Published in final edited form as:

Methods Enzymol. 2011 ; 500: 411–433. doi:10.1016/B978-0-12-385118-5.00021-9.

Whole-genome metabolic network reconstruction and constraint-based modeling

Charles R. Haggart^{*,†,1}, Jennifer A. Bartell^{*,1}, Jeffrey J. Saucerman^{*,†}, and Jason A. Papin^{*,2}

^{*}Department of Biomedical Engineering, University of Virginia, Charlottesville, Virginia, USA

[†]Robert M. Berne Cardiovascular Research Center, University of Virginia, Charlottesville, Virginia, USA

Abstract

With the advent of modern high throughput genomics, there is a significant need for genome-scale analysis techniques that can assist in complex systems analysis. Metabolic genome-scale network reconstructions (GENREs) paired with constraint-based modeling are an efficient method to integrate genomics, transcriptomics, and proteomics to conduct organism-specific analysis. This text explains key steps in the GENRE construction process and several methods of constraint-based modeling that can help elucidate basic life processes and development of disease treatment, bioenergy solutions, and industrial bioproduction applications.

1. Introduction

The rapid expansion of methods to utilize organism-specific whole genome sequences provided by high throughput sequencing technology has provided clarity in areas scientists have long puzzled over and improved our ability to probe for answers on a system-wide level in addition to our classic reductionist investigative strategy. The substantial amounts of data now being harvested on phenotypic, genetic, protein, and molecular scales is driving the development of computational systems analysis ever faster in a search for ways to organize and contextualize raw data into a coherent picture. One promising area within the field of systems biology is reconstruction of organism-specific genome-scale metabolic networks that has been accompanied by the development of a wide range of constraint-based modeling approaches. The metabolic genome-scale network reconstruction (GENRE) provides a framework to organize all available information about an organism's metabolism through careful construction and curation of a computational network that links the cell's genome and gene expression to metabolic reaction fluxes, biomass and energy production and consumption. Once completed, the power of each GENRE can be realized as a model for probing a cell's genotype-phenotype relationship via constraint-based modeling.

Building an organism-specific GENRE is a promising opportunity to improve metabolic understanding. If the reconstructed network model can be validated through experimental investigations of growth rates, phenotypes, reaction fluxes, and gene expression, it may be immediately useful as a method for drug discovery or the development of strategies for optimizing byproduct metabolite yields as examples. If the model predictions vary from experimental validation, the discrepancies offer the researcher a roadmap for iterative *in vivo* experimentation, *in silico* modeling, and annotation refinement that further the understanding of the organism's metabolism. Also, the GENRE itself, if well-constructed

²Corresponding author: papin@virginia.edu, phone: 434-924-8195, fax: 434-982-3870.

¹These authors contributed equally to this work

and curated, is a directed method of collecting and organizing all currently available knowledge of the organism's metabolism in a functional manner.

A cell's phenotype is subject to the constraints imposed by its genome and extracellular environment. By quantifying these constraints in an appropriate manner, one can predict the range of possible cellular phenotypes *in silico* without acquiring large quantities of kinetic metabolic data. While constraint-based modeling of metabolic networks is far more quantitative than interaction-based (*i.e.*, graph-based) network models (Papin *et al.*, 2003), it does not allow for characterization of the dynamics of metabolism with the same level of confidence as a mechanistic, kinetic model (Raman and Chandra, 2009). However, given the lack of experimentally measured metabolic reaction parameters for most organisms, kinetic modeling is not a widely applicable approach to modeling cellular biochemical networks. Constraint-based modeling presents an alternative approach to better understanding and predicting the behavior of a biochemical network under various environmental and genetic perturbations.

While excellent reviews of the reconstruction process are available (Durot *et al.*, 2009; Feist *et al.*, 2009; Rocha *et al.*, 2008), including highly detailed protocols for reconstructing genome-scale metabolic networks (Thiele and Palsson, 2010), the present work focuses on providing an introduction to the reconstruction process, including important methods, tools, and validation techniques, as well as the constraint-based modeling approaches that utilize the curated information within each GENRE.

2. Metabolic Network Reconstruction

Metabolic network reconstructions are comprised of the stoichiometry of reactions necessary for nutrient usage in synthesis and degradation of basic metabolites and more complex compounds, specific genes whose protein products are associated with these biochemical reactions, and supporting annotation and literature references. This list can then be converted to mathematical form and combined with constraint-based modeling approaches detailed in the second half of this manuscript to predict metabolic phenotype. The reconstruction of a metabolic network of an organism with a completed genome sequence consists of the following steps: 1) genome annotation, 2) automated network reconstruction, 3) network refinement, 4) *in vitro* experimentation, and 5) gap analysis. These steps are often completed concurrently or iteratively to improve the accuracy of the model as shown in Figure 1.

2.1 Genome annotation

Since the first genome was sequenced via high-throughput technology in 1998, technological advancement has decreased the required cost and time to obtain a full genome sequence (Kircher and Kelso, 2010). However, for completed genome sequences to be useful, they need to be annotated using standardized gene ontologies that provide consistency while identifying genes and cataloging their function and regulation in relation to biological processes (Giglio *et al.*, 2009). Automated annotation programs (ERGO, J. Craig Venter Institute, Integrated Microbial Genomes) provide annotations that usually require manual curation to supplement organism-specific depth and detail as well as verify the automated annotation assignments (Yang *et al.*, 2010). Some effort has been made to design methods for evaluating annotation quality dependent on the degree of sequence match and phylogenetic relatedness to other annotated genomes (Yang *et al.*, 2010). Alignment tools such as BLAST are commonly used to annotate gene function based on orthology with previously annotated genomes provided in online databases, followed by other more intensive methods relying on phylogenetic grouping (Kuzniar *et al.*, 2008). Both general and organism-specific databases of genome annotations are available as listed in the

appropriate section of Table 1, which also details other useful tools in the reconstruction process.

2.2 Automated network reconstruction

Most metabolic reconstruction efforts begin with an automated draft reconstruction from a publicly annotated genome available online. Much effort is being directed to development of comprehensive automation suites. One such effort, the Model SEED project, moves from an unannotated, sequenced genome to a draft metabolic network with gap filling and verification features. Programs such as Pathway Tools and others listed in Table 1 require a previously annotated genome to build a draft reconstruction and offer various visualization and analysis tools. These programs produce at minimum lists of genes, their associated reactions, and corresponding Enzyme Commission (EC) numbers. It is also possible to use appropriate ontology keywords to retrieve core metabolic genes and reactions from an annotation. However, regardless of the automated tool used, manual reconstruction efforts are always necessary to ensure organism-specific metabolic characteristics are included (DeJongh *et al.*, 2007; Karp *et al.*, 2010).

