Chapter 9
Genome-Scale Reconstruction, Modeling, and Simulation of *E. coli's* Metabolic Network

Adam M. Feist, Ines Thiele, and Bernhard Ø. Palsson

Contents

Abstract Since the release of the first genome-scale metabolic reconstruction of the *E. coli* metabolic network in 2000, there has been a growing number of researchers around the world adapting it for a broad range of studies (Feist and Palsson 2008). The uses range from practical applications to obtaining basic biological understanding of cellular behavior. This range of uses is further expected to expand as the reconstruction broadens in scope and as new *in silico* methods are developed, implemented, and put to use.

In this chapter, we will describe foundational concepts central to the reconstruction process and model formulation, the history of reconstruction of the *E. coli* metabolic network, the development of reconstruction technology, genome-scale constraint based modeling with key exemplary case studies of uses of the *E. coli* metabolic reconstruction, and insights into the future of the field. As such,

---

B.Ø. Palsson (✉)
Department of Bioengineering, University of California San Diego, 9500 Gilman Drive,
La Jolla, CA 92093-0412, USA
e-mail: palsson@ucsd.edu

S.Y. Lee (ed.), *Systems Biology and Biotechnology of Escherichia coli*,
DOI 10.1007/978-1-4020-9394-4_9,

this chapter should serve as a guide to those interested in either expanding the application of the *E. coli* reconstruction or adapting established applications to other organisms.

9.1 Foundational Concepts

The reconstruction of the *E. coli* metabolic network has led to the development of *'bottom-up'* reconstruction technology, genome-scale modeling methods, and basic and practical uses. A number of foundational concepts have also been developed during the period that we introduce here and provide background and a conceptual framework for the reader (see Palsson 2006, Price et al. 2004a).

Forming a BiGG knowledge base: A network reconstruction is based on a highly curated set of primary biological information for a particular organism; a biochemically, genetically and genomically structured (BiGG) knowledge base (Reed et al. 2006a). Such a knowledge base represents a large body of experimental data that is meticulously assembled and curated through the systems biology and reconstruction approaches detailed herein.

Genome-scale network reconstruction (GENRE): An organism-specific BiGG knowledge base is the basis for a GENRE. A GENRE is specific to a particular organism, for example, GENRE of *Escherichia coli* (below we will see four of these, specifically called *i*JE660, *i*JR904, *i*MBEL979, and *i*AF1260). A GENRE contains a list of all the known (and some predicted) chemical transformations that are believed to take place in the particular network (e.g. metabolic, transcriptional regulatory network, etc.).

The central role of network reconstruction in systems biology: Systems biology research generally can be conceptualized as a four-step process (Fig. 9.1). Foundational to the field is the generation of global, or genome-scale, data. The growing number of available 'omics' data types has created the need for formal and structured multi-'omic' data integration (Joyce and Palsson 2006). Omics data, along with legacy information (i.e., the 'bibliome') and detailed small-scale experiments, can be used to define the interactions among biological components that are used to reconstruct networks in particular organisms (Reed et al. 2006a). Network reconstruction is also an iterative, on-going process that continually integrates data in a formal fashion as it becomes available (Reed and Palsson 2003). These characteristics render the network reconstruction as a common denominator for those studying systems biology. The reconstruction effectively represents a 2-D annotation of a genome detailing not only the parts for an organism, but the interactions between specific components (Palsson 2004). Genome-scale reconstruction technologies for metabolic (Reed et al. 2006a), transcriptional regulation (Covert et al. 2004, Gianchandani et al. 2006, Herrgard et al. 2004) and signaling networks (Papin et al. 2005) have been established, and transcriptional/translational network reconstruction methods are currently under development (Thiele et al. 2009). An in depth review on the bottom-up reconstruction process (Palsson 2006) as well as a current review of biological network reconstruction (Feist et al. 2009) have been generated.

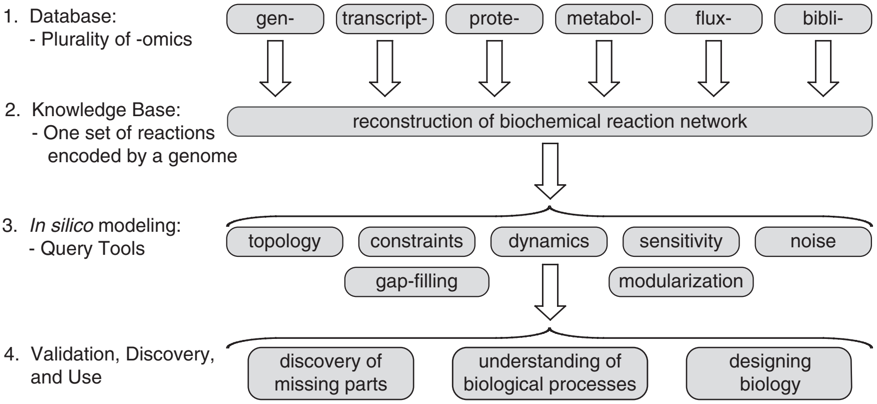

Fig. 9.1 Systems Biology as a 4-step Process. Step 1, the process is based on a variety of high-throughput data sets (i.e., 'omics' data) and a comprehensive assessment of the literature (i.e., bibliomic data). Step 2, all of the data types are used to reconstruct the list of biochemical transformations that make up a network as well as their genetic basis (Reed et al. 2006a). In principal, the network is unique. Step 3, the data contained in the reconstruction can be formally represented (i.e., in the form of matrices and logical statements) that can be mathematically characterized by a variety of methods. Step 4, the computational model enables a broad spectrum of applications, as reviewed in this chapter. Figure adapted from (Feist and Palsson 2008, Palsson 2006)

Constraint-based reconstruction and analysis (COBRA): COBRA is the overall philosophy and approach of applying constraints to limit the range of achievable functional (phenotypic) states of GENREs (outlined below). A GENRE operates under defined constraints. These constraints fall into at least four categories (Palsson 2006): physico-chemical, topological, regulatory, and environmental. Such constraints can be mathematically represented and imposed on the functional states that a GENRE can take on. Functional states can be assessed using a variety of computational methods (Palsson 2006, Price et al. 2004a) and have been disseminated in the form of a COBRA Toolbox (Becker et al. 2007) that is a MATLAB (The MathWorks Inc., Natick, MA) based software package.

Converting network reconstructions into a Genome-scale Model (GEM): A GENRE can be converted into a mathematical form (i.e., an *in silico* model) and used to computationally assess phenotypic properties (reviewed in (Price et al. 2004a)). The COBRA approach is used to analyze the properties of GENREs by assessing allowable functional states. Genome-scale reconstructions are thus a key step in quantifying the genotype-phenotype relationship and can be used to 'bring genomes to life' (Frazier et al. 2003). The availability of reconstructed metabolic networks for microorganisms has increased rapidly in recent years and a growing number of research groups are synthesizing GENREs for target organisms of interest (see Fig. 9.4) (Feist et al. 2009, Reed et al. 2006a).

The conversion of a reconstruction (GENRE) to an *in silico* model (GEM), represented by the arrow from step 2 to step 3 in Fig. 9.1, involves a subtle, but critical, transition. The chemical transformations of which a GENRE is comprised can be represented stoichiometrically (as well as other formats, e.g., a directed graph). Stoichiometric representations form a matrix, the rows of which represent the compounds, the columns of which represent the chemical transformations, and the entries of which are the stoichiometric coefficients (see section below and Fig. 9.6) With the definition of systems boundaries and other details, a network reconstruction can be converted into a mathematical format that can be computationally interrogated. The process that this arrow represents is the bridge between the realms of high-throughput data/bioinformatics and systems science.

9.2 History of the *E. coli* Metabolic Network Reconstruction: An Ongoing and Iterative Process

The 18-year history of metabolic reconstruction for *E. coli* is outlined in Fig. 9.2 (Feist and Palsson 2008, Reed and Palsson 2003). *E. coli* served as a model organism in the era of discovery of metabolic biochemistry, and thus, comprehensive metabolic reconstructions were developed before its genome sequence was available (Varma et al. 1993a,b).With the publication of the *E. coli* genomic sequence in 1997 (Blattner et al. 1997), the development and use of the metabolic reconstruction in *E. coli* grew rapidly in scope.

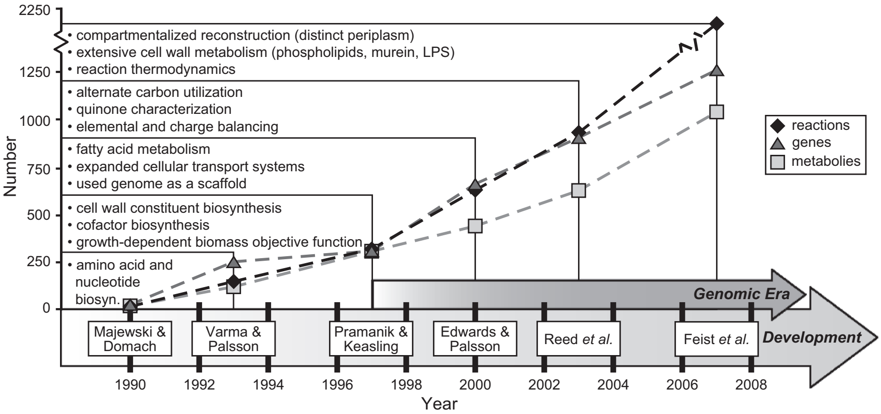

Fig. 9.2 The ongoing reconstruction of the E. coli metabolic network. History of the *E. coli* metabolic reconstruction. Shown are six milestone efforts contributing to the reconstruction of the *E. coli* metabolic network. For each of the six reconstructions (Edwards and Palsson 2000, Feist et al. 2007, Majewski and Domach 1990, Pramanik and Keasling 1997, 1998, Reed et al. 2003, Varma et al. 1993a,b) (see text for details), the number of included reactions (*diamonds*), genes (*triangles*), and metabolites (*squares*) are displayed. Also listed is the expansion in scope in each successive reconstruction. The start of the genome era in 1997 (Blattner et al. 1997) marked a significant increase in scope. The reaction, gene, and metabolite values for pre-genomic era reconstructions were estimated from the content outlined in each publication and in some cases, encoding genes for reactions were unclear. Fig. adapted from (Feist and Palsson 2008)

Pre-genome era: Beginning in 1990, a network reconstruction consisting of 14 reactions (characterizing primarily the TCA cycle and partially glycolysis) was generated to analyze the production and secretion of acetate during aerobic growth on glucose (Majewski and Domach 1990). This example demonstrates the scope of initial uses of network reconstructions of *E. coli*. Later, in 1993, a larger metabolic reconstruction consisting of 146 reactions was generated, representing key catabolic and anabolic metabolic pathways (Varma et al. 1993a,b). This reconstruction was used for computing (Varma et al. 1993a, Varma and Palsson 1993, 1994, 1995):Optimal production of cofactors and biosynthetic precursors, Maximum allowable generation of amino acids and nucleic acids, and Internal network flux distributions for optimal and sub-optimal growth.

