1. Introduction

The overall goal of systems biology is to interpret and predict whole-cell physiology brought about by a highly interconnected network of biochemical components. It can be achieved by quantitatively studying cellular processes as a whole system through the integration of experimental and computational approaches\(^{(50)}\). Systems biology differs from traditional biological research approaches in that it aims to study the interactions between the biochemical network components rather than the properties of the individual components themselves\(^{(5)}\). This field can be viewed as a genome-scale science that is enabled by the development of high-throughput, or “omics” technologies\(^{(52,79)}\).

Recent advances in high-throughput experimental technologies have led to an explosion of diverse, genome-based datasets for a large number of organisms. The complete genome sequence for over 250 species is already available and this number is growing rapidly\(^{(55)}\). The genome sequence data and its bioinformatic analyses provide information about open reading frames and their associated chromosomal location, putative function, and associated regulatory sequences\(^{(27)}\). DNA or oligonucleotide microarray technology\(^{(18)}\) affords the ability for researchers to probe gene expression and regulation patterns of cells and tissues on a genome-scale. The combination of chromatin immunoprecipitation and microarray technologies, commonly known as ChIP-chip, provides binding site information of transcription factors for the entire cell\(^{(18)}\). Proteomics\(^{(47)}\), interactomics\(^{(17)}\),
and metabolomics (Kell, 2004) allow scientists to quantitatively examine global protein expression, protein-protein interactions, and metabolite concentrations in a cell, respectively. Metabolism-wide flux measurements, or fluxomics, allows for the observation of the combined functional output of the transcriptome, proteome and metabolome (67). Furthermore, a two-dimensional phenotype microarray technology (7) can be used to measure the effects of hundreds or thousands of environmental conditions on cellular fitness simultaneously, yielding so-called phenomics data. These diverse data types provide a detailed (and heterogeneous) molecular components catalog and interaction list, as well as the functional capabilities of a living organism.

A major challenge currently facing biologists is identifying, extracting, and using the most relevant information from these large-scale data as well as extensive existing reductionist datasets. This is important in order to gaining insight into complex biological processes of interest (52). To overcome this challenge, researchers will benefit from the use of mathematical or computational representations of cellular processes because of: 1) the ability to evaluate and compare different types of “omics” data that consist of tens of thousands of measurements; 2) the capacity for rapid and objective data processing; 3) the suitability for analysis of organisms that cannot be easily cultured in the laboratory, including human pathogens; and 4) the generation of specific hypotheses that can then be tested experimentally (52,76). Therefore, cellular model development is at the core of systems biology. It can be viewed as an iterative process where a model is developed and continuously improved through integrative analyses with high-throughput data (8,50).

Many different mathematical approaches have been formulated to reconstruct cellular processes, including kinetic (42,75), stochastic (4,44) and cybernetic (78,29) methods. These approaches are well developed for small-scale biochemical networks, but it is currently difficult to use them to model genome-scale networks. These models generally require a large number of kinetic parameters. These parameters are often difficult, if not impossible, to determine experimentally. They may differ between in vivo and in vitro environments, and they can change over time due to evolutionary pressure. Accordingly, obtaining accurate values for these parameters remains a key challenge (51). Therefore, at present, these modeling strategies are not very useful for studying systems properties of overall cellular functions and for integrating genome-scale, “omics” data (8,28,56).

To date, the data-driven, constraint-based modeling of cellular metabolism represents a successful story in developing genome-scale models. Unlike other modeling strategies, this constraint-based modeling approach does not attempt to calculate and predict exactly what a biochemical network does. Rather, it seeks to clearly distinguish these network states that a system can achieve from those that it cannot, based on the successive imposition of governing physicochemical constraints (50,51). This approach has been employed to generate many genome-scale networks from each of the three major domains of the tree of life: archaea
(i.e. \textit{Methanococcus jannaschii} \cite{76}), bacteria (i.e. \textit{Escherichia coli} \cite{65}); and eukarya (i.e. \textit{Saccharomyces cerevisiae} \cite{19}). Furthermore, a variety of analytical tools have been developed to probe these models at the systems level \cite{64}. These \textit{in silico} models have applications both in studying biological phenomena (evolutionary adaptation) \cite{32}; two-dimensional annotation of genomes, \cite{53}; and, uncovering transcriptional regulation of metabolism, \cite{51}; as well as in designing microbial strains for the industrial production of biochemicals \cite{10,25,58}. This review describes the constraint-based modeling approach for genome-scale reconstruction. Also discussed are recent achievements in “omic” data analysis guided by metabolic models and some of the potential biotechnological applications.