2.3 Network refinement

Manual reconstruction begins with an initial evaluation of the completeness of the automated draft reconstruction. Each reaction must be evaluated for its necessity to the model, accurate stoichiometry, reaction direction and reversibility, and role in metabolite production, usage and recycling. Other concerns include thermodynamic feasibility and energy constraints. Integral to a successful reconstruction, this manual evaluation can take a significant amount of dedicated work to complete.

2.3.1 Key organizational tools—Research groups commonly use the few techniques listed below to speed the reconstruction process and provide consistency between different reconstruction projects.

(i) Spreadsheet organization: A thorough and well organized annotation record, stored in a spreadsheet or as a file in Systems Biology Markup Language (SBML) format with embedded annotations (Herrgard *et al.*, 2008), should include the gene name, all pertinent gene abbreviations, the reaction equation with consistent reactant, substrate, and product symbols, balanced stoichiometry, an indicator of reversibility, the metabolic subsystem that utilizes the reaction, the associated protein and its EC number, any literature references used to adjust or add the reaction, a rating of confidence in the annotation entry and comments about any adjustments or questions about the entry.

(ii) Reaction confidence level: In curating the reconstruction, it is helpful to keep track of the quality of the various annotations and changes. A standard method is to apply a classification system to each entry, ranging from class 1 (high confidence in the data) to class 4 (low confidence). This judgment is still qualitative, so consistency in class assignment within the reconstruction is important. When adding a reaction based solely on sequence matches with another organism, the confidence level of the added reaction should be based on the phylogenetic relatedness of the organism. It is also vital to record the addition of enzymes that have not been associated with a particular gene in the organism's sequence data but are known to be necessary in a certain pathway. Examples of metabolic reconstructions that have utilized reaction confidence ratings include iMO1056 (Oberhardt *et al.*, 2008) and the human metabolic network reconstruction (Duarte *et al.*, 2007).

2.3.2 Organism-specific curation—After canonical metabolic reactions have been collected via the automated draft reconstruction, a review of literature pertaining to the

metabolism and function of the organism is necessary to identify organism-specific characteristics that should be integrated. Reconstructions of related organisms can be searched for specific functionality such as substrate usage preference and unique metabolic products that may be replicated in the organism of interest. For example, the reconstruction of *Aspergillus niger* included pathways related to steroid synthesis based on the biology of other aspergilli fungi, while a gap in lipopolysaccharide synthesis in the reconstruction of *Pseudomonas aeruginosa* was filled using gene similarity in the *Pseudomonas fluorescence* annotation (Andersen *et al.*, 2008; Oberhardt *et al.*, 2008).

(i) Compartmentalization and exchange reactions: When building the reconstruction, care must be taken to check the transport and exchange potential of a particular metabolite so that appropriate limits can be implemented as constraints on model behavior. Localization and transport of a metabolite between the cytosol and organelles or within the cytosol may be complicated, particularly in eukaryotic organisms, and the order of transport may be critical to metabolic processes (Hao *et al.*, 2010). Exchange reactions can be implemented that account for the movement of metabolites across intracellular and intercellular membranes, creating compartments within the *in silico* network to represent these physical boundaries.

(ii) Gene-protein-reaction relationships: While data for the reconstruction is based on gene annotations, constraint-based modeling can proceed using solely the enzymes and associated reaction stoichiometries. However, to simulate the phenotypic effects of gene knockout (KO) experiments and enable integration of gene expression microarray data, genes can be connected to their associated enzymes. Boolean logic is often used to define this relationship. When multiple genes encode enzyme subunits that are each required for catalysis of a reaction, an “AND” statement is used to convey necessity of each gene for that reaction. Alternately, an “OR” statement is used to reflect the requirement of genes that encode isozymes that catalyze the same reaction independently of one another. Some enzymes may perform similar functions with unclear substrate specificity where it is difficult to correlate a particular gene with a particular pathway (Andersen *et al.*, 2008) and may require approximation during modeling. The incorporation of gene-protein-reaction relationships (GPRs) have been incredibly valuable to researchers interested in studying the impact of different growth states, environment, or degree of infection measured with microarrays and connecting the gene expression results to a modeled reconstruction to investigate changes in metabolism (Oberhardt *et al.*, 2010).

(iii) Biomass composition: Biomass, the key components from which an organism is built and sustains itself, may be determined from experimental measures, described here, or through more computational methods described in Section 3. DNA, RNA, lipids, fatty acids, cell wall components, cytosolic solutes and ions, and cofactors must be present in sufficient amounts to enable cell operation (Senger, 2010). Experimentally, biomass composition can be determined through fitting carbon content and amino acid content measured at different population levels and growth states (Gonzalez *et al.*, 2010). The number of factors that should be included varies with the organism’s metabolic characteristics, but even biomass composition of the minimal cell model *Mycoplasma genitalium* includes 61 different factors (Suthers *et al.*, 2009).

(iv) Thermodynamics: Assessing reaction thermodynamics can alleviate excessive constraining of the model’s directionality (Henry *et al.*, 2009). Directionality of reactions (reversible vs. irreversible) in the model and feasible rates of reaction are based on thermodynamic favorability determined from Gibbs free energy changes. Experimentally determined values can be found in literature, but available data is rarely sufficient to address

a genome-scale reconstruction and more often has been obtained for smaller sections of metabolism or major canonical pathways. Methods for computationally estimating the free energy values of reactions include group contribution methods that can be performed online at Web GCM using reaction information from KEGG and BRENDA (Jankowski *et al.*, 2008).