The computational predictions based on the model were compared to experimental data and found to be consistent with measurements under both aerobic and anaerobic glucose minimal media conditions (Varma and Palsson 1994). The comparison of computation and experimental findings in this work demonstrated the important concept of comparison to *in vivo* data as computational outcomes have to be considered as hypotheses that need experimental confirmation.

Following these developments in the early 1990s, an expanded reconstruction consisting of 317 reactions was generated in 1997. It included cofactor and cell wall biosynthesis, and other additional metabolic pathways (Pramanik and Keasling 1997, 1998). This expanded reconstruction was used for computations that incorporated measured metabolite uptake and secretion rates to predict central metabolic fluxes which were found to be consistent with enzymatic flux values determined from isotopomer-based measurements (Pramanik and Keasling 1997, 1998). These studies also incorporated a growth rate dependent biomass objective function that had not been considered in previous studies. It should be noted that isotopomer-based measurements are also network dependent and studies are currently emerging looking specifically at this issue (Suthers et al. 2007).

Note that these pre-genome era reconstructions of *E. coli* metabolism were based solely on biochemical information and provided an important foundation for subsequent work at the genomic scale.

Genome era: The complete genome sequence for *E. coli* K-12 MG1655 was published in 1997 (Blattner et al. 1997). Its availability fueled a significant increase in network reconstruction content and scope as the genome sequence directly provided a list of parts (components) present in *E. coli* (Fig. 9.2). Utilizing the annotated sequence, a genome-scale metabolic reconstruction was generated for *E. coli* consisting of 627 unique reactions catalyzed by 660 gene products (Edwards and Palsson 2000). This reconstruction, later titled *i*JE660, was initially used to:Predict the phenotypes for knock-out mutants of the central metabolic pathways (Edwards and Palsson 2000), Design quantitative experiments (Edwards et al. 2001), and Predict the outcome of adaptive evolution in the context of the metabolic machinery available to the cell (Ibarra et al. 2002). These results demonstrated the utility of the reconstruction to understand growth characteristics of *E. coli*, the effects of gene deletions, and to point to areas of computational and experimental disagreement that identify targets for further biochemical characterization (see below).

An updated annotation of the *E. coli* K-12 MG1655 genome (Serres et al. 2001) and continual functional characterization of *E. coli* metabolic content enabled an expansion of the reconstruction in 2003, which consisted of 931 reactions catalyzed by 904 gene products (Reed et al. 2003). This reconstruction, titled *i*JR904, was an improvement over previous efforts in that contained both charge and elemental balancing of all reactions, expanded the various carbon source utilization pathways, contained a larger number of characterized transport systems and their encoding genes, better accounted for quinone usage in the electron transport chain, and better detailed the relationship between given genes, proteins, and reactions contained in the reconstruction (the GPR associations).

This reconstruction has been utilized for a broad number of applications reviewed later in this chapter. Utilizing the *i*JR904 (Reed et al. 2003) reconstruction, an expanded reconstruction of *E. coli* was generated (containing 979 reactions and titled *i*MBEL979) for the purpose of designing overproducing strains in the software framework MetaFluxNet (Lee et al. 2005).

The most recent metabolic reconstruction for *E. coli*, titled *i*AF1260, incorporates data from the most recent *E. coli* K-12 MG1655 genome annotation (Riley et al. 2006) and consists of 2,077 reactions and 1,260 genes (Feist et al. 2007). The advancements represented by *i*AF1260 over *i*JR904 lie in five main areas: an increased scope with the inclusion of 357 additional ORFs; compartmentalization into three distinct compartments (cytoplasmic, periplasmic and extra-cellular); the detailing of all grouped, or lumped, reactions (most often associated with lipid and lipopolysaccharide biosynthesis); the incorporation of reaction thermodynamics, calculated Gibbs free energy (ΔG°) values for 950 metabolites and 1935 reactions; and alignment with the EcoCyc database (Keseler et al. 2005) which provided expanded coverage for the network and content mappings for further computational analyses.

This 18-year history of reconstruction of the *E. coli* metabolic network has culminated in a network containing a total number of 1,260 metabolic genes covering 28% of the 4,453 identified ORFs on the *E. coli* genome. More importantly, the 1260 ORFs represent 48% of the functionally annotated ORFs that have been confirmed by experimental data (Table 9.1). Thus, 92% of the 1,260 gene products included in *i*AF1260 have been experimentally verified (Riley et al. 2006) with the balance of 8% having a computationally predicted function which necessitate confirmation with focused experimentation. Model-aided gap-filling and discovery will aid in this process (see Section 9.5.2). In addition, protein structures (computed or experimental) are available for a large fraction of the proteins in *i*AF1260 (Berman et al. 2000). Integration of protein structural data with the functional content of the reconstruction will lead to a better understanding of structural motifs and their properties.

Reconstruction of the *E. coli* metabolic network is thus approaching exhaustion of known metabolic gene functions and is now being used in a prospective fashion to discover new metabolic capabilities in *E. coli* (see below). As a result of this endeavour, the reconstruction of the *E. coli* metabolic network represents the best-developed genome-scale network to date.

Table 9.1 Properties of the most current *E. coli* metabolic reconstruction

	*i*AF1260 this study
Included genes	1260 (28%)[d]
Experimentally–based function	1161 (92%)
Computationally predicted function	99 (8%)
Unique functional proteins	1148
Multigene complexes	167
Genes involved in complexes	415
Instances of isozymes[a]	346
Reactions	2077
Metabolic Reactions	1387
Unique metabolic reactions[b]	1339
Cytoplasmic	1187
Periplasmic	192
Extracellular	8
Transport Reactions	690
Cytoplasm to periplasm	390
Periplasm to extracellular	298
Cytoplasm to extracellular	2
Gene - protein - reaction associations	
Gene associated (met./trans.)	1294/625
Spontaneous/diffusion reactions[c]	16/9
Total gene associated and no association needed (met./trans.)	1310/634 (94%)
No gene association (metabolic/transport)	77/56 (6%)
Exchange reactions	304
Metabolites	
Unique Metabolites[b]	1039
Cytoplasmic	951
Periplasm	418
Extracellular	299

[a] tabulated on a reaction basis, not counting outer membrane non-specific porin transport.
[b] reactions can occur in or between multiple compartments and metabolites can be present in more than one compartment.
[c] diffusion reactions do not include facilitated diffusion reactions and are not included in this total if they can also be catalyzed by a gene product at a higher rate.
[d] overall genome coverage based on 4453 total ORFs in *E. coli*; *i*AF1260 contains 48% of the ORFs in *E. coli* that have been characterized experimentally (2403 ORFs).

9.3 Continuing Development of Reconstruction Technology

Development of the reconstruction process for metabolic networks: As illustrated in the previous section, the reconstruction process for metabolic networks is an iterative procedure that requires different types of experimental data and techniques at each phase of reconstruction. The experience with *E. coli* has led to

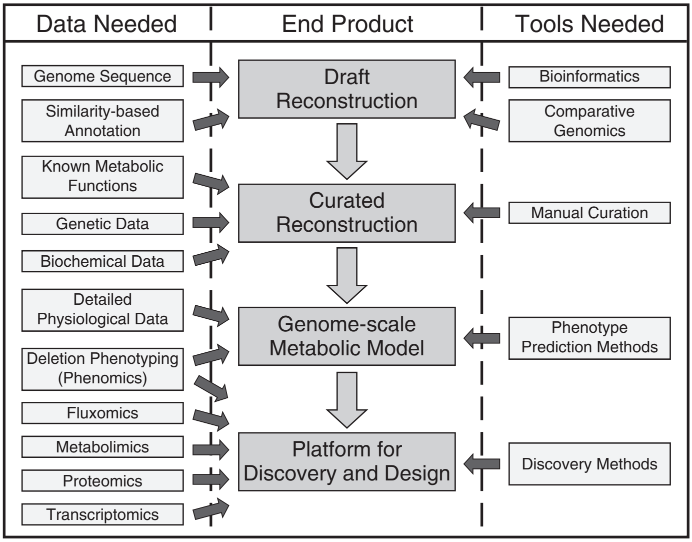

Fig. 9.3 The phases and tools necessary to generate a metabolic reconstruction. The genome-scale metabolic reconstruction process can be broken down into four major phases (center column), with each of the latter phases building off the previous. This process is iterative and driven by experimental data (primarily in the three latter phases). For each phase, specific data types are necessary and these range from high-throughput data types (e.g., phenomics, metabolomics, etc.), to detailed studies characterizing individual components (e.g., biochemical data for a particular reaction). For example, the genome annotation can provide a parts list of a cell, whereas genetic data can provide information about the contribution of each gene product towards a phenotype (e.g., when removed or mutated). The product generated from each reconstruction phase can be utilized and applied to examine a growing number of questions with the final product having the broadest applications

the formulation of the workflows that underlie metabolic reconstruction. The four phases of the reconstruction process are depicted in Fig. 9.3 and the product at each phase can be used for different applications, with the number of applications increasing with network development. This procedure represents the current status of network reconstruction, and the most recent *E. coli* reconstruction, *i*AF1260, was built accordingly (Feist et al. 2007) with the advantage of starting from an already well-established reconstruction, *i*JR904. The end product of this reconstruction effort is a platform for design and discovery, and key examples of use are given later in this chapter. More extensive descriptions exist, which outline the conceptual basis (Reed et al. 2006a) and the detailed process to generate genome-scale biological networks, (Feist et al. 2009) and these will not be repeated here.