2. Building constraint-based genome-scale \textit{in silico} models

This section outlines the general procedure followed in reconstructing constraint-based genome-scale model. The approach can be divided approximately into three successive steps (Fig. 1):

I. Network reconstruction.

II. Stoichiometric (S) matrix compilation and imposition of constraints.

III. Linear optimization and other analytical techniques.

2.1. Network reconstruction

The first step in constraint-based modeling, known as network reconstruction, involves generating a model that describes the system of interest. This process can be decomposed into three parts typically performed simultaneously during model construction. These components are known as data collection, metabolic reaction list generation, and gene-protein-reaction relationship determination.

2.1.1. Data collection

Perhaps the most critical component of the constraint-based modeling approach involves data collection relevant to the system of interest. Not long ago, this was one of the most challenging steps. Researchers had access to very limited amounts of genomic and biochemical data. However, the success of recent genome sequencing and annotation projects, advances in high-throughput technologies, as well as the development of detailed and extensive online database resources, have dramatically improved this process.

After identifying the system or organism of interest, relevant data sources must be identified to begin compiling the appropriate metabolites, biochemical reactions, and associated genes to be included in the model. The three primary types of resources are the biochemical literature, high-throughput data, and integrative biochemical database resources.
Biochemical literature

Direct biochemical information found in the primary literature usually contains the best quality data for use in reconstructing biochemical networks. Important details, such as precise reaction stoichiometry, in addition to its reversibility, are often directly available. Given that scrutinizing each study individually is an excessively time-consuming and tedious task, biochemical textbooks and review articles should be utilized, when available, with the primary literature used to resolve conflicts. Furthermore, many volumes devoted to individual organisms such as *E. coli* (48) and *Bacillus subtilis* (72) are available and are typically excellent resources.

High-throughput data

Genomic and proteomic data are useful sources of information for identifying relevant metabolic network components. In recent years, the complete genome sequence for over 250 species has been determined (55). Furthermore, extensive bioinformatics-based annotation efforts have made great strides toward identifying all coding regions contained within the sequence. For those biochemical reactions known to occur in an organism, but whose corresponding genes are unknown, sequence alignment tools such as BLAST and FASTA (12) can be utilized to assign putative function based on similarity to orthologous genes and proteins of known function. However, putative assignments are hypothetical and subject to revision upon direct biochemical characterization. The SEED (82) is another excellent resource that provides a platform to use chromosomal clustering combined with other genomic information to identify enzyme orthologs. Efforts are also underway to reconstruct networks based on annotated sequence information alone (35, 76).

The proteome of a biological system defines the full complement, localization, and abundance of proteins. Although these data are generally difficult to obtain, some organelle and bacterial data are available (33). Protein localization data are of particular importance in eukaryotic systems modeling (19), in which care must be taken to assign reactions to their appropriate sub-cellular compartment or organelle. Similarly, when modeling a system under a single condition, these protein abundance data are important in identifying active components.

Integrative biochemical database resources

In addition to the primary literature, genomic, transcriptomic, and proteomic data repositories can be accessed via the internet. Some popular resources are provided in Table 1. Of particular interest are resources that have incorporated these disparate data sources into metabolic pathway maps. The Kyoto Encyclopedia of Genes and Genomes (KEGG) (34) is perhaps the most extensive and well known among these resource types. Pathway maps for numerous metabolic processes are available. KEGG also provides information regarding orthologous genes for a variety of organisms, thus greatly enhancing the power of this resource. Additional organism-specific
database resources are also available. EcoCyc (36) incorporates gene and regulatory information, as well as enzyme reaction pathways particular to E. coli. MetaCyc (40) is an associated resource that details genetic, regulatory, and enzymatic information from the primary literature and computational prediction tools for a variety of organisms. The Comprehensive Yeast Genome Database (CYGD) (45) and Saccharomyces Genome Database (SGD) (13) are other examples of S. cerevisiae-specific comprehensive resources. SubtiList (46) and DBTBS (43) provide information about genome and transcriptional regulation for B. subtilis, respectively. Other useful biochemical databases are available elsewhere (5).