2.4 In vitro experimentation and validation

In conjunction with *in silico* reconstruction efforts that rely on annotations, databases, and literature review, valuable information can also be collected through *in vivo* and *in vitro* experimentation. Growing an organism on specific carbon sources via high throughput Biolog microplates or more traditional techniques is an easy method for initial model validation because carbon source can be limited *in silico*. The organism's carbon source usage and production of byproducts during growth can be investigated by measuring media metabolite concentration using high performance chromatography. Many organisms have also been examined to determine survival genes through mutagenesis studies that create single gene KO mutants (Cameron *et al.*, 2008; Gallagher *et al.*, 2007; Jacobs *et al.*, 2003). Phenotypic characteristics such as substrate usage and byproduct synthesis of a particular mutant can be predicted *in silico* and then compared to *in vitro* experiments to investigate and improve GENRE accuracy. Investigating reaction fluxes and static and dynamic localization of specific enzymes in a living cell for comparison and incorporation with a GENRE is possible via metabolic mapping of fluorescent substrates and cofactors within live cells (Van Noorden, 2010) and pulse labeling, using ^{13}C as a radiotracer to investigate distribution of labeled molecules (Niittylae *et al.*, 2009).

3. Constraint-based modeling methods

Thus far, we have introduced the necessary steps to assemble an *in silico* metabolic GENRE whose components include stoichiometric coefficients for the reactants and products of each metabolic reaction, a Boolean GPR rule-set which defines the genes associated with each reaction, and notation for reversible and irreversible reactions. Each of these constrains the behavior of the metabolic model they comprise and limits the attainable metabolic phenotype of the cell. An organism's metabolic phenotype is often quantified by growth rate, the amount of biomass added per unit time. Because the reactions that produce individual biomass components are integrated into the GENRE (*e.g.*, DNA, RNA, amino acids, lipids, carbohydrates), we can also characterize metabolic phenotype by the amount of flux (metabolite mass per dry organism mass per unit time) that is carried through these reactions in the GENRE. Constraint-based modeling aims to reduce the number of possible flux profiles and identify one that best predicts the metabolic phenotype of the organism under specified genetic and environmental conditions.

Constraints limit the number of possible flux profiles for a given organism to some finite number and additional constraints further reduce the flux space and allow us to make even more accurate *in silico* predictions of metabolic phenotype. These and other constraints are generally assigned to one of four categories: physico-chemical constraints (*e.g.*, mass and energy balance), environmental constraints (*e.g.*, temperature, pH, substrate availability), spatial/topological constraints (*e.g.*, organelle compartmentalization), and self-regulatory constraints (Price *et al.*, 2004). The rapid growth of complete genome sequencing and curation of metabolic GENREs and the utility of constraint-based modeling have driven the improvement and innovation of methods for constraint-based modeling. Here, we will introduce a range of these methods developed for constraint-based modeling of genome-scale metabolism using GENREs constructed with the methods detailed in Section 2. In particular, we will focus on the quantitative constraints and motivation for method development.

3.1 GENRE-to-model implementation

To test model predictions against experiments, the constraining elements of the GENRE are given mathematical structure, primarily within the sparse S matrix, which quantifies the stoichiometric relationship between reactant and product metabolites for each reaction in the GENRE (Figure 2). The non-zero elements of S are the stoichiometric coefficients collected in the reconstruction process that correspond to the i^{th} metabolite within the j^{th} metabolic reaction, such that each of the m rows represents a metabolite ($M_{I,2,3}$) and each of n columns represents a reaction, including intracellular biochemical conversions ($R_{I,2}$), metabolite exchange (uptake, $X_{I,2}$; secretion, X_3), metabolite demand (cell maintenance, D_I), and inter-compartment metabolite transport (extracellular to cytosolic space, $T_{I,2}$). To account for a metabolite's presence in separate compartments, one row should be added to the S matrix for each compartment within which the metabolite exists. By convention, reactants have a negative coefficient and products have a positive one, while flux, v_i , is positive for forward biochemical conversions, transport reactions into the cell, and exchange reactions out of the cell (*i.e.*, secretion), and negative flux indicates a reverse intracellular reaction, an uptake exchange reaction and transport from the cytosol to another compartment. The GPR relationships between genes and reactions should be coded as a Boolean vector with an element for each reaction, while flux bounds and an objective function should be implemented in vector form, as described in the next section. One can carry out most of the constraint-based methods that follow with these few mathematical and Boolean structures.

3.2 Flux Balance Analysis

The most widely used constraint-based method in the field of metabolic engineering and systems analysis of metabolism is flux balance analysis (FBA), and most of the methods described herein are either explicit alternatives or add-on improvements to FBA. The number of review papers dedicated solely to FBA in recent years (Kauffman *et al.*, 2003; Lee *et al.*, 2006; Raman and Chandra, 2009) gives an idea of the wide applicability of this approach. Two recent articles provide greater detail to support carrying out an FBA (Oberhardt *et al.*, 2009) as well as a concise, visually aided description of what is achieved by performing an FBA (Orth *et al.*, 2010). FBA allows for the *in silico* prediction of a flux profile that optimizes some pre-defined cellular objective without making any experimental measurements. It has become a standard tool for simulating the effects of genetic perturbations or environmental conditions on an organism's growth rate or rate of byproduct synthesis (Raman and Chandra, 2009). FBA has also provided a means to simulate the metabolic effects of global gene expression and other large "omics" datasets for an organism or even a particular human tissue type (Becker and Palsson, 2008; Jerby *et al.*, 2010; Shlomi *et al.*, 2008).

The underlying theory for FBA assumes that cellular metabolism is at a steady-state (*i.e.*, constant growth rate) (Savinell and Palsson, 1992; Varma and Palsson, 1994), and all reaction fluxes (v_i) and metabolite concentrations (C_i) have reached a steady state, such that the net amount of each metabolite added to and removed from the cell must equal zero (Eq. 1). Given this, FBA uses linear programming (LP) to find the steady-state flux profile (v) that optimizes the organism's objective function (*e.g.*, maximal growth rate) while satisfying a set of physico-chemical and environmental constraints. While the objective function can be unique for each FBA problem, a common choice for microorganisms is maximization of biomass (*e.g.*, proteins, nucleic acids, carbohydrates, and lipids necessary for growth), based on the idea that evolutionary pressures select for higher growth rates of these organisms. In some organisms, however, alternative objective functions such as maximization (minimization) of ATP production (utilization) are more appropriate, as will be further discussed below. The primary physical constraint that an FBA problem considers is mass

balance, requiring that the amount of each metabolite that is (1) transported into or (2) produced within the cell be balanced by the amount of that metabolite that is (3) consumed within or (4) secreted from the cell. The mass balance constraint (Eq. 1) provides m metabolite balance equations to solve for n unknown reaction fluxes, and because there are usually more reactions than metabolites, the system of equations is under-determined. As such, the collection of flux vectors that satisfy the mass balance constraints span a multidimensional space, often referred to as the *null space* of the S matrix.

