ability of the model to qualitatively predict cell viability in a variety of environmental conditions and with a variety of genetic perturbations. Depending on the type of predictions needed, curation of the BCR should be prioritized accordingly. Although the draft metabolic model is now complete, it will almost never include all of the reactions needed to produce every component of the BCR, meaning the model will still be unable to predict growth conditions at this time.

3.7. Auto-completion of a Metabolic Model in the Model SEED

The core metabolic model and biomass objective function that are assembled by the Model SEED during the model reconstruction process contain all the data required to begin the analysis of microbial phenotypes and capabilities using approaches such as flux balance analysis (13–16). However, even when working with the most well known of microbial organisms (e.g., *E. coli*, *B. subtilis*), these core models will be of limited use initially, as they will be lacking many of

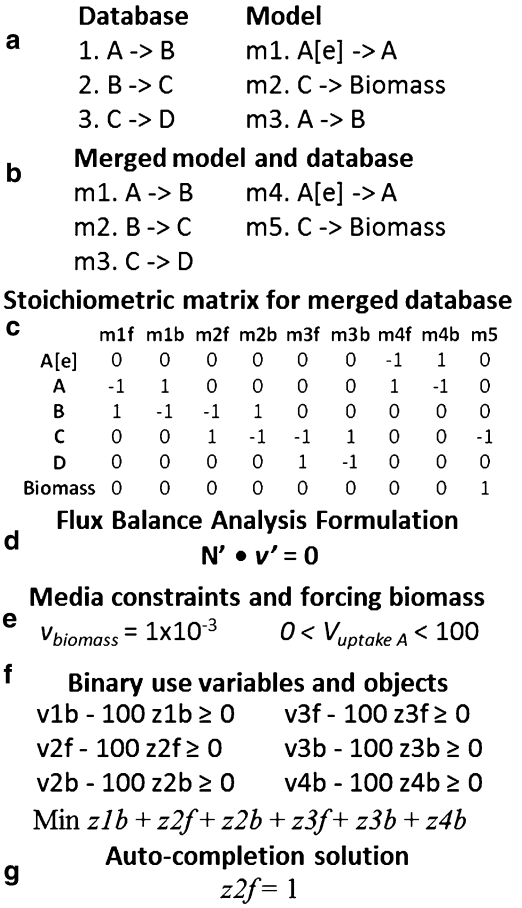

Fig. 4. Model auto-completion process.

the enzymatic steps required to produce all biomass building blocks including in the biomass objective function of the model. All core models generated by the Model SEED undergo an auto-completion process, whereby additional reactions are added to the model as needed to enable the production of all biomass components. The auto-completion process is automated by the Model SEED, but here we outline the steps of this process used by the Model SEED.

1. In the first step of the auto-completion process, a biochemistry database is prepared to serve as the source of reactions to be added to the core model to enable biomass production (Fig. 4a). Prior to use in auto-completion, all generic reactions, lumped reactions, and unbalanced reactions must be removed from the database, as these reactions can cause physiologically irrelevant pathways to be added by the auto-completion algorithm. In the Model SEED, the database used for auto-completion includes 10,516 reactions and 8,355 compounds.