2.1.2. Metabolic reaction list generation

The next step in defining a constraint-based model requires clearly specifying the reactions to be included based on the metabolite and enzyme information collected in the previous step. A metabolic reaction can be viewed as substrate(s) conversion to product(s), typically by enzyme-mediated catalysis. Each reaction in a biochemical network must always adhere to the fundamental laws of physics and chemistry; therefore, reactions must be balanced in terms of charge and elemental composition (65).

Biological boundaries must also be considered when defining reaction lists. Metabolic networks comprised both intracellular and extracellular reactions. For example, the reactions of glycolysis and the tricarboxylic acid (TCA) cycle take place intracellularly in the cytosol. However, glucose must be transported into the cell via an extracellular reaction in which a glucose transporter takes up extracellular glucose. Furthermore, reactions must be compartmentalized properly when modeling eukaryotic cells given that certain metabolic reactions take place in the cytosol and others take place in various organelles (19). Finally, reaction reversibility must be defined, depending on the intracellular condition.
2.1.3. Gene-protein-reaction relationship determination

Upon completing the reaction list, the protein or protein complexes that catalyze each reaction must be determined (6,65). Each subunit from a protein complex must be assigned to the same reaction. Additionally, some reactions can be catalyzed by different enzymes. These isozymes must all be assigned to the same appropriate reaction. Biochemical textbooks often provide the general name of the responsible enzyme(s); however, the precise gene and associated gene product specific for the model organisms of interest must be identified. The database resources provided in Table 1 can assist in this process. In particular, KEGG provides considerable enzyme-reaction information for a variety of organisms.

2.2. Stoichiometric (S) matrix compilation and constraints addition

The compiled reaction list can be represented mathematically in the form of a stoichiometric (S) matrix (54). The S matrix is formed from the stoichiometric coefficients of the reactions that participate in a reaction network. It has m × n dimensions, where m is the number of metabolites and n is the number of reactions. Therefore, the S matrix is organized such that every column corresponds to a reaction, and every row corresponds to a metabolite. The S matrix completely describes the components of a metabolic network and the interactions between them.

Having developed a mathematical representation of a metabolic network, the next step is to identify and apply appropriate constraints to the network. Cells are subject to a variety of constraints from environmental, physicochemical, evolutionary, and regulatory sources. The S matrix is also a constraint in that it describes all possible metabolic reactions available to the cell. These stoichiometric constraints establish a geometric solution space that contains all possible metabolic behaviors. Additional constraints such as enzyme or transporter capacities also can be imposed on the model, further limiting the metabolic behavior solution space (64). Other types of constraints can be applied, including energy balance and transcriptional regulation. These constraints, however, are not as easily assembled and are difficult to apply to metabolic networks (30,37).

2.3. Linear optimization

The solution space defined by constraint-based models can be explored via linear optimization with an approach commonly referred to as flux balance analysis (22,37). The linear programming problem corresponding to the optimal flux distribution determination through a metabolic network can be formulated as follows:

Maximize (or minimize) \( Z = c^T \mathbf{v} \)

Subject to \( S \mathbf{v} = 0 \)

\( \alpha_i \leq v_i \leq \beta_i \) for all reactions \( i \)
In the above representation, Z represents the objective function, described in more detail below, and the latter statements impose steady state as well as lower and upper bound flux constraints for the metabolic network. Given that multiple possible flux distributions exist for any given network, linear optimization can identify a particular solution that maximizes or minimizes a defined objective function. Commonly used objective functions include cell growth rate (as defined by the weighted consumption of metabolites needed to make biomass); (21,23), ATP production; (63), and synthesis of a particular metabolite (77). In addition, by focusing only on the steady-state condition, assumptions regarding reaction kinetics are not needed, and it is possible to determine all chemically balanced metabolic routes through the metabolic network. The solution results in an optimal flux distribution (v) that specifies the optimal flux through the chosen objective function.