Development of the reconstruction process beyond metabolism: The development and use of genome-scale reconstruction was rapid and many computational models were developed to address a growing spectrum of basic research and applied problems. Still, further development of reconstruction technology is necessary. The scope of reconstructions is bound to grow, representing more and more BiGG knowledge in the structured format of GEMs (Breitling et al. 2008). Growth in scope is likely to proceed in phases (Feist and Palsson 2008). Growth in scope in the near-term will involve the transcriptional and translational machinery (Allen and Palsson 2003, Mehra and Hatzimanikatis 2006, Thiele et al. 2009, Thomas et al. 2007). Such an extension will enable a range of studies including the direct inclusion of proteomic data, fine graining of growth requirements, and the explicit consideration of secreted protein products.

Another expansion in scope in the near-term is the reconstruction of the genome-scale transcriptional regulatory network (TRN). Such reconstruction at the genome-scale is now enabled by new experimental technologies, such as ChIP-chip (Lee et al. 2002). Experimental interrogation of the currently available TRN suggests that we know about one-fourth to one-third of its content (Covert et al. 2004), indicating that there is much to be discovered. This expectation is being confirmed with high-resolution ChIP-Chip data for *E. coli* (Cho et al. 2008). Once reconstructed, the TRN will allow computational predictions of the context-specific uses of the *E. coli* genome and the responses of two-component signaling systems.

Mid-term expansions in scope are likely to include the growth cycle, shock responses (e.g. heat and acid shock), and additional cellular functions (e.g. DNA replication and flagellar biosynthesis). Such a reconstruction should eventually be a comprehensive representation of the chemical reactions and transformations enabled by *E. coli*'s gene products.

Longer-term reconstruction may begin to address the 3-dimensional organization of the bacterial cell. In particular, high-resolution ChIP-chip data on the DNA binding protein could enable the estimation of the topological arrangement of the genome, and potentially elucidate the structure of the cell wall and other cellular structures that will allow a full 3-dimensional reconstruction of *E. coli*.

The two near-term expansions in content will encompass the activity of approximately 2000 ORFs in the *E. coli* genome. Clearly, quality-controlled reconstructions will help in guiding us to comprehensive genome-scale representation of all major cellular processes in bacteria at the BiGG data level of resolution that, in turn, enables GEMs of growing coverage and resolution. The scope of this effort has been described as being; "... 10 times more ambitious and 100 times more important for mankind [compared with Human Genome Project]..." Hans Westerhoff (Holden 2002).

Influence of the *E. coli* reconstruction on the *in silico* analysis of other micro-organisms: The metabolic network reconstruction of *E. coli* has been influential in the generation of other organism-specific metabolic networks. The *E. coli* metabolic reconstruction has served: As a content database where stoichiometrically and charge balanced reactions, and even pathways, have been incorporated into new reconstructions, As a database for defined metabolites, and as a source for a biomass objective function to query network content and functionality.

This influence has sparked an increase in the number of genome-scale network reconstructions that have been generated to formulate GEMs for a number of organisms. A detailed list of GEMs that have been developed, curated, and used for computation is given in Table 9.2. This table is a current snapshot of the

Table 9.2 Available predictive genome-scale metabolic network reconstructions

		Organism properties	Reconstruction properties				
Name	Strain	Genes	Genes	Metabolites	Reactions	Compartments	References
BACTERIA							
Bacillus subtilis		4, 225	844	988	1020	2 (c,e)	(Oh et al. 2007)
Clostridium acetobutylicum	ATCC 824	3, 848	474	422	552	2 (c,e)	(Senger and Papoutsakis 2008)
Clostridium acetobutylicum	ATCC 824	3, 848	432	479	502	2 (c,e)	(Lee et al. 2008)
Escherichia coli	K12 MG1655	4, 405	660	438	627	2 (c,e)	(Edwards and Palsson 2000)
Escherichia coli	K12 MG1655	4, 405	904	625	931	2 (c,e)	(Reed et al. 2003)
Escherichia coli	K12 MG1655	4, 405	1260	1039	2077	3 (c,e,p)	(Feist et al. 2007)
Geobacter sulfurreducens		3, 530	588	541	523	2 (c,e)	(Mahadevan et al. 2006)
Haemophilus influenzae	Rd	1, 775	296	343	488	2 (c,e)	(Edwards and Palsson 1999)
Haemophilus influenzae	Rd	1, 775	400	451	461	2 (c,e)	(Schilling and Palsson 2000)
Helicobacter pylori	26695	1, 632	341	485	476	2 (c,e)	(Thiele et al. 2005b)
Helicobacter pylori	26695	1, 632	291	340	388	2 (c,e)	(Schilling et al. 2002)
Lactobacillus plantarum	WCFS1	3, 009	721	531	643	2 (c,e)	(Teusink et al. 2006)
Lactococcus lactis	ssp. lactis IL1403	2, 310	358	422	621	2 (c,e)	(Oliveira et al. 2005)
Mannheimia succiniciproducens	MBEL55E	2, 384	425	519	686	2 (c,e)	(Kim et al. 2007)
Mycobacterium tuberculosis	H37Rv	4, 402	726	739	849	2 (c,e)	(Beste et al. 2007)
Mycobacterium tuberculosis	H37Rv	4, 402	661	828	939	2 (c,e)	(Jamshidi and Palsson 2007)
Mycoplasma genitalium	G-37	521	189	276	264	2 (c,e)	Personal Comm.: Patrick F. Suthers
Neisseria meningitidis	serogroup B	2, 226	555	471	496	2 (c,e)	(Baart et al. 2007)
Pseudomonas aeruginosa	PA01	5, 640	1056		883	2 (c,e)	(Oberhardt et al. 2008)
Pseudomonas putida	KT2440	5, 350	746	911	950	3 (c,e,p)	(Nogales 2008)
Rhizobium etli	CFN42	3, 168	363	371	387	2 (c,e)	(Resendis-Antonio et al. 2007)

Table 9.2 (continued)

		Organism properties	Reconstruction properties				
Name	Strain	Genes	Genes	Metabolites	Reactions	Compartments	References
Staphylococcus aureus	N315	2, 588	619	571	641	2 (c,e)	(Becker and Palsson 2005)
Staphylococcus aureus	N315	2, 588	551	604	712	2 (c,e)	(Heinemann et al. 2005)
Streptomyces coelicolor	A3(2)	8, 042	700	500	700	2 (c,e)	(Borodina et al. 2005)
ARCHAEA							
Methanosarcina barkeri	Fusaro	5, 072	692	558	619	2 (c,e)	(Feist et al. 2006)
Halobacterium salinarum	R-1	2, 867	490	557	711	2 (c,e)	(Gonzalez et al. 2008)
EUKARYOTES							
Aspergillus nidulans		9, 451	666	732	794	4	(David et al. 2008)
Homo sapiens		28, 783	1, 496	2, 766	3, 311	8	(Duarte et al. 2007)
Leishmania major	Friedlin	8, 370	560	1, 101	1, 112	8	(Chavali et al. 2008)
Mus musculus		28, 287	473	872	1, 220	3 (c,e,m)	(Sheikh et al. 2005)
Saccharomyces cerevisiae	Sc288	6, 183	708	584	1, 175	3 (c,e,m)	(Forster et al. 2003)
Saccharomyces cerevisiae	Sc288	6, 183	750	646	1, 149	8	(Duarte et al. 2004)
Saccharomyces cerevisiae	Sc288	6, 183	672	636	1, 038	3 (c,e,m)	(Kuepfer et al. 2005)

This list includes genome-scale metabolic network reconstructions that have been converted into predictive genome-scale models and whose predictive power has been validated against experimental data. Compartments: c – cytosol, e – extraorganism, p – periplasm, m – mitochondrion.

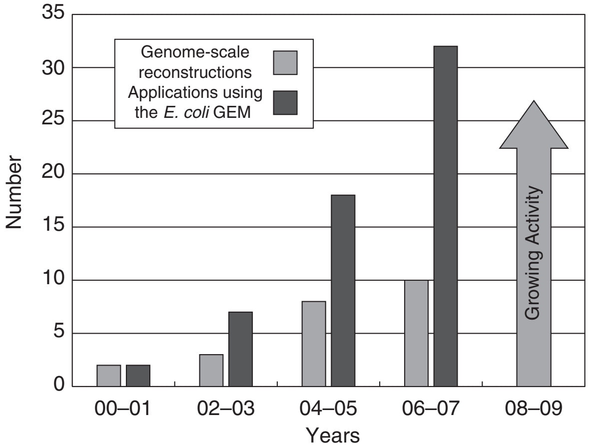

Fig. 9.4 Appearance of organism-specific genome-scale reconstructions and applications of the E. coli metabolism reconstruction. The genome-scale reconstructions for metabolic networks that have appeared every two years since the release of the first GEMs in 2000 (see Table 9.2) and the number of published studies the have appeared utilizing the *E. coli* GEM (Feist and Palsson 2008). Since the release of the first GEMs for *E. coli* (Edwards and Palsson 2000) and that of *Haemophilus influenzae* (Edwards and Palsson 1999), there has been a significant increase in both the number of genome-scale reconstructions and studies focused on the *E. coli* GEM for every time period

available reconstructions and a continually updated version can be found online (http://systemsbiology.ucsd.edu/*In_Silico*_Organisms/Other_Organisms). Additionally, Fig. 9.4 shows the number of genome-scale reconstructions that have been developed over two year periods (for the reconstructions listed in Table 9.2). The number of reconstructions generated for each period has increased since the release of the first genome-scale reconstructions for *Haemophilus influenzae* in 1999 (Edwards and Palsson 1999) and *E. coli* in 2000 (Edwards and Palsson 2000). Furthermore, the number of published studies utilizing the *E. coli* GEM has also increased significantly over time resulting in the applications outlined in the sections below (Feist and Palsson 2008).