2. Next, the auto-completion database is merged with the reactions and compounds of the core metabolic model while ensuring that identical compounds and reactions in the model and database are unified to produce a single nonredundant biochemical network (Fig. 4b). In the Model SEED, this process is simple as all reactions included in Model SEED models come from the Model SEED biochemistry database, meaning the models and database have a common namespace. When the model and biochemistry database namespaces are different, this merging process can be the most difficult step in the auto-completion process, due to inconsistencies in how compounds and reactions are named and represented.
3. Now the unified biochemical network and model BCR are translated into a stoichiometric matrix, where the columns are reactions, the rows are compounds, and the elements are the stoichiometric coefficients of the compounds in the reactions. As this matrix is formulated, every reaction is decomposed (regardless of reversibility) into separate forward and reverse component reactions (Fig. 4c), so that the flux through every component reaction is always greater than or equal to zero.
4. This matrix is then used to form the linear mass balance constraints of a flux balance analysis problem by setting the product of the stoichiometric matrix and the vector of fluxes through the forward and reverse component reactions to be equal to zero (Fig. 4d).
5. Next, an additional constraint is added to the linear optimization problem that forces the flux through the BCR to a positive nonzero value. Uptake and drain fluxes are also added for all metabolites that occur in the extracellular compartment. Bounds on these fluxes are adjusted as needed to represent the media conditions in which the auto-completion is being performed (Fig. 4e). Because the specific defined growth conditions for organisms are unknown, in the Model SEED auto-completion is performed in complete media, where uptake of all transportable metabolites is allowed (see Note 6).
6. To track which new reactions in the auto-completion database are to be used when flux is forced through the BCR, binary variables are associated with each component reaction that did not appear in the original model either because annotated reactions were irreversible or because no gene was annotated to perform the reaction. Each binary use variable is equal to "1" if its associated reaction is active and "0" otherwise. The objective function for the auto-completion optimization problem then becomes the minimization of the sum of the binary use

variables multiplied by a set of cost coefficients. Cost coefficients are computed for every component reaction based on thermodynamic feasibility, completion of existing pathways, the confidence in the biochemistry, and the amount of information available for the biochemistry (Fig. 4f). Costs are also commonly calculated based on blast scores for genes associated with the gap-filled reactions in other genomes (26).

7. Each solution to the auto-completion optimization problem (Fig. 4 g) represents a set of reactions that must be either added or made reversible in order to enable the metabolic model to produce biomass in the media condition selected for auto-completion (see Notes 7–8). The Model SEED will select a solution that best minimizes the auto-completion cost function, and that solution will be integrated into the core model as gap-filling reactions. However, it's important to note that many equivalent optimal solutions often exist for the auto-completion optimization, meaning the solution must be manually curated to determine if it is correct.
   Once the auto-completion process is complete, the metabolic model will be capable of producing biomass in the media condition in which the auto-completion was performed (typically complete media). At this stage, flux balance analysis may be used to generate qualitative predictions of essential genes, growth conditions, growth phenotypes, and metabolic capabilities. But this metabolic model is still a draft model, and substantial curation must be performed before the model is capable of generating accurate quantitative predictions. In the next section, the tools available for viewing and curating metabolic models will be explored, with an emphasis on the Model SEED website.

3.8. Reviewing and Curating a Model SEED Model

Once a model has been build and auto-completed in the Model SEED, the final step is to review the model and begin the process of curating the model. The Model SEED website (http://www.theseed.org/models) provides numerous interfaces for viewing metabolic model data and for comparing the model to other models in the Model SEED database. Below, we will walk through the Model Viewer interface of the Model SEED, the Cytoscape SEED interface for model viewing, and the methods for downloading model data for offline analysis.

1. The Model Viewer interface of the Model SEED is divided into two main frames: an upper frame and a lower frame. The upper frame contains tools for selecting and running analyses on models. It is split into six tabbed panes, and initially the tab labeled "Selected models and run FBA" is displayed. This tab contains a text box displaying the text "type here to see available models," which can be used to search for existing models

that are publically available. For example, if you type "k12" in this text box, you will see several publically available metabolic models for *E. coli K12*. To select your model for viewing, first make sure that you are logged into the Model SEED site. The login menu is located on the upper right-hand corner of the Model Viewer page. Next, type the name of the genome, genome ID, or name of the model you constructed in the model selection text box and find your model from among the options that appear in the filter select. Select your model and click the submit button. The Model Viewer page will reload with your model selected. see Note 9 for other methods of selecting models in the Model Viewer.