Problems of this type can be readily formulated and solved by commercial software packages, such as Matlab (The MathWorks, Inc., Natick, MA), Mathematica (Wolfram Research, Inc., Champaign, IL), LINDO (Lindo Systems, Inc., Chicago, IL), as well as tools available through the General Algebraic Modeling System (GAMS Development Co., Washington, DC). In recent years, the SimPhenyTM modeling platform (Genomatica, San Diego, CA) has been developed specifically to address the integrative data management and computational challenges inherent in building large-scale cellular models. This versatile platform provides network visualization (including explicit association logic between genes, proteins and reactions), database management, and various analytical tools (optimization, perturbation, flux variability, gene deletion analysis, phase plane analysis, network robustness, and pathway flux distributions, comparative modeling, etc.) that greatly facilitate the construction and study of genome-scale cellular models (60,20). Non-commercial program packages (free for academic users) such as FluxAnalyzer (39) and MetaFluxNet (41) are also available, and can be used for quantitative metabolic flux analysis.

The in silico simulation results can be experimentally verified and used to further strengthen the predictive capabilities of model. For example, in silico predictions with E. coli cell growth rate as an objective function are consistent with experimental data approximately 86% of the time (21), and could be further enhanced to approximately 91% with the addition of transcriptional regulatory constraints (14).

3. Current genome-scale in silico models

Several genome-scale reconstructions of metabolic networks have appeared, representing all three domains of life: archaea, bacteria, and eukarya. “Genome-scale” is used to describe these models because all of the metabolic reactions included in the model could be determined to take place in an organism based on genome annotation and organism-specific
biochemical literature. The number of genes, metabolites, and reactions included in each of these models is summarized in Table 2.

The largest and most mature reconstructions to date are for *S. cerevisiae* (19) and *E. coli* (65). In the case of *E. coli*, these reconstructions have been used to extensively study this model organism for over a decade (64). Based on the annotated genome sequence and biochemical data the *E. coli* metabolic model now accounts for 904 genes and catalyzes 931 internal metabolic reactions and transport processes operating on a network of 625 metabolites (65). Included in these genome-scale reconstructions are reactions and genes involved in glycolysis, the TCA cycle, the pentose phosphate pathway, electron transport chain, anaplerotic reactions, fermentative reactions, amino acid biosynthesis and degradation, nucleotide biosynthesis and interconversions, fatty acid biosynthesis and degradation, phospholipid biosynthesis, cofactor biosynthesis, and metabolite transport.

Several useful predictions have been obtained from such *in silico* models, including substrate preference; (23), minimal medium requirements; (69,71), essential metabolites and genes for cell growth; (16,33), optimal growth patterns (70), global metabolic flux patterns; (3), and outcomes of adaptive evolution and shifts in expression profiles. (32). Other organisms, including *B. subtilis*, *Pseudomonas aeruginosa*, *P. putida*, *Geobacter sulfurreducens* (71), and *Methanosarcina barkeri* have been fully reconstructed, but not yet published.

### 4. Large-scale mutant growth phenotype data

Genome-scale metabolic models have been shown to be very useful for large-scale growth phenotype analysis of single knock-out mutants under various growth conditions. Gene deletions (or equivalently, loss of gene product function) are simulated by constraining the flux through the associated reaction to zero, and the simulation results compared with experimental growth data from the known mutant. The *S. cerevisiae*
metabolic network was used to accurately predict or to explain 83% of 4154 mutant growth phenotypes in different media compositions (19). The E. coli stoichiometric network could predict the phenotype in 65% of 13,750 cases without considering transcriptional regulation (16). Results from such studies also enable the comparison of metabolic network robustness among microorganisms; for instance, Haemophilus influenzae seems to be less resistant to network perturbation than E. coli, as a higher percentage of central metabolic genes are predicted to be essential (54).