Modeling strategy and philosophy: Models are a formal way of accounting for our knowledge about the phenomena being described. When describing biochemical reaction networks formally, we need to deal with the 'links' (i.e., the reactions) between 'nodes' (i.e., the compounds). Our knowledge about links between biological molecules varies; from the abstract to the specific (Fig. 9.5). Statistical models are built on correlations and a black box approach that is not mechanism based. Specific mechanism-based models are based on knowledge of chemistry, kinetics, and thermodynamics. Given the fact that kinetic and thermodynamic information is hard to obtain on a large-scale, stoichiometric models stop one step short of full specification (in the spectrum conveyed in Fig. 9.5). The result is that we have chemistry (and its genetic basis) and network structure used as the foundation for building a mathematical description of network functions. Such models do not have a unique solution

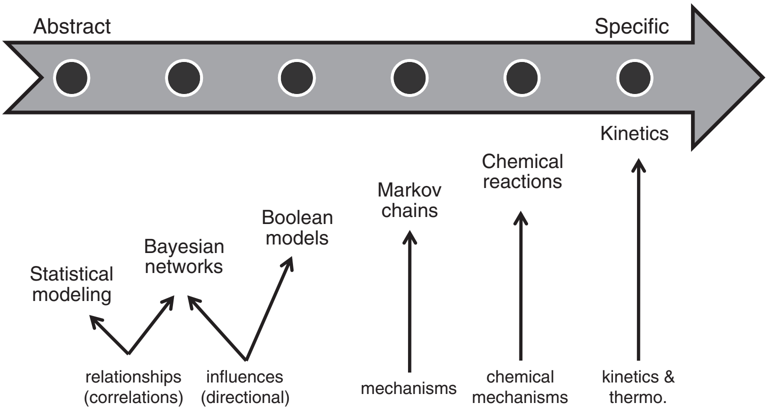

Fig. 9.5 The different levels of knowledge used to generate biological models. Our knowledge about links in biochemical networks varies. At one extreme, the information is abstract and often takes the form of *black-box* correlations. At the other, we have detailed chemical mechanisms with kinetic and thermodynamic information. Stoichiometric models would be second from the *right*, accounting for mechanisms, but not incorporating kinetic and thermodynamic information

(e.g., see (Palsson 2006) and below). The lack of kinetic information can be dealt with by: (1) examining the properties of the entire set of solutions (i.e., the solutions space) or (2) by using constraint-based optimization to find specific solutions in the space (Price et al. 2004a). The latter can be successful if we know the prevailing selection pressure on an organism. The combination of a network reconstruction that is based on a knowledge-base at the genome-scale and the inherent optimality properties of the selection process underlie the success of COBRA for a number of applications.

Constraint-based modeling methods: Over the past quarter century, there has been a growing number of computational tools developed to interrogate biological networks and models (Breitling et al. 2008, Palsson 2006, Price et al. 2004a). Owing to its early development, the *E. coli* reconstruction and model has been a popular target for initial screening and development of a number of these methods. In this section, we introduce basic concepts common to most of these methods and describe in more detail those methods that were used in the studies presented in this chapter. The interested reader is encouraged to refer to recently published reviews presenting the constraint-based modeling methods in more detail (Breitling et al. 2008, Palsson 2006, Price et al. 2004a).

Mathematical description of the reconstruction: The metabolic reconstruction consists of a list of biochemical transformations known to take place in the target organism. This reaction list can be readily converted into a mathematical, computable format by using any available parser (e.g. in COBRA toolbox (Becker et al. 2007)). Using a parser, the stoichiometric coefficients are extracted for the individual reactions and entered in the cell of the stoichiometric matrix, also called the S matrix (Fig. 9.6). In this S matrix, every row corresponds to a metabolite and every column corresponds to a network reaction. Note that a typical S matrix

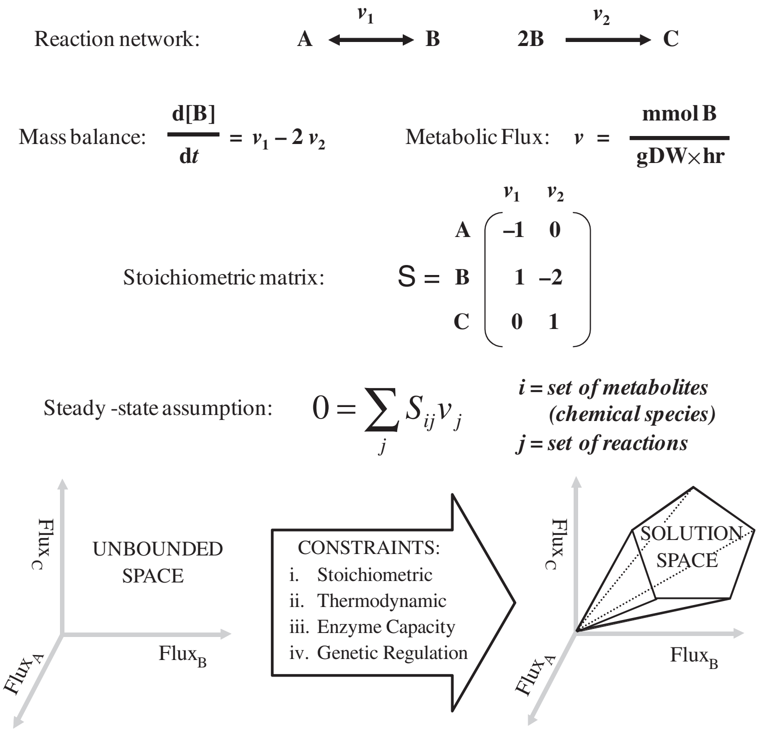

Fig. 9.6 The structure and application of constraints to networks. Shown are the components (Reaction network) and the engineering approaches and equations used to model a reconstructed network. The stoichiometric matrix is a mathematical representation of a reconstructed network and the steady-state assumption is used in a number of COBRA approaches, including flux balance analysis. The *bottom* of the diagram depicts how an unbound space can be confided to a solution space in which a network must behave by imposing the governing physiochemical constraints on a system (e.g., thermodynamic constraints)

is very sparse ($<$ 1% non-zero entries) as many biochemical transformations are bi-linear, and the majority of metabolites appear only in few metabolic reactions. Only a few metabolites, such as protons, water, and ATP, are highly connected in a metabolic network, and participate in many metabolic reactions. Many studies have concentrated on studying the topological features of metabolic networks and the S matrix (see Section 9.5.4 or (Feist and Palsson 2008)).

The multiplication of this S matrix with a flux vector v, containing flux values for all reactions v_j in S, results in a vector listing the changes in concentrations of all metabolites x_i over time:

$$S \bullet v = \frac{dx}{dt} \tag{9.1}$$

The constraint–based modeling approaches are based on the steady state assumption (Fig. 9.6), which assumes that the change of metabolite concentration over time is zero:

$$S \bullet v = \frac{dx}{dt} = 0 \tag{9.2}$$

This assumption is valid for the metabolic reactions as the time scale of the reaction rates is much smaller (milliseconds range) than the doubling time of a cell, which is on the order of hours. Due to this time-scale separation, the metabolic network is essentially in a steady state during cell replication, and as a consequence, intracellular metabolites are not allowed to accumulate. This restriction, imposed by Equation (9.2), is known as the mass-balance constraint (Fig. 9.6).

Further constraints may be added to the reconstruction, leading to the conversion of the reconstruction to a condition-specific model. Such constraints can include thermodynamic (i.e., reaction reversibility), regulatory (e.g., expression of an enzyme), topological (i.e., composition and connectivity of network), and environmental (e.g., presence/absence of a specific carbon source).

Interrogation of the steady state solution space: In most cases, the set of equations encoded in the S matrix are underdetermined, meaning that there are more variables (fluxes v_j for $j = 1 \ldots n$) than there are equations (mass-balances for each metabolite x_i for $i = 1 \ldots m$). As a consequence, there is no single solution or flux vector v satisfying all the equations, but rather there are many possible flux vectors. This set of possible flux vectors is called the steady-state solution space. Each flux vector v, satisfying the given model constraints, is called a functional state of the network. This term functional state can be seen as analogous to the traffic pattern of the road mesh in a large city. The road mesh would correspond to the metabolic network and the traffic pattern, which shows high traffic and low traffic on the highways, corresponds to the functional state of the road system. Clearly this traffic pattern will be very different in the afternoon during rush hour versus the traffic pattern found late into the night. This example highlights the idea that one network can have many distinct functional states.

Functional states of a network can be determined using different mathematical approaches. In the COBRA approach, there is a distinction between biased and unbiased methods. Biased methods require the statement of an objective function, such as a biomass formation reaction or a byproduct secretion reaction by the metabolic network. This objective function is then maximized (or minimized) to obtain a functional state leading to the maximal (or minimal) flux value of the objective function. In contrast, unbiased methods explore the entire steady state solution space by determining a representative subset of possible functional states that can be analyzed in a statistical manner. Examples of unbiased methods are uniform sampling (Almaas et al. 2004, Price et al. 2004b, Thiele et al. 2005a, Wiback et al. 2004) and extreme pathway analysis (Papin et al. 2002, Price et al. 2003).