2. The upper frame will now contain some summary statistics for your model, including the number of genes, reactions, and compounds. The upper frame will also contain controls for running flux balance analysis on your model within the Model SEED environment. These will be discussed later. For now, we will review the bottom frame, where all other data related to your model will be displayed. The bottom frame is a tabbed display containing the following tabs: Map, Reactions, Compounds, Biomass Reactions, Genes, and Media formulations. Initially, the "Map" tab is selected, which enables selection of KEGG pathway maps (23) from a table showing the names of the pathway maps and the number of reactions, compounds, and EC numbers that occur in your model in each map. You can search from the list of available maps for viewing using the text box immediately below the "Name" header in the table. see Note 10 for general information about Model SEED tables. Once you've identified a metabolic map of interest, simply click on the map link. The metabolic map should immediately begin loading in the area beneath the map table. Note that you can open multiple maps at once by clicking on different map links in the map table. Maps can also be opened by clicking on map links in the metabolic maps themselves, in the reaction table, and in the compound table.

3. Scroll to the bottom of the page after a selected map has loaded. All of the reactions that are present in your model are highlighted. Hover the mouse over any reaction in the pathway map to see associated information, such as the reaction ID, the corresponding KEGG ID, and the associated gene ID. The KEGG pathway maps will only highlight reactions from the metabolic model that correspond to the KEGG reactions in the map. Metabolic models sometimes contain reactions that do not have corresponding KEGG reaction IDs; these reactions will not be displayed in the KEGG pathway maps. This may cause apparent gaps in the pathway maps (e.g., the pyruvate dehydrogenase complex in the iJR904 glycolysis/gluconeogenesis map).

To determine whether the gap is apparent or real, select the "Reactions" tab in the bottom frame of the Model SEED webpage and type the name of a compound adjacent to the gap (e.g., "pyruvate") in the text box below the "Equation" column header and press the "enter" key. The table will show all reactions containing that compound as a substrate or product (e.g., rxn00154 which represents a condensed version of the reaction catalyzed by the pyruvate dehydrogenase complex). The map view is the first location where model curation can begin. It's useful to examine the maps associated with the central metabolic pathways. Look for pathways where many reactions are present and only single steps are missing. These pathways are prime candidates for additional gap filling. When filling gaps in this manner, look at the gene page in the SEED for genes associated with reactions around the gap. These regions of the chromosome often contain the genes that may be associated with the reaction gap. Also look for reactions that are isolated from the rest of the model in the metabolic maps. These "island" reactions are prime candidates for removal.

4. Now select the "Reactions" tab on the lower frame of the Model Viewer. Once the tab finishes loading, you should see a table of all reactions currently included in your model. This table includes Model SEED reaction IDs (which you can click on to load a separate reaction page with images of all reactants), reaction names, reaction equations (you can click on compound names to load a separate compound page showing structure and listing all reactions the compound takes part in), functional roles and subsystems mapped to the reaction in the SEED ontology (these mappings were used to generate the core model as described in Subheading 3.5), KEGG maps (click on these links to load the associated map in the "Maps" tab), EC numbers, KEGG IDs, notes for the reaction in your model, and a list of the genes mapped to your reaction in your model (full GPR rules for reactions are shown on the reaction pages). This view is the best location to systematically see all the reactions that were added to the model during the auto-completion process. Simply go to text box in the model column of the reaction table, type "Gapfilling," and press enter. The table will now list all reactions that were gap filled in the model. Now critically examine each gap-filled reaction. If the reaction appears to be a correct addition to the model, go to the KEGG maps associated with the gap-filled reaction and repeat the curation steps described in step 3. If it appears that the gap-filled reaction is wrong, consider why that gap-filled reaction was added. For example, if a folate transporter was incorrectly added, this still indicates that the folate biosynthesis pathways

in the model are incomplete. This provides a hint as to which correct gap-filling reactions must be added.