$$\frac{dC_i}{dt} = \sum_{j=1}^n S_{ij} \cdot v_j = 0 \quad \forall i \in M \quad (1)$$

$$\alpha_j \leq v_j \leq \beta_j \quad \forall j \in N \quad (2)$$

It is important to note that many of the flux profile vectors, v , in the null space are physically impossible to achieve (Orth *et al.*, 2010). In order to refine and reduce the size of the predicted flux space and exclude irrelevant flux profiles, the standard FBA implementation constrains individual reaction fluxes, v_j , by defining lower and upper bounds (α_j and β_j , respectively; Eq. 2). Some of these bounds are based on literature and experimentation and are collected in the GENRE process. These bounds can be applied to enforce environmental constraints, such as the availability of a substrate in growth media, by setting the upper and lower uptake exchange bounds to the same negative value (exchange flux is negative for uptake and positive for secretion). Similarly setting the lower and upper bounds on an exchange reaction to zero and some arbitrarily high value, respectively, will prevent metabolite uptake, while leaving the rate of secretion of a desired output metabolite unconstrained. Bounds can also be set on internal biochemical reactions by constraining the lower bound to zero to enforce thermodynamic irreversibility of a reaction, by setting the bounds to reflect the kinetic limitations of a particular enzyme or membrane transporter, or by constraining the flux to zero to simulate a gene KO. This approach is often used to identify genes essential for growth, both for identification of drug targets and maximizing the production of targeted byproducts (Feist and Palsson, 2008; Oberhardt *et al.*, 2009). Several of the methodologies described below impose more sophisticated or problem-specific constraints on the flux bounds to minimize the possible flux space and provide more accurate predictions of metabolic network behavior, though a standard FBA implementation typically enforces constraints on substrate uptake, reaction irreversibility, and flux capacity.

Once the mass balance constraint and flux bounds are set, an objective function (v_{obj} , Eq. 3) must be defined for the LP optimization problem. The optimization problem will find a flux profile, v , which maximizes or minimizes the objective function and satisfies the mass balance constraints and flux bounds described above. In those GENREs for which maximization of biomass production is the organism's objective, a "biomass reaction" is added as a column to the S matrix with stoichiometric coefficients in proportion to each metabolite's *in vivo* contribution to one unit of biomass determined as explained in Section 2.3.2 (Oberhardt *et al.*, 2009). To implement the chosen objective function, a "cost" vector (c) is defined, whose elements correspond to each reaction (*i.e.*, columns of S , elements of v). To maximize the desired reaction(s), all elements of the cost vector are set equal to zero except those corresponding to the chosen objective reaction(s) (Eq. 4). FBA identifies an optimal flux vector whose elements corresponding to the non-zero elements of the cost vector are maximized, subject to the defined constraints. For this example, the first and the second-to-last reactions comprise the objective function and would thus be maximized.

$$\max(v_{obj}) = c \cdot v \quad (3)$$

$$c = [1 \ 0 \ 0 \ \dots \ 0 \ 1 \ 0] \quad (4)$$

3.3 Dynamic Flux Balance Analysis

The steady-state assumption required of FBA is not appropriate for *in silico* modeling of organisms in all situations, most notably during a “diauxic shift” from one carbon source to another. To address this, Dynamic Flux Balance Analysis (DFBA) was developed, based on an earlier approach (Varma and Palsson, 1994), implementing both dynamic (non-linear programming, NLP) and static (LP) optimization of an objective function and the addition of constraints to the rates of change of flux in addition to the typical FBA constraints (Mahadevan *et al.*, 2002). The static optimization – a more tractable computational approach for large biochemical networks – performs a series of FBA problems at the beginning of each discrete time interval. Using both approaches, Mahadevan and colleagues correctly predicted the timing of acetate production and the sequence of substrate utilization (glucose before acetate) in *E. coli*, validating the use of this constraint-based method to simulate the dynamic behavior of genome-scale metabolism.

3.4 Flux Variability Analysis

The redundancy inherent in most biochemical networks corresponds to the possibility for numerous alternative flux distributions that yield the same maximal growth rate or ATP synthesis rate (Price *et al.*, 2002). Put in graphical terms, the multidimensional flux space of possible flux profiles often includes multiple points with an identical optimal objective function value. In metabolic pathway terms, different combinations of intermediate reaction fluxes give rise to the same flux through the objective reaction (*e.g.*, biomass), and each of these alternative pathways may be biologically meaningful. FBA is often used to identify a reference *in silico* flux profile with which to compare constraint-based behavior of a genetically or environmentally perturbed counterpart (see Minimization of Metabolic Adjustment below). As such, it is critical to understand flux variability, as two alternate FBA reference flux profiles could lead to a significantly different assessment of the perturbed state.

A recent report by Smallbone and Simeonidis introduced a computational method to identify a unique flux profile by analyzing the geometry of the multidimensional flux space defined by the standard FBA constraints (Smallbone and Simeonidis, 2009). This method identifies a minimal flux solution without any thermodynamically infeasible fluxes. An earlier approach by Mahadevan and co-workers called Flux Variability Analysis (FVA), uses constraints to assess the variability of each reaction flux toward an optimized objective flux value (Mahadevan and Schilling, 2003). FVA first uses FBA to determine the optimal objective flux (Eqs. 3–5), which is used with mass balance and flux bounds (Eqs. 1–2) as constraints for the subsequent one-by-one maximization and minimization of each reaction flux within the GENRE (Eqs. 6–7). This method quantifies the maximum and minimum fluxes for each reaction that are consistent with the optimal objective function value.