In many COBRA applications, it is assumed that the aim of a living cell is to grow as fast as possible to outgrow competitors and thus to use available nutrients mainly for biomass production. Hence, many COBRA applications are used in conjunction with the maximization of the biomass production rate. For example, gene essentiality can be determined *in silico* where the essentiality of every gene is tested to see whether the metabolic network is still able to produce biomass despite the *in silico* disruption of a gene (see Fig. 9.9). Other examples in this chapter discuss metabolic engineering applications, where the metabolic network is modified in such way that it produces a desired byproduct while maintaining a certain biomass production capability. Many industrially-interesting byproducts are produced by cells when they cannot produce biomass (e.g., due to nitrogen or phosphate limitations). Thus, the byproduct and biomass production are competitive, or ‘orthogonal’ to each other. COBRA has been successfully used to couple the byproduct production with the biomass production by deleting certain metabolic genes, thereby redirecting carbon fluxes in the metabolic networks (see below). The byproduct coupling to biomass production forces the organism to produce the desired byproduct in order to obtain the cellular objective of biomass production.

9.4 Applications and Uses of the *E. coli* Metabolic Reconstruction

Ask not what you can do for a reconstruction, but what a reconstruction can do for you: The *E. coli* reconstruction and GEM has been adapted for a broad number of uses by research groups around the world. Studies utilizing the reconstructed *E. coli* network range from pragmatic to theoretical applications and address a wide range of questions. These uses can be further categorized into five areas which include: (1) metabolic engineering, (2) biological discovery, (3) assessment of phenotypic behavior, (4) biological network analysis, and (5) studies of bacterial evolution (Fig. 9.7). A more extensive review of these uses has recently appeared (Feist and Palsson 2008), as well as an additional review on metabolic engineering efforts with *E. coli* and other organisms (Kim et al. 2008). Here, key examples of uses of the *E. coli* reconstruction in each of these fields will be presented to demonstrate the utility of the reconstruction and modeling process.

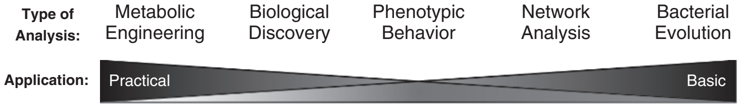

Fig. 9.7 Spectrum of uses of the of the genome-scale E. coli metabolic network reconstruction. Uses of the *E. coli* metabolic reconstruction can be categorized into 5 different areas. Furthermore, these categories can be arranged in order of addressing more practical (e.g., generating a production strain) or more basic (e.g., understanding horizontal gene transfer) questions

9.4.1 Metabolic Engineering

Metabolic engineering efforts utilizing the GEM of *E. coli* have focused on exploring overproduction for a number of products. Three examples in which computation and experimental construction were used to achieve overproduction will be discussed here. The first two examples utilized the *E. coli* GEM to explore the production of the amino acids L-valine (Park et al. 2007) and L-threonine (Lee et al. 2007) in *E. coli*, and each has demonstrated the broad usage of GEM-aided computation for strain design.

Production of L-threonine: In the first study, GEM-aided modeling was employed in three different areas to increase the production of L-threonine to industrial titers (Fig. 9.8) (Lee et al. 2007). In one instance, *in silico* modeling was used to identify the optimal activity of a key enzymatic reaction towards maximum L-threonine production using a parametric sensitivity analysis that compared reaction activity to L-threonine production rate. The optimal activity prediction was subsequently used to tune the over-expression of the gene that encodes for this enzymatic reaction through comparison to base line activity, and the result was a production increase. This method proved to be vital to the success of this strain, as a previous transcription profiling guided attempt at over-expression resulted in an undesirable surplus of activity that was detrimental to L-threonine production.

For the same strain, a GEM-aided flux analysis in conjunction with mRNA expression data levels guided the elimination of negative regulation on a gene, which encoded for a reaction that channeled flux towards the final product. The third use of the GEM for the design of this strain occurred when an unwanted byproduct was observed in the culture medium and computation was utilized to divert the flux from this byproduct to L-threonine (Lee et al. 2007) through over-expression of another key gene encoded activity.

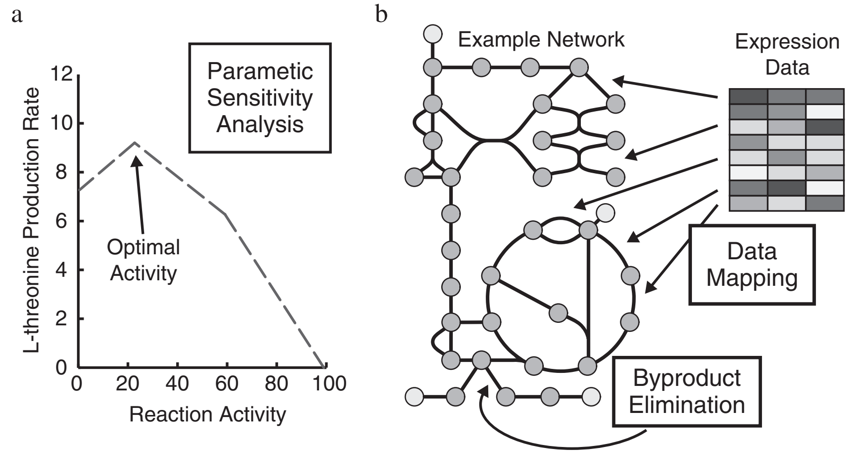

Fig. 9.8 Three different areas where modeling was incorporated to increase strain production. Areas of model-driven strain improvement utilized to overproduce L-threonine in *E. coli* (Lee et al. 2007). (**a**) Shown is a graph that provides the computed relationship between L-threonine production and the activity of particular reaction. This *in silico* parametric sensitivity analysis guided the level of expression necessary for increased production of the amino acid in the strain. (**b**) Given is a map of central metabolism representing the metabolic reconstruction of *E. coli*. In the analysis, expression data was mapped onto the network to guide the elimination of negative regulation and the network was used to overexpress a reaction that diverted flux away from a byproduct (byproduct elimination) towards the desired product

Production of lycopene: Lycopene is an important intermediate in the biosynthesis of many carotenoids, and it is used for food coloring as it possesses a strong color (bright red) and is non-toxic. To increase the production of an already high-producing strain, a systematic computational search was developed (Alper et al. 2005b) to explore the *E. coli* metabolic network and report gene deletions that diverted metabolic flux towards the desired product. This process resulted in a knock-out strain that, when constructed, showed a two fold increase in the production of lycopene over the parental strain. In this analysis, the minimization of metabolic adjustment (MOMA) computational algorithm (Segre et al. 2002) and the IJE660 (Edwards and Palsson 2000) *E. coli* GEM were utilized to sequentially examine additive genetic deletions that would improve lycopene production while maintaining cell viability. It was found that this computational approach yielded a twofold increase in production rate over a previously engineered overproducing strain and an eightfold increase over wild-type production harboring only a lycopene biosynthesis plasmid (Alper et al. 2005b). In addition, the strain designs identified computationally were compared to mixed combinatorial transposon mutagenesis, and it was found that the maximum production observed could be designed solely using the systematic GEM-aided computational method (Alper et al. 2005a,b). Furthermore, a deleterious effect was observed when targets identified in individual computational designs were combined in an attempt to achieve an overall more desirable phenotype. Thus, the overall systematic effects from individual designs were not additive and needed to be interpreted in the context of the entire network.

Production of L-valine: This model-driven example of metabolic engineering demonstrates the use of applying a systematic computational search algorithm (Alper et al. 2005b) to the updated *E. coli* GEM MBEL979 (Lee et al. 2005) (similar to the *i*JR904 GEM (Reed et al. 2003)) to improve L-valine production. In this analysis, the *in silico* computation of beneficial knock-outs to divert flux towards the desired product once again resulted in a significant increase (greater than twofold) in the production of the desired metabolite over an existing overproducing strain (Park et al. 2007). A number of additional metabolic engineering approaches to increase overproduction were performed by, (i) relieving feedback inhibition and regulation through attenuation, (ii) removing competing pathways, (iii) up-regulation of primary biosynthetic pathways, and (iv) over-expression of export machinery. When compared to each of the other individual strain modifications, the *in silico* GEM aided interventions resulted in the greatest increase in L-valine production (Park et al. 2007). Taken together, this and the previous study demonstrate the broad applications for which GEMs can be utilized to design strains not only in a *de novo* fashion, but to make further improvements on strains through integrating and interpreting experimental data.

9.4.2 Biological Discovery

The GEM of *E. coli* can be used as a guide to discovery. There is still a significant amount of information missing relating to gene functions in *E. coli* (Riley et al.

2006), and the content contained in the *E. coli* reconstruction can be queried and analyzed to first, determine the current gaps in our knowledge of the organism and second, design experiments to specifically fill uncovered gaps in the knowledge landscape. Two examples of model-driven discovery are presented, and these studies should form the basis for further analysis. To uncover the genetic basis for experimentally observed functions in *E. coli*, the studies combined GEM-aided computation with guided experimentation.

Systems approach to refining genome annotation: The first study utilized an iterative process (Reed et al. 2006b) in which, (i) differences in modeling predictions and high-throughput growth phenotype data were identified, (ii) potential missing reactions that remedy these disagreements were algorithmically determined, (iii) bioinformatics was utilized to identify likely encoding ORFs, and (iv) resulting targeted ORFs were cloned and experimentally characterized. Application of this process led to the functional characterization of eight ORFs that are involved in transport, regulatory and metabolic functions in *E. coli* (Reed et al. 2006b). The discovery process was aided by a high-throughput growth phenotyping analysis and the genome-wide single-gene mutant collection (Baba et al. 2006), along with other characterization analyses such as targeted expression profiling. This work was the first such example of model-driven discovery of genome content aided by a metabolic network reconstruction.