5. Now select the "Compounds" tab on the lower frame of the Model Viewer. This tab contains a table of all compounds currently included in your model. This table includes Model SEED compound IDs (which you can click on to load the compound page), compound names, molecular formula, molecular weight, molecular charge, KEGG maps (click on these links to load the associated map in the "Maps" tab), KEGG IDs, and an indication of the compartments where each compound appears in the model. Note that all molecular properties displayed for compounds in this table were computed at pH 7.
6. Now select the "Biomass Components" tab on the lower frame of the Model Viewer. This tab contains a table of all compounds currently included in the biomass reaction of your model. This table includes the compound ID, names, formula, mass, charge, KEGG map, KEGG ID, and coefficient for each compound in the biomass reaction. The entries in the biomass component table should be reviewed, with special attention paid to the cofactors, lipids, and cell wall components included in the BCR. Compare the BCR of the draft model with BCR from other models of phylogenetically close organisms to identify if important components were excluded or if components were included that should not be there.
7. Now select the "Genes" tab on the lower frame of the Model Viewer. This tab contains a table of all genome features that appear in the annotated genome for which the model was reconstructed. This table includes gene IDs (click on these to go to the gene page in SEED), start, length, direction of transcription, functional annotation, predicted and experimental essentiality, and a list of the reactions mapped to each gene in your model. This is the best place to view the entire genome annotation that was used to assemble the model. Explore the annotation, check the gene calls, and look for large blocks of unannotated genes that might be indicative of a problem. These issues can then be resolved using the annotation curation tools available in RAST and PubSEED (see Subheading 3.3).
8. Finally, select the "Media formulations" tab on the lower frame of the Model Viewer. This tab contains a table of all media formulations currently loaded into the Model SEED database. The table includes media IDs and a list of media compounds in terms of names and compound IDs. This table is useful for quickly identifying the media conditions on which you want to simulate for model using flux balance analysis.

9. To run flux balance analysis on your model, return to the "Selected models and run FBA" tab of the upper frame of the Model Viewer website. Below the model information in this tab is a header entitled "Click here to run FBA on selected models" – click on this header to reveal a text box where you can select from a set a predefined media formulations. The default media condition for FBA is "Complete," meaning that all compounds for which the metabolic model has transport reactions are present in the medium. Click the "Run" button, and the top frame will switch to the "Flux Balance Results" tab. After FBA has completed, the upper frame will display a table containing the FBA results. If the model predicts growth on the selected medium, the "Growth" column will display the growth rate. Select the corresponding checkbox in the "Select" column and click on "View Selected Results." The "Reactions" tab in the bottom frame is updated to display the flux for each reaction in the rightmost column.
10. Another mechanism for viewing model content is the Cytoscape SEED plugin for Cytoscape. The CytoSEED viewer (27) for Model SEED models provides a more flexible environment for viewing metabolic models. CytoSEED is a plugin for the Cytoscape biological network viewer (28), and instructions for installing and using CytoSEED are available at http://www.cs.hope.edu/cytoseed/.
11. One of the most powerful mechanisms for model curation is the comparison of your model with the model of another more well studied by phylogenetically close organism. The Model SEED site facilitates such comparison by enabling the user to select multiple models at once. To compare models, simply use the model filter select in the "Selected models and run FBA" tab of the upper frame of the Model Viewer website to select another model. Once the desired model is selected, once again click on the "Select model" button. The Model Viewer site will now reload, but this time, both your model and the new model will be selected at the same time. You can use this mechanism to select up to five models at once for side-by-side comparison. All views described above will then contain data for all selected models.
12. Model SEED models may also be downloaded to enable their use with other available flux balance analysis platforms. In the "Selected models and run FBA" tab of the upper frame of the Model Viewer website, the model information table includes a series of download link in the far right-hand column. Three formats are available for model download: LP format, SBML format, and Excel format. The LP format defines the constraints, variables, and objective function for the linear optimization problem defined by running flux balance analysis on your model. LP files can be manipulated and simulated using command line

interfaces for most linear optimization solvers (e.g., CPLEX, GLPK, SCIP). The SBML format is a standard format for metabolic models. These files can be loaded into FBA software platforms such as the COBRA Toolbox (29) or OptFlux (30) for simulation and analysis. These FBA software packages feature numerous studies that apply the model to predict phenotypes and behavior. They also include algorithms to refine a model based on experimental data (8). The Excel format download includes three worksheets containing the reaction, compound, and gene data for the model. This format mirrors much of the data displayed on the Model Viewer website.