$$v_{obj} = \max(c_j \cdot v_j) \quad (5)$$

$$\max v_j, \quad \forall j \in N, j \neq obj \quad (6)$$

$$\min v_j, \quad \forall j \in N, j \neq obj \quad (7)$$

3.5 Minimization of Metabolic Adjustment

Flux balance analysis assumes that each organism's metabolic network has been tuned through evolution for some objective function, be it maximal growth rate or energy efficiency (*e.g.*, minimal ATP utilization). While this assumption may be valid for wild-type (WT) organisms that have evolved over many hundreds or thousands of generations, it may be less appropriate for engineered mutants, genetically modified in a controlled laboratory environment and unexposed to the same evolutionary forces or number of reproduction cycles. As such, Segre and colleagues developed Minimization of Metabolic Adjustment (MOMA), hypothesizing that mutant organisms are unable to immediately adapt their metabolic network to achieve the WT objective function, but instead display some suboptimal flux profile intermediate to the FBA-determined optima of both the WT (WT-FBA) and genetically perturbed (KO-FBA) organisms (Segre *et al.*, 2002). This approach finds a suboptimal flux profile (v) that is a minimal Euclidean distance from the WT-FBA flux profile (v^{FBA} , Eq. 8), and is mathematically formalized as a quadratic programming (QP) problem (Eq. 9), subject to the standard FBA mass balance and flux capacity constraints (Eqs. 1–2), including constraining all KO gene-associated reaction fluxes to zero (Eq. 10). Ultimately, FBA combined with MOMA provides a more accurate prediction of the immediate metabolic response to KO than FBA does on its own (Segre *et al.*, 2002).

$$D(v^{FBA}, v) = \sqrt{\sum_{j=1}^N (v_j^{FBA} - v_j)^2} \quad \forall j \in N \quad (8)$$

$$\min (v^{FBA} - v)^T (v^{FBA} - v) \quad (9)$$

$$v_k = 0, \quad \forall k \in A \quad (10)$$

3.6 Regulatory On/Off Minimization

The Euclidean distance metric introduced in MOMA finds a sub-optimal KO-associated flux distribution with many component fluxes slightly altered from their WT-FBA state. Shlomi and colleagues developed an alternative method, Regulatory On/Off Minimization (ROOM), which seeks a flux profile for the mutant organism with a minimal number of significant flux changes from the WT-FBA flux profile (Shlomi *et al.*, 2005). This method is based on the assumptions that a KO organism will minimize the costs associated with adapting its gene regulatory network, and that these costs are quantifiable – using a Boolean on/off framework – and independent of the magnitude of change in gene expression and associated reaction flux.

ROOM does not explicitly incorporate any gene regulatory constraints into the model, but accounts for them via a binary cost variable for each reaction (y_j set to one for fluxes that are significantly different from the WT-FBA; Eq. 12). This quantification of an altered flux is assumed to reflect a concordant change in expression of the necessary enzyme-encoding gene, and the associated cost to the organism of regulating that gene's expression. The objective of this mixed-integer linear programming (MILP) problem is to minimize the number of fluxes in the mutant GENRE that are different from their WT-FBA counterparts

(Eq. 11), while satisfying the FBA mass balance, irreversibility, and flux capacity constraints (Eqs. 1–2), and constraining the set of fluxes (A) associated with the given knockouts to zero (Eq. 10). The upper and lower thresholds by which flux changes are deemed significant (v_j^{FBA+} and v_j^{FBA-} , respectively) are user-defined via relative (δ) and absolute (ϵ) flux tolerance terms (Eqs. 13–14). Accordingly, if y_j is set to zero, v_j is constrained by these bounds, while setting y_j to one adds to the cost sum (Eq. 11) and leaves v_j unbounded (Eqs. 15–16). This method was shown to improve predictions of flux and steady-state growth rate over MOMA, though it performed very similarly to FBA (Shlomi *et al.*, 2005).

$$\min \sum_{j=1}^n y_j \quad (11)$$

$$y_j \in \{0, 1\} \quad (12)$$

$$v_j^{FBA+} = v_j^{FBA} + \delta |v_j^{FBA}| + \epsilon \quad (13)$$

$$v_j^{FBA-} = v_j^{FBA} - \delta |v_j^{FBA}| - \epsilon \quad (14)$$

$$v_j - y_j(v_{\max,j} - v_j^{FBA+}) \leq v_j^{FBA+} \quad (15)$$

$$v_j - y_j(v_{\max,j} - v_j^{FBA-}) \leq v_j^{FBA-} \quad (16)$$

3.7 Objective function search methods

Both MOMA and ROOM proposed an alternative to the FBA assumption that organisms are evolved for a specific objective function. Though MOMA and ROOM were developed to address this assumption in the context of a gene knockout, the question of a universal WT objective function for an organism remains. To address this, Schuetz and colleagues tested the ability of 11 unique FBA objective functions (or combinations thereof) to accurately predict fluxes in an *E. coli* model of central metabolism (Schuetz *et al.*, 2007), and reported that no single *in silico* objective function was the most accurate predictor of ^{13}C -determined *in vivo* fluxes across many different growth conditions. Several constraint-based approaches have been developed to identify the best objective function for an organism in a specified environment, including ObjFind, BOSS, and a Bayesian probability-based selection method.

3.7.1 ObjFind—ObjFind (Burgard and Maranas, 2003), uses bi-level programming to find a set of positive weights, c_p , that maximize the sum of optimal fluxes (Eqs. 18–19), while minimizing the sum-squared difference between the optimal flux profile, v_p , and the experimentally measured fluxes, v_j^* (Eq. 17, where E is the set of all experimentally measured fluxes and P is the set of all reactions that could potentially be cellular objectives). By comparing these optimal weights to those given by a hypothesized objective function, such as biomass, a researcher is able to assess the accuracy of the hypothesized objective function. However, this approach requires that all components of the true objective function be included within the S matrix, *a priori*, and will return the wrong objective if a component reaction is not included in the GENRE.

$$\min_{c_j} \sum_{j \in E} (v_j - v_j^*)^2 \quad (17)$$

$$\max_{v_j} \sum_{j \in P} c_j v_j \quad (18)$$

$$\sum_{j \in P} c_j = 1, \quad c_j \geq 0, \quad \forall j \in P \quad (19)$$

3.7.2 Biological Objective Solution Search—Gianchandani and co-workers addressed the *a priori* requirement of ObjFind by adding a generic “objective reaction,” v_{obj} , to the stoichiometric S matrix (Gianchandani *et al.*, 2008). This Biological Objective Solution Search (BOSS) requires only the S matrix and experimental isotopomer flux data. BOSS adds the generic objective reaction with unknown stoichiometry as an additional column in the S matrix (Eq. 20, where m is the total number of metabolites) and performs an FBA on this updated set of mass balance constraints and flux bounds (Eqs. 1–2), defining v_{obj} as the objective function to maximize (Eq. 22).

$$S_{i,obj} = [S_{1,obj} \ S_{2,obj} \ \dots \ S_{m,obj}] \quad (20)$$

$$\min \sum_{j \in N} (v_j^{BOSS} - v_j^*)^2 \quad (21)$$

$$\max \ v_{obj} \quad (22)$$

Similar to ObjFind, this inner optimization serves as a constraint on the outer optimization, which minimizes the sum-squared difference between the *in silico* flux profile (v^{BOSS}), and the measured *in vivo* fluxes (v^*). This approach will identify the optimal objective reaction if it already exists in the network (*i.e.*, with identical stoichiometry as an existing column in S), but also if it is a combination of existing reactions or a reaction that was omitted from the GENRE altogether. To validate this method, BOSS was used to identify the objective function in the central metabolic network of *S. cerevisiae*, and found the best two reactions to be nearly identical to the precursor biomass synthesis reaction and the ATP maintenance reaction, both commonly used FBA objective functions.