Genetic basis of orphan reactions: The second GEM-based analysis that resulted in ORF discovery utilized network topology to examine orphan reactions in the *E. coli* network (i.e., reactions known to exist in *E. coli* that have not been linked to an encoding gene) identified by network topology-based gap-filling algorithms (Chen and Vitkup 2006, Kharchenko et al. 2006, 2004). The basic premise behind these algorithms is the utilization of an orphan reaction's network neighbors as constraints to assign metabolic function. With the resulting tentative ORF assignments, biochemical characterization studies utilizing genetic mutants (Baba et al. 2006), analysis of growth under different substrate conditions, and expression data were all utilized to characterize and assign function to an orphan ORF that is responsible for a metabolic conversion that has been known for 25 years (Fuhrer et al. 2007). These two studies are early examples of how GEM computation can lead to the discovery of new genetic and biochemical content in an organism.

9.4.3 Assessment of Phenotypic Behavior

Researchers have utilized the *E. coli* GEM to better understand the coordinated functions of the cell and observed physiological outcomes. Computations seeking to predict cellular phenotypes have been performed under a range of genetic and environmental conditions, and phenotypic assessment has received the most attention in terms of publication and tool development. Here, we outline computational tools developed to analyze the *E. coli* GEM in each of the two major areas of phenotypic assessment, studies of (i) network perturbation/essentiality, and (ii) the incorporation of thermodynamic information.

a

Gene	Glucose	Glycerol	Succinate	Acetate
aceA	+ / +		+ / +	− / −
aceB				− / −
aceEF	− / +			
ackA				+ / +
acn	− / −			− / −
acs				+ / +
cyd	+ / +			
cyo	+ / +			
eno	− / +	− / +	− / −	− / −
fba	− / +			
fbp	+ / +	− / −	− / −	− / −
frd	+ / +		+ / +	+ / +
gap	− / −	− / −	− / −	− / −
glk	+ / +			
gltA	− / −			− / −
gnd	+ / +			
idh	− / −			− / −
mdh	+ / +	+ / +	+ / +	
ndh	+ / +	+ / +		
nuo	+ / +	+ / +		
pfk	− / +			
pgi	+ / +	+ / −	+ / −	
pgk	− / −	− / −	− / −	− / −
pgl	+ / +			
pntAB	+ / +	+ / +	+ / +	
ppc	± / +	− / +	+ / +	
pta				+ / +
pts	+ / +			
pyk	+ / +			
rpi	− / −	− / −	− / −	− / −
sdhABCD	+ / +		− / −	− / −
sucAB	+ / +		− / +	− / +
tktAB	− / −			
tpi	− / +	− / −	− / −	− / −
unc	+ / +		± / +	− / −
zwf	+ / +	+ / +	+ / +	

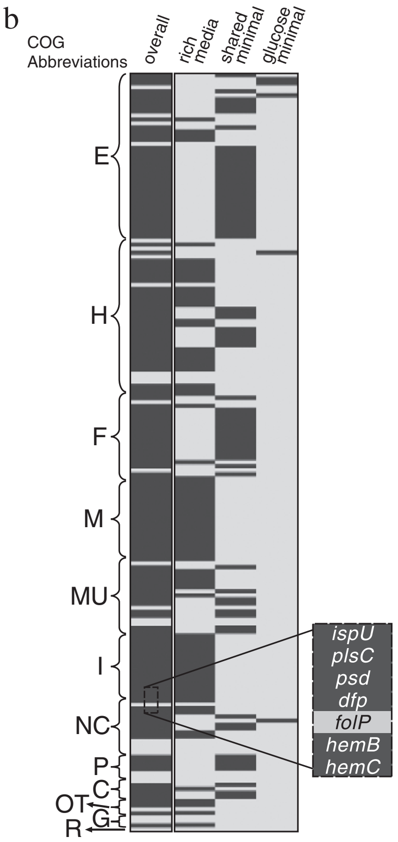

Fig. 9.9 Gene-deletion analyses utilizing the E. coli GEM. Analyses of gene essentiality in the *E. coli* metabolic network. (**a**) A table of results from an analysis performed using the *i*JE660 GEM of *E. coli* where experimental phenotypes were collected from bibliomic data. Results are scored as + or − meaning growth or no growth determined from *in vivo*/*in silico data*. The ± indicates that suppressor mutations have been observed that allow the mutant strain to grow. In 68 of 79 cases the *in silico* behavior is the same as the experimentally observed behavior. Each column represents a different carbon source. (**b**) This heat map characterizes the agreement between ORFs predicted to be essential using the *i*AF1260 GEM of *E. coli* (Feist et al. 2007) and those experimentally determined (Baba et al. 2006, Joyce et al. 2006). The enlarged region details how each row corresponds to a computationally predicted essential ORF (188 total). A dark row indicates the condition under which each ORF was found to be essential. For example, *folP* was predicted to be an essential ORF for the biosynthesis of folate in *i*AF1260 under these conditions, but was not identified as essential by Baba et al. (2006). The different columns show at which level each gene in the overall column was found to be essential on. With the advancement of both experimental data and model coverage, analyses of this type have reached the genomic scale

***In silico* perturbations**: A set of distinct computational methods using GEMs has been developed to determine the physiological state of *E. coli* (and other cells for which a GEM exists) after genetic perturbations (Segre et al. 2002, Shlomi et al. 2005, Wunderlich and Mirny 2006). These methods were analyzed to examine the effectiveness of predictions when compared to experimental data (Fig. 9.9). Whereas comparisons to flux data from wild-type and *E. coli* mutants reveals that one of the computational algorithms, MOMA (Segre et al. 2002), provided better predictions for transient growth rates (early post perturbation state), another algorithm, ROOM (Shlomi et al. 2005) (and basic FBA), was found to be more successful in predicting final steady-state growth rates and overall lethality (Shlomi et al. 2005). These two algorithms have been utilized, in addition to basic FBA, for genome-wide essentiality screens. Aiding the effort is the recent availability of a comprehensive single-gene knock-out library for *E. coli* (Baba et al. 2006) which has been utilized for comparison with GEM computation (Feist et al. 2007, Joyce et al. 2006). Touching on the predictive capability of GEM computations, it was found that the *E. coli* GEM was able to predict the outcomes of adaptively evolved strains to a high degree (78%) when knock-out *E. coli* strains were grown in a number of different substrate environments by examining growth rates at the beginning and end of adaptive evolution (Fong and Palsson 2004). Genetic perturbations have played a key role in the study of the genotype-phenotype relationship in biology, and GEMs can be used to mechanistically interpret the results and predict the outcomes of such perturbations.

Adding thermodynamic information: The incorporation of thermodynamic information with GEMs is an effort that is progressing rapidly and should increase the predictive capabilities of genome-scale modeling through the addition of further governing physico-chemical constraints. Furthermore, the addition of thermodynamics enables the analysis of metabolomic data in the context of a reconstruction. A study utilizing high-throughput metabolomic data and GEMs resulted in the proposition of likely regulatory interactions by deciphering the metabolite concentrations in the context of overall network functionality (Kümmel et al. 2006). Not only did the metabolomic data benefit computations by constraining the system using physiological measurements, but the computational predictions were also able to validate quantitative metabolomic data sets for consistency through providing a functional context to relate metabolite concentrations. This application is one example of how metabolomic data will directly influence modeling. Metabolite concentration data is likely to greatly influence future metabolic modeling due to its intimate connection with GEM content.

9.4.4 Biological Network Analysis

Although there is still much to learn about the metabolism of *E. coli* and how a model-driven approach can be used to uncover these unknowns, the wealth of knowledge collected and represented in the current *E. coli* reconstruction makes it an ideal platform for network analyses. Researchers have been taking advantage of this fact and have centered network analyses on probing and uncovering the properties of biological networks in general. In this section, we discuss a key analysis based on the *E. coli* GEM and the implications drawn from such analyses.

One noteworthy study utilizing the *E. coli* network examined thousands of different potential growth conditions and resulted in the observation of a ‘high-flux backbone’ in *E. coli* that both, (i) carried high levels of flux across the different environmental conditions, and (ii) was composed of a relatively small set of enzymatic reactions (Almaas et al. 2004). This result can be of practical importance for synthetic biology efforts aimed towards manipulating flux within biological systems. Furthermore, this finding was hypothesized to be a universal feature of metabolic activity in all cells and was consistent with flux measurements from 13C labeling experiments (Almaas et al. 2004).

Overall, studies of network analyses have a common systems biology theme: the development and subsequent demonstration of methods that identify sets of reactions or metabolites with correlated or coordinated functions and systematic relationships. The systems biology that these methods enable and demonstrate has the potential to influence the more practical applications already outlined. The role that the *E. coli* GEM has taken is a comprehensive and curated set of up-to-date metabolic knowledge that provides a scaffold for large-scale computations.