At this stage, the draft Model SEED model can be applied to predicting phenotypes that may be compared with phenotypic data available in the literature. Whole-genome transposon mutagenesis, gene knockout studies, and Biolog phenotyping arrays are all valuable tools for model testing and validation. Powerful computational techniques such as GrowMatch (8) now exist for reconciling metabolic models with experimental phenotype data. In addition to testing the capacity of the model to correctly predict microbial viability, the model can also be applied to predict growth yields, which can be compared to experimentally observed growth curves. Prior to attempting to predict growth yields, first ensure that the ATP biosynthesis mechanism being employed by the model is physiologically reasonable. Often the auto-completion process will produce models that generate ATP in physiologically unreasonable ways. Correct auto-completion of electron transport chains in metabolic models remains an open problem and an active area of work. Finally, take advantage of the comparative tools in the Model SEED that compare the draft models with other more well-curated models. Comparison with published models of similar organisms is the fastest way to identify and correct errors that occur in the annotation and automated model reconstruction process.

A detailed protocol (10) does exist for manually creating new genome-scale metabolic models. In this chapter, we have focused on the tools and steps taken to automatically produce a draft metabolic model, but this reconstruction protocol remains one of the best resources available for any researcher contemplating the development of a new metabolic model. We recommend reading this protocol in detail, comparing the protocol with the content of this book chapter, and using the protocol as a guide to the process of curating and revising your draft metabolic model.

4. Notes

1. The biochemistry database used by the Model SEED to construct metabolic models (see Subheading 3.5) is available for browsing at http://seed-viewer.theseed.org/models. If no model is selected, the "Reactions" and "Compounds" tabs will contain all reactions and compounds available in the database (13,000 and 16,000 entities, respectively). Where possible, we have also preserved connections to the KEGG pathway maps, which may be viewed under the "Maps" tab. Additional information about each reaction is available through the link on the reaction id, e.g., "rxn00123." This page includes thermodynamic reversibility, Gibbs free energy estimates, and a table of alternate names under the "Database Links" tab. A similar set of information is available for compounds under the compound ID, e.g., "cpd00456."
2. There is free software available, called MarvinBeans (http://www.chemaxon.com/products/calculator-plugins/), for predicting the predominant charged form of compounds at a specified pH. The software requires only a MOL file with the compound molecular structure as input. The command to determine the predominant charge at pH 7 is "cxcalc -N hi majorms -H 7 -f mol:-a example.mol > chargedExample.mol".
3. We mention that the specific metabolite coefficients in the BCR are only approximations garnered from representative organisms on which the template BCRs are based. The same is true for the specific metabolites included in the BCR as well, although we do strive to be as comprehensive as possible with that list. We also emphasize the need to refine the BCR metabolite coefficients based on experimental data before precise quantitative flux predictions can be obtained. Here, we note the substantially greater importance of obtaining correct values for the growth-related ATP consumption over obtaining correct values for specific biomass composition of proteins, RNA, cofactors, lipids, and cell walls. This is important to emphasize, because calculating growth-related ATP consumption requires only simple growth curves measured in a variety of growth conditions. Full biomass compositions can be much more challenging to obtain exact values and have less value in models.
4. Why are there products besides biomass in my Model SEED BCR? Often products other than biomass are added to the BCR of metabolic models. In some cases, this is done for modeling reasons. For example, two of the most common products in BCR are ADP and phosphate, which are present because they

are the products of hydrolysis of ATP and water. Most BCR include the reactants and products of ATP hydrolysis to represent the energy consumed by the synthesis of biomass from small molecule metabolites. For example, DNA, RNA, and protein polymerization reactions all hydrolyze ATP into ADP to provide energy for the hydrolysis process. In other cases, products are added to biomass to recycle metabolites that are consumed during the synthesis of biomass metabolites, but for which no biosynthesis pathways exist. Examples are the protein components attached to CoA and ACP (the protein components are attached in the metabolic pathways, so we can capture the essentiality of the enzymes that perform this ligation step) or the molecule dimethylbenzimidazole, for which the biosynthesis pathway is unknown.