3.7.3 Bayesian discrimination—Knorr and colleagues introduced a method which uses FBA and *in vivo* flux measurements to compare any number of candidate objective functions (Knorr *et al.*, 2007). This method determines the most probable objective function by calculating posterior probabilities. Specifically it calculates the probability of each objective function (F_x), given the product of differences between predicted and measured data (Y), normalized by the sum of all posterior probabilities (Eq. 23). While this approach again requires the definition of candidate objective functions *a priori*, the validation used only a few *in vivo* measurements (growth rate, oxygen uptake rate, succinate uptake rate and acetate production rate) to compare to *in silico* predictions

$$\pi(F_x|Y) = \frac{p(F_x|Y)}{\sum_z p(F_z|Y)} \quad (23)$$

3.8 Multiple metabolic objectives

Selecting the most biologically relevant objective function is critical to accurately predicting cellular metabolism by FBA or other constraint-based methods, though more than one objective function is occasionally desired. For example, metabolic engineering often aims to identify the genetic manipulations for an organism that will optimize its synthesis rate of a desired byproduct (*e.g.*, ethanol in *E. coli*). An appropriate approach must optimize the organism's intrinsic objective function (*e.g.*, max growth) in parallel with this secondary engineered objective. Burgard and colleagues developed OptKnock, a constraint-based method, to tackle this problem of multiple objective functions (Burgard *et al.*, 2003). Similar to the bi-level optimizations of ObjFind and BOSS, OptKnock uses an FBA framework to maximize the cellular objective, v_{obj} , subject to mass balance and flux bounds constraints (Eqs. 1–2), as well as gene knockout constraints (Eqs. 25–28, where K is the maximum number of allowable knock-outs across the set of all reactions, N). In parallel, this approach varies the number and identity of knockout genes to maximize byproduct secretion, $v_{byproduct}$ (Eq. 24). More recently, numerous constraint-based metabolic engineering methods have been developed to identify optimal gene deletions with more computational efficiency than OptKnock (OptGene and GDLS), and to identify optimal gene knock-ins (OptStrain) or manipulations to increase or decrease gene expression (OptReg) (Lun *et al.*, 2009; Patil *et al.*, 2005; Pharkya *et al.*, 2004; Pharkya and Maranas, 2006).

$$\max_{y_j} v_{byproduct} \quad (24)$$

$$\max_{y_j} v_{obj} \quad (25)$$

$$v_j^{\min} \cdot y_j \leq v_j \leq v_j^{\max} \cdot y_j, \quad \forall j \in N \quad (26)$$

$$y_j \in \{0, 1\} \quad (27)$$

$$\sum_{j \in N} (1 - y_j) \leq K \quad (28)$$

3.9 Constraint-based modeling software

The bulk of the aforementioned constraint-based modeling approaches are LP problems, which can be solved using a standard solver of which there are many open source (LP_Solve, glpk) and proprietary versions (Gurobi, CPLEX, LINDO). The constraint-based reconstruction and analysis (COBRA) toolbox has become a popular means to apply many of these constraint-based methods (*e.g.*, FBA, MOMA, and FVA) to GENRE-derived models (Becker *et al.*, 2007). This toolbox is open source and has been written to operate in the MATLAB programming environment using the systems biology markup language (SBML). A recently developed open source software platform, OptFlux, combines strain

optimization methods (*e.g.*, OptKnock) as well as classic constraint-based approaches FBA, MOMA and ROOM (Rocha *et al.*, 2010).

4. Summary

In this chapter we have attempted to detail the key steps in the metabolic reconstruction process as well as introduce several constraint-based modeling techniques useful for various applications and desired outcomes. Again, it is critical to use the highest quality annotation available and dedicate significant time to refinement of the network reconstructions with well-curated databases and quality literature references in addition to the results of automated reconstruction techniques. Some common methods of collecting experimental data for both model improvement and validation are detailed, but many more are available that may be more appropriate for specific applications. The constraint-based modeling techniques presented are standards in the FBA field and accompanied by additional information helpful during difficult portions of the modeling process such as objective function definition and variability analysis. The field is a fast growing one, with an exponential rise in new reconstructions and further investigations into comparative model building techniques and reconstruction consistency, and we hope this text will be a helpful reference in these efforts.

Acknowledgments

This work was supported by funding from the National Science Foundation (CAREER Grant #0643548 to JAP), the National Institutes of Health and National Institute of General Medical Sciences (Grant #GM088244 to JAP), the Cystic Fibrosis Research Foundation (Grant #1060 to JAP), and the Robert M. Berne Cardiovascular Research Center (USPHS Grant #HL007284-33 Basic Cardiovascular Research Training Grant providing postdoctoral support to CRH).

References

- Andersen MR, Nielsen ML, Nielsen J. Metabolic model integration of the bibliome, genome, metabolome and reactome of *Aspergillus niger*. *Mol Syst Biol*. 2008; 4:178. [PubMed: 18364712]
- Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat Protoc*. 2007; 2(3):727–738. [PubMed: 17406635]
- Becker SA, Palsson BO. Context-specific metabolic networks are consistent with experiments. *Plos Computational Biology*. 2008; 4(5)
- Burgard AP, Maranas CD. Optimization-based framework for inferring and testing hypothesized metabolic objective functions. *Biotechnol Bioeng*. 2003; 82(6):670–677. [PubMed: 12673766]
- Burgard AP, Pharkya P, Maranas CD. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng*. 2003; 84(6):647–657. [PubMed: 14595777]
- Cameron DE, Urbach JM, Mekalanos JJ. A defined transposon mutant library and its use in identifying motility genes in *Vibrio cholerae*. *Proc Natl Acad Sci U S A*. 2008; 105(25):8736–8741. [PubMed: 18574146]
- DeJongh M, Formsma K, Boillot P, Gould J, Rycenga M, et al. Toward the automated generation of genome-scale metabolic networks in the SEED. *BMC Bioinformatics*. 2007; 8:139. [PubMed: 17462086]
- Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A*. 2007; 104(6):1777–1782. [PubMed: 17267599]
- Durot M, Bourguignon PY, Schachter V. Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiol Rev*. 2009; 33(1):164–190. [PubMed: 19067749]
- Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO. Reconstruction of biochemical networks in microorganisms. *Nat Rev Microbiol*. 2009; 7(2):129–143. [PubMed: 19116616]