9.4.5 Studies of Bacterial Evolution

The GEMs of *E. coli* have been used to examine the process of bacterial evolution (Pal et al. 2005a,b, 2006). Specifically, the network reconstructions have been used to interpret adaptive evolution events (Pal et al. 2005a), horizontal gene transfer (Pal et al. 2005a,b) and evolution to minimal metabolic networks (Pal et al. 2006). These studies, which utilize the *E. coli* reconstruction as an organism-specific genetic and metabolic content database and the corresponding GEM have been able to provide insight into evolutionary events through combining known physiological data (e.g., in various environmental conditions) with hypotheses and *in silico* computation. Examination of the evolution of minimal metabolic networks through simulation demonstrated that it was possible to predict the gene content of close relatives of *E. coli* by examining the necessity of genes and reactions in the overall context of the system functionality for a specific lifestyle (Pal et al. 2006). Similarly, by re-examining network functionality in a number of different environments, and through the utilization of comparative genomics, it was shown that recent evolutionary events (i.e., horizontal gene transfer) likely resulted from a response to a change in environment (Pal et al. 2005a). Furthermore, computational analysis led to the additional conclusion that these horizontal gene transfer events are more likely when the host organism contains an enzyme that catalyzes a coupled metabolic flux related to the transferred enzyme’s function (Pal et al. 2005a,b). Taken together, these studies demonstrate the importance of having high-quality curated reconstructions to enable studies on an organism’s response to environmental changes and on the fundamental forces driving bacterial evolution.

9.5 Need for New *In Silico* Methods and Applications

We now know how to represent BiGG data in either a stoichiometric format or in the form of causal relationships (Gianchandani et al. 2006) and how to use this data to perform several lines of computational inquiries. Computational query tools of GEMs will continue to be developed. New advances in these query tools will likely include, (i) modularization methods, (ii) use of fluxomic data, and (iii) eventually, kinetic information.

Modularization: As the scope and content of the reconstruction grows, the need to modularize its content becomes more pressing. Fine or coarse-grained views of cellular processes are needed for different applications.

Fluxomics: Currently, computational limitations force the reduction in network size for the analysis of isotopomer data. Given the systemic nature of fluxomic data and its phenotypic relevance, there is a pressing need to increase the size of the networks that can be utilized for experimental measurement and estimation of flux states. A network reconstruction will both guide the content that is needed for analyzing fluxomic data and offer a starting point for a rational reduction to generate relevant models in the meantime.

Kinetics/thermodynamics: Although detailed kinetic models of microbial functions may currently be mostly of academic interest, they will most likely be able to be constructed in the mid-term based on advances with metabolomic and fluxomic data, in addition to the developments that are occurring with the incorporation of thermodynamic information. Such large-scale kinetic models are likely to differ from those resulting from traditional approaches for construction of kinetic models as they come with different challenges.

9.6 Closing

The process underlying the *E. coli* metabolic reconstruction has pioneered many approaches, methods, and studies in the systems biology of microbial metabolism. This effort has effectively put a mechanistic basis into the genotype-phenotype relationship. In fact, this relationship is now broken down into four steps:

(1) Components (a large knowledge base, BiGG), leading to networks (the reconstruction process resulting GENRE), leading to *In silico* Models (GEMs), leading to Phenotypic States (estimated by COBRA methods).
(2) GEMs will allow for gap-filling and systematic biological discovery (Breitling et al. 2008) and for understanding of complex biological processes (see Chapter 15).

Predictive models also allow for experimental strain design. In fact, in engineering, there is '*nothing more practical than a good theory*.' As this chapter demonstrated, genomics and high-throughput technologies have enabled the construction of predictive computational models. The scope of such predictions is limited at the moment, but with the growing scope and coverage of genome-scale reconstructions and advancements in the development of computational tools, this scope will broaden. Not only will GEMs influence design in synthetic biology, but also their help with discovering cellular content will provide a more complete picture of the intra-cellular environment in which future synthetically engineered constructs and circuits will be placed. The impact of GEMs on synthetic biology is thus likely to be notable, ranging from the provision of the cellular context of a small-scale gene circuit design to engineering of the entire genome-scale network towards fundamentally new and useful (i.e., production) phenotypes.

Finally, we can speculate about the deep scientific impact that comprehensive predictive GEMs will have on our understanding of the living process. A comprehensive view of cellular functions will allow us to study the fundamental properties of both the underlying energy and information flows in living organisms. Such a view is likely to deeply affect our understanding of both distal and proximal causation in biology.

References

Allen TE, Palsson BO (2003) Sequenced-Based Analysis of Metabolic Demands for Protein Synthesis in Prokaryotes. J Theor Biol 220(1):1–18

Almaas E, Kovacs B, Vicsek T et al. (2004) Global organization of metabolic fluxes in the bacterium *Escherichia coli*. Nature 427(6977):839–43

Alper H, Jin YS, Moxley JF et al. (2005a) Identifying gene targets for the metabolic engineering of lycopene biosynthesis in *Escherichia coli*. Metab Eng 7(3):155–64

Alper H, Miyaoku K, Stephanopoulos G (2005b) Construction of lycopene-overproducing *E. coli* strains by combining systematic and combinatorial gene knockout targets. Nat Biotechnol 23(5):612–6

Baart GJ, Zomer B, de Haan A et al. (2007) Modeling Neisseria meningitidis metabolism: from genome to metabolic fluxes. Genome Biol 8(7):R136

Baba T, Ara T, Hasegawa M et al. (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol 2:2006.0008

Becker SA, Feist AM, Mo ML et al. (2007) Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox. Nat Protocols 2(3):727–38

Becker SA, Palsson BO (2005) Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation. BMC Microbiol 5(1):8

Berman HM, Westbrook J, Feng Z et al. (2000) The Protein Data Bank. Nucleic Acids Res 28(1):235–42

Beste DJ, Hooper T, Stewart G et al. (2007) GSMN-TB: a web-based genome-scale network model of Mycobacterium tuberculosis metabolism. Genome Biol 8(5):R89

Blattner FR, Plunkett G, 3rd, Bloch CA et al. (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277(5331):1453–74

Borodina I, Krabben P, Nielsen J (2005) Genome-scale analysis of Streptomyces coelicolor A3(2) metabolism. Genome Res 15(6):820–9

Breitling R, Vitkup D, Barrett MP (2008) New surveyor tools for charting microbial metabolic maps. Nat Rev Microbiol 6(2):156–61

Chavali AK, Whittemore JD, Eddy JA et al. (2008) Systems analysis of metabolism in the pathogenic trypanosomatid Leishmania major. Mol Syst Biol 4:177

Chen L, Vitkup D (2006) Predicting genes for orphan metabolic activities using phylogenetic profiles. Genome Biol 7(2):R17

Cho BK, Knight EM, Barrett CL et al. (2008) Genome-wide Analysis of Fis Binding in *Escherichia coli* Indicates a Causative Role for A-/AT-tracts. Genome Res 18(6):900–10

Covert MW, Knight EM, Reed JL et al. (2004) Integrating high-throughput and computational data elucidates bacterial networks. Nature 429(6987):92–6

David H, Ozcelik IS, Hofmann G et al. (2008) Analysis of Aspergillus nidulans metabolism at the genome-scale. BMC Genomics 9:163

Duarte NC, Becker SA, Jamshidi N et al. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci USA 104(6):1777–82

Duarte NC, Herrgard MJ, Palsson B (2004) Reconstruction and Validation of *Saccharomyces cerevisiae* iND750, a Fully Compartmentalized Genome-Scale Metabolic Model. Genome Res 14(7):1298–309

Edwards JS, and Palsson, B.O. (2000a) Metabolic flux balance analysis and the in silico analysis of *Escherichia coli* K-12 gene deletions. BMC Bioinformatics 1(1)

Edwards JS, Ibarra RU, Palsson BO (2001) *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. Nat Biotechnol 19(2):125–30

Edwards JS, Palsson BO (1999) Systems properties of the Haemophilus influenzae Rd metabolic genotype. J Biol Chem 274(25):17410–6

Edwards JS, Palsson BO (2000b) The *Escherichia coli* MG1655 *in silico* metabolic genotype: Its definition, characteristics, and capabilities. Proc Natl Acad Sci USA 97(10):5528–33

Feist AM, Henry CS, Reed JL et al. (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 3(121)

Feist AM, Herrgard MJ, Thiele I et al. (2009) Reconstruction of biochemical networks in microbial organisms. Nat Rev Microbiol 7(2)

Feist AM, Palsson BO (2008) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. Nat Biotech 26(6):659–67

Feist AM, Scholten JCM, Palsson BO et al. (2006) Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*. Mol Syst Biol 2(2006.0004):1–14

Fong SS, Palsson BO (2004) Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes. Nat Genet 36(10):1056–58

Forster J, Famili I, Fu PC et al. (2003) Genome-Scale Reconstruction of the *Saccharomyces cerevisiae* Metabolic Network. Genome Res 13(2):244–53

Frazier ME, Johnson GM, Thomassen DG et al. (2003) Realizing the potential of the Genome Revolution: The Genomes to life Program. Science 300(5617):290–3

Fuhrer T, Chen L, Sauer U et al. (2007) Computational prediction and experimental verification of the gene encoding the NAD+/NADP+-dependent succinate semialdehyde dehydrogenase in *Escherichia coli*. J Bacteriol 189(22):8073–8

Gianchandani EP, Papin JA, Price ND et al. (2006) Matrix Formalism to Describe Functional States of Transcriptional Regulatory Systems. PLoS Comput Biol 2(8):e101

Gonzalez O, Gronau S, Falb M et al. (2008) Reconstruction, modeling & analysis of *Halobacterium salinarum R-1* metabolism. Mol Biosyst 4(2):148–59

Heinemann M, Kummel A, Ruinatscha R et al. (2005) In silico genome-scale reconstruction and validation of the *Staphylococcus aureus* metabolic network. Biotechnol Bioeng 92(7):850–64

Herrgard MJ, Covert MW, Palsson BO (2004) Reconstruction of Microbial Transcriptional Regulatory Networks. Curr Opin Biotechnol 15(1):70–7

Holden C (2002) Alliance launched to model *E. coli*. Science 297(5586):1459–60

Ibarra RU, Edwards JS, Palsson BO (2002) *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. Nature 420(6912):186–9

Jamshidi N, Palsson BO (2007) Investigating the metabolic capabilities of *Mycobacterium tuberculosis H37Rv* using the in silico strain iNJ661 and proposing alternative drug targets. BMC Syst Biol 1:26