5. Students learning about metabolic modeling and biomass objective functions often ask why amino acid and RNA nucleotide BCR coefficients cannot be calculated from the direct translation of the DNA sequence as is done for DNA deoxynucleotide coefficients. This view would be correct if all genes and proteins were expressed by the cell in equal quantities at all times. Unfortunately, this is not the case, and in fact, it is so far from reality as to be an extremely poor assumption to make.
6. In the model auto-completion process, choosing a minimal media for the auto-completion is preferred as it reduces the available solution space during the auto-completion optimization. This will result in more specific gap-filling predictions, and it will produce a model that is more likely to grow in many of the observed growth conditions.
7. The model auto-completion process involves solving a large mixed-integer linear optimization problem with potentially tens of thousands of binary variables. The most common open-source optimization package, GLPK, is not capable of solving this problem in a reasonable amount of time. Alternative MILP optimization packages must be used. Open-source alternatives include SCIP (http://scip.zib.de/), CBC (http://www.coin-or.org/projects/Cbc.xml), and SYMPHONY (http://www.coin-or.org/projects/SYMPHONY.xml). Commercial software includes CPLEX (www.ibm.com/software/integration/optimization/cplex-optimizer/). In our experience, the CPLEX solver outperforms all others by a significant margin, although all the open-source solvers we list above will solve most auto-completion problems in less than 24 h.
8. When solving the mixed-integer optimization problem during the auto-completion process, extreme caution must be taken

with the enforcement of the binary use variables. Although the problem constraints state that the flux through a component reaction must be zero if its corresponding use variable is zero, all optimization solvers have a tolerance setting that dictates the maximum amount by which a constraint may be violated. This tolerance can provide numerical "wiggle room" that allows reactions to carry a small amount of flux while keeping the binary use variable at zero. This small amount of flux can be sufficient to satisfy the minimal flux through the biomass reaction, while in fact, the current solution is not feasible. To avoid this problem, we recommend performing auto-completion with very low tolerance settings (e.g., 1×10^{-9}) and with large bounds on reaction flux and metabolite uptake (e.g., 10,000).

9. Some of the other tabs in the upper frame of the Model Viewer website provide alternative mechanisms for loading model summary statistics and selecting models for detailed viewing. The "User models" tab contains a table of all private models currently owned by the logged in user. This tab is a useful mechanism for quickly checking the status and availability of all your private models in the Model SEED. The "Model Statistics/select" tab includes a table of all models in the Model SEED that you currently have access to (including public models and private models). This tab is useful for quickly viewing, comparing, and querying metadata for all Model SEED models.
10. In general, all tables in the Model SEED website have controls for searching and sorting based on columns. When a text box is present immediately below a column header, type your search text in the text box and press the "enter" key; the table contents will be filtered to display only the entries that match your search text. The table can be restored by deleting the search text from the text box and pressing the "enter" key. Additionally, some table headers have two arrows immediately to their right; click the "up" arrow to sort in ascending order and the "down" arrow to sort in descending order.

Acknowledgements

We acknowledge the entire SEED, Model SEED, and CytoSEED teams at Argonne National Laboratory, Fellowship for Interpretation of Genomes, Hope College, and University of Chicago for efforts on the frameworks described in this chapter. This work was supported by the US Department of Energy under contract DE-

ACO2-06CH11357 (SD, CH), the National Institute of Allergy and Infectious Diseases under contract HHSN266200400042C (RO), and the National Science Foundation under grants MCB-0745100 and DBI-0850546 (MD, AB, VV, RO).