- Feist AM, Palsson BO. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat Biotechnol*. 2008; 26(6):659–667. [PubMed: 18536691]
- Gallagher LA, Ramage E, Jacobs MA, Kaul R, Brittnacher M, et al. A comprehensive transposon mutant library of *Francisella novicida*, a bioweapon surrogate. *Proc Natl Acad Sci U S A*. 2007; 104(3):1009–1014. [PubMed: 17215359]
- Gianchandani EP, Oberhardt MA, Burgard AP, Maranas CD, Papin JA. Predicting biological system objectives de novo from internal state measurements. *BMC Bioinformatics*. 2008; 9:43. [PubMed: 18218092]
- Giglio MG, Collmer CW, Lomax J, Ireland A. Applying the Gene Ontology in microbial annotation. *Trends Microbiol*. 2009; 17(7):262–268. [PubMed: 19577473]
- Gonzalez O, Oberwinkler T, Mansueto L, Pfeiffer F, Mendoza E, et al. Characterization of growth and metabolism of the haloalkaliphile *Natronomonas pharaonis*. *PLoS Comput Biol*. 2010; 6(6):e1000799. [PubMed: 20543878]
- Hao T, Ma HW, Zhao XM, Goryanin I. Compartmentalization of the Edinburgh Human Metabolic Network. *BMC Bioinformatics*. 2010; 11:393. [PubMed: 20649990]
- Henry CS, Zinner JF, Cohoon MP, Stevens RL. iBsu1103: a new genome-scale metabolic model of *Bacillus subtilis* based on SEED annotations. *Genome Biol*. 2009; 10(6):R69. [PubMed: 19555510]
- Herrgard MJ, Swainston N, Dobson P, Dunn WB, Arga KY, et al. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat Biotechnol*. 2008; 26(10):1155–1160. [PubMed: 18846089]
- Jacobs MA, Alwood A, Thaipisuttikul I, Spencer D, Haugen E, et al. Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. *Proc Natl Acad Sci U S A*. 2003; 100(24):14339–14344. [PubMed: 14617778]
- Jankowski MD, Henry CS, Broadbelt LJ, Hatzimanikatis V. Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys J*. 2008; 95(3):1487–1499. [PubMed: 18645197]
- Jerby L, Shlomi T, Ruppin E. Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Mol Syst Biol*. 2010; 6:401. [PubMed: 20823844]
- Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, et al. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief Bioinform*. 2010; 11(1):40–79. [PubMed: 19955237]
- Kauffman KJ, Prakash P, Edwards JS. Advances in flux balance analysis. *Curr Opin Biotechnol*. 2003; 14(5):491–496. [PubMed: 14580578]
- Kircher M, Kelso J. High-throughput DNA sequencing--concepts and limitations. *Bioessays*. 2010; 32(6):524–536. [PubMed: 20486139]
- Knorr AL, Jain R, Srivastava R. Bayesian-based selection of metabolic objective functions. *Bioinformatics*. 2007; 23(3):351–357. [PubMed: 17150997]
- Kuzniar A, van Ham RC, Pongor S, Leunissen JA. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet*. 2008; 24(11):539–551. [PubMed: 18819722]
- Lee JM, Gianchandani EP, Papin JA. Flux balance analysis in the era of metabolomics. *Brief Bioinform*. 2006; 7(2):140–150. [PubMed: 16772264]
- Lun DS, Rockwell G, Guido NJ, Baym M, Kelner JA, et al. Large-scale identification of genetic design strategies using local search. *Mol Syst Biol*. 2009; 5:296. [PubMed: 19690565]
- Mahadevan R, Edwards JS, Doyle FJ 3rd. Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. *Biophys J*. 2002; 83(3):1331–1340. [PubMed: 12202358]
- Mahadevan R, Schilling CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng*. 2003; 5(4):264–276. [PubMed: 14642354]
- Niittylä T, Chaudhuri B, Sauer U, Frommer WB. Comparison of quantitative metabolite imaging tools and carbon-13 techniques for fluxomics. *Methods Mol Biol*. 2009; 553:355–372. [PubMed: 19588116]
- Oberhardt MA, Chavali AK, Papin JA. Flux balance analysis: interrogating genome-scale metabolic networks. *Methods Mol Biol*. 2009; 500:61–80. [PubMed: 19399432]