Joyce AR, Palsson BO (2006) The model organism as a system: integrating 'omics' data sets. Nat Rev Mol Cell Biol 7(3):198–210

Joyce AR, Reed JL, White A et al. (2006) Experimental and Computational Assessment of Conditionally Essential Genes in *Escherichia coli*. J Bacteriol 188(23):8259–71

Kümmel A, Panke S, Heinemann M (2006) Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. Mol Syst Biol 2:2006.0034

Keseler IM, Collado-Vides J, Gama-Castro S et al. (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. Nucleic Acids Res 33(Database Issue):D334–7

Kharchenko P, Chen L, Freund Y et al. (2006) Identifying metabolic enzymes with multiple types of association evidence. BMC Bioinformatics 7(177)

Kharchenko P, Vitkup D, Church GM (2004) Filling gaps in a metabolic network using expression information. Bioinformatics 20(Suppl 1):I178-I185

Kim HU, Kim TY, Lee SY (2008) Metabolic flux analysis and metabolic engineering of microorganisms. Mol BioSyst 4(2):113–20

Kim TY, Kim HU, Park JM et al. (2007) Genome-scale analysis of *Mannheimia succiniciproducens* metabolism. Biotechnol Bioeng 97(4):657–71

Kuepfer L, Sauer U, Blank LM (2005) Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. Genome Res 15(10):1421–30

Lee J, Yun H, Feist AM et al. (2008) Genome-scale reconstruction and in silico analysis of the *Clostridium acetobutylicum* ATCC 824 metabolic network. Appl Microbiol Biotechnol 80(5):849–52

Lee KH, Park JH, Kim TY et al. (2007) Systems metabolic engineering of *Escherichia coli* for L-threonine production. Mol Syst Biol 3:149

Lee SY, Woo HM, Lee D-Y et al. (2005) Systems-level analysis of genome-scale *in silico* metabolic models using MetaFluxNet. Biotechnol Bioproc Eng 10:425–31

Lee TI, Rinaldi NJ, Robert F et al. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 298(5594):799–804.

Mahadevan R, Bond DR, Butler JE et al. (2006) Characterization of Metabolism in the Fe(III)-Reducing Organism *Geobacter sulfurreducens* by Constraint-Based Modeling. Appl Environ Microbiol 72(2):1558–68

Majewski RA, Domach MM (1990) Simple constrained optimization view of acetate overflow in *E. coli*. Biotechnol Bioeng 35:732–8

Mehra A, Hatzimanikatis V (2006) An algorithmic framework for genome-wide modeling and analysis of translation networks. Biophys J 90(4):1136–46

Nogales J, Thiele, I.*, Palsson, B. Ø. (2008) A genome-scale metabolic reconstruction for *P. putida* KT2440: *i*JN746 as cell factory

Oberhardt MA, Puchalka J, Fryer KE et al. (2008) Genome-scale metabolic network analysis of the opportunistic pathogen *Pseudomonas aeruginosa PAO1*. J Bacteriol 190(8):2790–803

Oh YK, Palsson BO, Park SM et al. (2007) Genome-scale reconstruction of metabolic network in bacillus subtilis based on high-throughput phenotyping and gene essentiality data. J Biol Chem 282(89):28791–9

Oliveira AP, Nielsen J, Forster J (2005) Modeling Lactococcus lactis using a genome-scale flux model. BMC Microbiol 5:39

Pal C, Papp B, Lercher MJ (2005a) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet 37(12):1372–5

Pal C, Papp B, Lercher MJ (2005b) Horizontal gene transfer depends on gene content of the host. Bioinformatics 21 Suppl 2:ii222–3

Pal C, Papp B, Lercher MJ et al. (2006) Chance and necessity in the evolution of minimal metabolic networks. Nature 440(7084):667–70

Palsson BO (2004) Two-dimensional annotation of genomes. Nat Biotechnol 22(10):1218–9

Palsson BO (2006) Systems biology: properties of reconstructed networks. Cambridge University Press, New York

Papin JA, Hunter T, Palsson BO et al. (2005) Reconstruction of cellular signalling networks and analysis of their properties. Nat Rev Mol Cell Biol 6(2):99–111

Papin JA, Price ND, Palsson BO (2002) Extreme pathway lengths and reaction participation in genome-scale metabolic networks. Genome Res 12(12):1889–900

Park JH, Lee KH, Kim TY et al. (2007) Metabolic engineering of *Escherichia coli* for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation. Proc Natl Acad Sci USA 104(19):7797–802

Pramanik J, Keasling JD (1997) Stoichiometric model of *Escherichia coli* metabolism: Incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. Biotechnol Bioeng 56(4):398–421

Pramanik J, Keasling JD (1998) Effect of *Escherichia coli* biomass composition on central metabolic fluxes predicted by a stoichiometric model. Biotechnol Bioeng 60(2):230–8

Price ND, Reed JL, Palsson BO (2004a) Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2(11):886–97

Price ND, Schellenberger J, Palsson BO (2004b) Uniform Sampling of Steady State Flux Spaces: Means to Design Experiments and to Interpret Enzymopathies. Biophys J 87(4):2172–86

Price ND, Reed JL, Papin JA et al. (2003) Analysis of metabolic capabilities using singular value decomposition of extreme pathway matrices. Biophys J 84(2):794–804

Reed JL, Famili I, Thiele I et al. (2006a) Towards multidimensional genome annotation. Nat Rev Genet 7(2):130–41

Reed JL, Palsson BO (2003) Thirteen Years of Building Constraint-Based In Silico Models of *Escherichia coli*. J Bacteriol 185(9):2692–9

Reed JL, Patel TR, Chen KH et al. (2006b) Systems approach to refining genome annotation. Proc Natl Acad Sci USA 103(46):17480–4

Reed JL, Vo TD, Schilling CH et al. (2003) An expanded genome-scale model of *Escherichia coli* K-12 (*i*JR904 GSM/GPR). Genome Biol 4(9):R54.1–R54.12

Resendis-Antonio O, Reed JL, Encarnacion S et al. (2007) Metabolic reconstruction and modeling of nitrogen fixation in Rhizobium etli. PLoS Comput Biol 3(10):1887–95

Riley M, Abe T, Arnaud MB et al. (2006) *Escherichia coli* K-12: a cooperatively developed annotation snapshot-2005. Nucleic Acids Res 34(1):1–9

Schilling CH, Covert MW, Famili I et al. (2002) Genome-scale metabolic model of *Helicobacter pylori* 26695. J Bacteriol 184(16):4582–93

Schilling CH, Palsson BO (2000) Assessment of the Metabolic Capabilities of *Haemophilus influenzae Rd* through a Genome-scale Pathway Analysis. J Theor Biol 203(3):249–83

Segre D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci USA 99(23):15112–7

Senger RS, Papoutsakis ET (2008) Genome-scale model for *Clostridium acetobutylicum*. Part 1: Metabolic network resolution and analysis. Biotechnol Bioeng 101(5):1036–52

Serres MH, Gopal S, Nahum LA et al. (2001) A functional update of the *Escherichia coli* K-12 genome. Genome Biol 2(9):RESEARCH0035

Sheikh K, Forster J, Nielsen LK (2005) Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of *Mus musculus*. Biotechnol Prog 21(1):112–21

Shlomi T, Berkman O, Ruppin E (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. Proc Natl Acad Sci USA 102(21):7695–700

Suthers PF, Burgard AP, Dasika MS et al. (2007) Metabolic flux elucidation for large-scale models using 13C labeled isotopes. Metab Eng 9(5–6):387–405

Teusink B, Wiersma A, Molenaar D et al. (2006) Analysis of growth of *Lactobacillus plantarum* WCFS1 on a complex medium using a genome-scale metabolic model. J Biol Chem 281(52):40041–8

Thiele I, Jamshidi N, Fleming RMT et al. (2009) Genome-scale reconstruction of *E. coli*'s transcriptional and translational machinery: A knowledge-base its mathematical formulation, and its functional characterization. PLOS Comp Biol. In press

Thiele I, Price ND, Vo TD et al. (2005a) Candidate metabolic network states in human mitochondria: Impact of diabetes, ischemia, and diet. J Biol Chem 280(12):11683–95

Thiele I, Vo TD, Price ND et al. (2005b) An Expanded Metabolic Reconstruction of *Helicobacter pylori* (*i*IT341 GSM/GPR): An *in silico* genome-scale characterization of single and double deletion mutants. J Bacteriol 187(16):5818–30

Thomas R, Paredes CJ, Mehrotra S et al. (2007) A model-based optimization framework for the inference of regulatory interactions using time-course DNA microarray expression data. BMC Bioinformatics 8:228

Varma A, Boesch BW, Palsson BO (1993a) Biochemical production capabilities of *Escherichia coli*. Biotechnol Bioeng 42(1):59–73

Varma A, Boesch BW, Palsson BO (1993b) Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates. Appl Environ Microbiol 59(8):2465–73

Varma A, Palsson BO (1993) Metabolic capabilities of *Escherichia coli*: I. Synthesis of biosynthetic precursors and cofactors. J Theor Biol 165(4):477–502

Varma A, Palsson BO (1994) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. Appl Environ Microbiol 60(10):3724–31

Varma A, Palsson BO (1995) Parametric sensitivity of stoichiometric flux balance models applied to wild-type *Escherichia coli* metabolism. Biotechnol Bioeng 45(1):69–79

Wiback SJ, Famili I, Greenberg HJ et al. (2004) Monte Carlo Sampling Can Be Used to Determine the Size and Shape of the Steady State Flux Space. J Theor Biol 228(4):437–47

Wunderlich Z, Mirny LA (2006) Using the topology of metabolic networks to predict viability of mutant strains. Biophys J 91(6):2304–11