References

1. Feist AM, Palsson BO (2008) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. Nat Biotechnol 26:659–667
2. 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 1672:1–6
3. 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
4. Overbeek R, Disz T, Stevens R (2004) The SEED: a peer-to-peer environment for genome annotation. Commun ACM 47:46–51
5. DeJongh M, Formsma K, Boillot P, Gould J, Rycenga M, Best A (2007) Toward the automated generation of genome-scale metabolic networks in the SEED. BMC Bioinformatics 8:139
6. Jankowski MD, Henry CS, Broadbelt LJ, Hatzimanikatis V (2008) Group contribution method for thermodynamic analysis of complex metabolic networks. Biophys J 95:1487–1499
7. Henry CS, Zinner J, Cohoon M, Stevens R (2009) *i*Bsu1103: a new genome scale metabolic model of *B. subtilis* based on SEED annotations. Genome Biol 10:R69
8. Kumar VS, Maranas CD (2009) GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. PLoS Comput Biol 5:e1000308
9. Suthers PF, Dasika MS, Kumar VS, Denisov G, Glass JI, Maranas CD (2009) A genome-scale metabolic reconstruction of *Mycoplasma genitalium*, *i*PS189. PLoS Comput Biol 5: e1000285
10. Thiele I, Palsson B (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc 5:93–121
11. Schuler GD, Epstein JA, Ohkawa H, Kans JA (1996) Entrez: molecular biology database and retrieval system. Methods Enzymol 266:141–162
12. Edwards JS, Palsson BO (2000) The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci U S A 97:5528–5533
13. Papoutsakis ET, Meyer CL (1985) Equations and calculations of product yields and preferred pathways for butanediol and mixed-acid fermentations. Biotechnol Bioeng 27:50–66
14. Jin YS, Jeffries TW (2004) Stoichiometric network constraints on xylose metabolism by recombinant *Saccharomyces cerevisiae*. Metab Eng 6:229–238
15. 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 60:3724–3731
16. Varma A, Palsson BO (1993) Metabolic capabilities of *Escherichia coli*. 2. Optimal-growth patterns. J Theor Biol 165:503–522
17. Varma A, Palsson BO (1993) Metabolic capabilities of *Escherichia coli*.1. Synthesis of biosynthetic precursors and cofactors. J Theor Biol 165:477–502
18. Edwards JS, Ibarra RU, Palsson BO (2001) In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. Nat Biotechnol 19:125–130
19. Meyer F, Overbeek R, Rodriguez A (2009) FIGfams: yet another set of protein families. Nucleic Acids Res 37:6643–6654
20. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27:4636–4641
21. Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23:673–679
22. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

23. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30
24. Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO (2009) Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol 7:129–143
25. Kummel A, Panke S, Heinemann M (2006) Systematic assignment of thermodynamic constraints in metabolic network models. BMC Bioinformatics 7:512
26. Krumholz EW, Yang H, Weisenhorn P, Henry CS, Libourel IG (2012) Genome-wide metabolic network reconstruction of the picoalga Ostreococcus. J Exp Bot 63:2353–2362
27. DeJongh M, Bockstege B, Frybarger P, Hazekamp N, Kammeraad J, McGeehan T (2012) CytoSEED: a Cytoscape plugin for viewing, manipulating and analyzing metabolic models created by the Model SEED. Bioinformatics 28:891–892
28. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27:431–432
29. 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 2:727–738
30. Rocha I, Maia P, Evangelista P, Vilaca P, Soares S, Pinto JP, Nielsen J, Patil KR, Ferreira EC, Rocha M (2010) OptFlux: an open-source software platform for in silico metabolic engineering. BMC Syst Biol 4:45

http://www.springer.com/978-1-62703-298-8

Systems Metabolic Engineering
Methods and Protocols
Alper, H.S. (Ed.)
2013, XII, 474 p. 61 illus., 47 illus. in color. With online files/update., Hardcover
ISBN: 978-1-62703-298-8
A product of Humana Press