- Oberhardt MA, Goldberg JB, Hogardt M, Papin JA. Metabolic network analysis of *Pseudomonas aeruginosa* during chronic cystic fibrosis lung infection. *J Bacteriol.* 2010; 192(20):5534–5548. [PubMed: 20709898]
- Oberhardt MA, Puchalka J, Fryer KE, Martins dos Santos VA, Papin JA. Genome-scale metabolic network analysis of the opportunistic pathogen *Pseudomonas aeruginosa* PAO1. *J Bacteriol.* 2008; 190(8):2790–2803. [PubMed: 18192387]
- Orth JD, Thiele I, Palsson BO. What is flux balance analysis? *Nat Biotechnol.* 2010; 28(3):245–248. [PubMed: 20212490]
- Papin JA, Price ND, Wiback SJ, Fell DA, Palsson BO. Metabolic pathways in the post-genome era. *Trends Biochem Sci.* 2003; 28(5):250–258. [PubMed: 12765837]
- Patil KR, Rocha I, Forster J, Nielsen J. Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics.* 2005; 6:308. [PubMed: 16375763]
- Pharkya P, Burgard AP, Maranas CD. OptStrain: a computational framework for redesign of microbial production systems. *Genome Res.* 2004; 14(11):2367–2376. [PubMed: 15520298]
- Pharkya P, Maranas CD. An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Metab Eng.* 2006; 8(1):1–13. [PubMed: 16199194]
- Price ND, Papin JA, Palsson BO. Determination of redundancy and systems properties of the metabolic network of *Helicobacter pylori* using genome-scale extreme pathway analysis. *Genome Res.* 2002; 12(5):760–769. [PubMed: 11997342]
- Price ND, Reed JL, Palsson BO. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol.* 2004; 2(11):886–897. [PubMed: 15494745]
- Raman K, Chandra N. Flux balance analysis of biological systems: applications and challenges. *Brief Bioinform.* 2009; 10(4):435–449. [PubMed: 19287049]
- Rocha I, Forster J, Nielsen J. Design and application of genome-scale reconstructed metabolic models. *Methods Mol Biol.* 2008; 416:409–431. [PubMed: 18392985]
- Rocha I, Maia P, Evangelista P, Vilaca P, Soares S, et al. OptFlux: an open-source software platform for in silico metabolic engineering. *BMC Syst Biol.* 2010; 4:45. [PubMed: 20403172]
- Savinell JM, Palsson BO. Optimal selection of metabolic fluxes for in vivo measurement. I. Development of mathematical methods. *J Theor Biol.* 1992; 155(2):201–214. [PubMed: 1453697]
- Schuetz R, Kuepfer L, Sauer U. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol Syst Biol.* 2007; 3:119. [PubMed: 17625511]
- Segre D, Vitkup D, Church GM. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A.* 2002; 99(23):15112–15117. [PubMed: 12415116]
- Senger RS. Biofuel production improvement with genome-scale models: The role of cell composition. *Biotechnol J.* 2010; 5(7):671–685. [PubMed: 20540108]
- Shlomi T, Berkman O, Ruppin E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci U S A.* 2005; 102(21):7695–7700. [PubMed: 15897462]
- Shlomi T, Cabili MN, Herrgard MJ, Palsson BO, Ruppin E. Network-based prediction of human tissue-specific metabolism. *Nature Biotechnology.* 2008; 26(9):1003–1010.
- Smallbone K, Simeonidis E. Flux balance analysis: a geometric perspective. *J Theor Biol.* 2009; 258(2):311–315. [PubMed: 19490860]
- Suthers PF, Dasika MS, Kumar VS, Denisov G, Glass JI, et al. A genome-scale metabolic reconstruction of *Mycoplasma genitalium*, iPS189. *PLoS Comput Biol.* 2009; 5(2):e1000285. [PubMed: 19214212]
- Thiele I, Palsson BO. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc.* 2010; 5(1):93–121. [PubMed: 20057383]
- Van Noorden CJ. Imaging enzymes at work: metabolic mapping by enzyme histochemistry. *J Histochem Cytochem.* 2010; 58(6):481–497. [PubMed: 20124092]
- Varma A, Palsson BO. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl Environ Microbiol.* 1994; 60(10):3724–3731. [PubMed: 7986045]

Yang Y, Gilbert D, Kim S. Annotation confidence score for genome annotation: a genome comparison approach. *Bioinformatics*. 2010; 26(1):22–29. [PubMed: 19855104]

Figure 1. Flow diagram illustrates the iterative multi-step methodology for reconstructing and simulating the behavior of an organism-specific genome-scale metabolic network model.

Figure 2. Conversion of GENRE reaction stoichiometries to S matrix. A - Schematic of a toy cell/organism with defined boundaries (E/I), metabolites (M), and reactions (X , T , R , D). B - Five metabolites and seven reactions comprise the S matrix converted from the cell/organism of panel A.

Table 1

Valuable resources for GENRE curation and analysis efforts.

Genome Annotation Resources		
ERGO	automated and manual annotation with pathway mapping and bioinformatics tools	http://ergo.integratedgenomics.com
Comprehensive Microbial Resource	database of all prokaryotic genomes sequenced to date with automated and/or manual annotation	http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi
Integrated Microbial Genomes	automated genome annotation and comparative genomics analysis	http://img.jgi.doe.gov
ENSEMBL	sequence access portal with annotation, visualization, and analysis tools	http://www.ensemblgenomes.org/
NCBI Entrez Genome	currently over 12,000 sequenced and/or annotated records covering all major organism groups	http://www.ncbi.nlm.nih.gov/sites/genome
KEGG	databases on sequenced genomes, orthology, hierarchy, reactions, proteins, and pathway maps	http://www.genome.jp/kegg/
UCSC Genome Bioinformatics	browser and database useful for comparing genomes and annotations	http://genome.ucsc.edu/
GO	database of standardized annotation terms	http://www.geneontology.org/
Gene Sequence Alignment Tools		
BLAST	local alignment search of protein/nucleotide sequences - does not allow gaps when matching sequences	http://blast.ncbi.nlm.nih.gov/
FASTA	local alignment search of protein/nucleotide sequences - allows gaps when matching sequences	http://www.ebi.ac.uk/Tools/fasta/
BLAT	local alignment search matching sequence greater than 25 bp long	http://genome.ucsc.edu/cgi-bin/hgBlat?org=human
Enzyme/Protein Databases		
UniProt/Swiss-Prot	manual annotation of protein sequence function	http://www.uniprot.org
UniProt/TrEMBL	automatic annotation of protein sequence function	http://www.uniprot.org
NCBI-RefSeq	curated DNA, RNA, and protein sequence annotations	http://www.ncbi.nlm.nih.gov/RefSeq/
BRENDA	enzymes with associated reactions, pathways, substrates, and cofactors	http://www.brenda-enzymes.org/
Automated Reconstruction		
SEED	automated annotation and draft reconstruction of user-input genomes	http://www.theSEED.org/
PathwayTools	draft reconstruction with visualization tools and pathway gap prediction	http://bioinformatics.ai.sri.com/ptools/
metaSHARK	draft reconstruction with ability to overlay information on generic metabolic networks	http://bmbpcu36.leeds.ac.uk/shark/
KOBAS	automated annotation and reconstruction using KEGG orthology	http://kobas.cbi.pku.edu.cn/
ASGARD	automated annotation and reconstruction	http://sourceforge.net/projects/asgard-bio/
Analysis		

CellNetAnalyzer	Matlab-based environment for topological network analysis of metabolism and cell signaling	http://www.mpi-magdeburg.mpg.de/projects/cna/cna.html
Metatool	analysis suite that finds a set of elementary modes (nondecomposable constrained steady state networks) for a reconstruction	http://penguin.biologie.uni-jena.de/bioinformatik/networks/