Although the predictions with metabolic models are generally quite good, there are also notable instances where they fail. Their failures are often due to the lack of regulatory events in the metabolic model. The metabolic models are subject to network topology constraints and flux balance analysis assumes their unfettered use to achieve optimal performance. However, cells use complex regulatory networks to achieve their goals that may or may not be consistent with assumptions of optimal performance (60). Therefore, metabolic models may yield incorrect predictions in situations where regulatory effects are a dominant influence on the behavior of the organism. Thus, there is a compelling need to include regulatory events within metabolic networks to broaden their scope and predictive capabilities, and initial progress toward this aim is being made (see below).

5. Transcriptomic data

Cellular networks have evolved over millions of years and, as a result, the cell has many interconnected pathways and elaborate transcriptional regulation networks that control which genes are expressed in response to various environmental and developmental signals. Transcriptional regulatory networks in microorganisms may have 200 or 300 transcriptional factors and of the order of 1000 defined links between components (53). For example, E. coli has been predicted to have 314 transcriptional factors and, based on the primary literature, 1468 regulatory interactions have been identified (30). The regulation of gene expression might lead to the repression of enzyme synthesis and, therefore, the effective removal of a reaction from the biological network. Regulatory constraints allow the cell to eliminate suboptimal phenotypic states and to confine itself to behaviors of increased fitness. Extensive efforts are being devoted to the elucidation of the components of transcriptional regulatory networks and the links between them. However, transcriptional regulatory network reconstruction in practice is not yet as well developed as metabolic network reconstruction (30,61).

To account for the effects of transcriptional regulation, a Boolean representation of the transcriptional regulatory network has been constructed for central and genome-scale metabolic models of E. coli, and verified with experimental data from biochemical literature and high-throughput phenotypic arrays (15,14). Within this framework, genes can only be found in two states, i.e. either expressed or not expressed. If the gene is not expressed, the enzyme will not be present in the cell and thus the
associated reaction is inactive \( (v_i = 0) \). Mathematically, this leads to a reduction in the solution space so that it contains only the solutions or phenotypes that the cell can choose from under a particular condition \( (63) \).

The combined regulatory and metabolic \textit{E. coli} model \( (16) \) accounts for 1010 genes, including 104 regulatory genes whose products together with other stimuli regulate the expression of 479 of 906 model genes. The integrated model improves the predictive capabilities of growth phenotypes for single gene deletions on various carbon and nitrogen sources (with 79\% accuracy compared with 65\% of 13750 cases when regulatory effects are ignored), and also predicts changes in gene expression in response to changes in environmental conditions. Akesson et al. \( (1) \) reported that the integration of transcriptomic data into stoichiometric \textit{S. cerevisiae} models resulted in improved predictions of metabolic behavior in batch cultures, enabling quantitative predictions of exchange fluxes as well as qualitative estimations of changes in intracellular fluxes. Effective promoter strength and metabolic resource requirements for genome expression and protein synthesis were also investigated by the integration of gene expression and mRNA half-life data in the context of an \textit{E. coli} model \( (2) \).

The predictions obtained from using Boolean logic regulation (“switch mechanism”) were in good agreement with experimental data. However, there are usually many ways in which a genome-scale network can attain optimal function, and many different transcription factors involved in the regulation of metabolic genes \( (64) \). Furthermore, more realistic simulations of signal transduction cascades often require representation by sigmoidal or other shaped functions \( (8) \). Therefore, reconstructed regulatory constraints should be viewed as testable hypotheses that ultimately require experimental verification.

6. Use of other high-throughput datasets

The availability of diverse “omics” datasets (transcriptomic, proteomic, metabolomic, fluxomic, and phenomic data) offers new challenges with genome-scale metabolic models. Although there is significant progress on transcriptional reconstruction and gene expression analysis with metabolic models, integrated analysis of other high-throughput data types is in the early stages of development. The next generation of constraint-based models is likely to account for the simultaneous reconciliation of multiple high-throughput data.

Recently, Raghunathan et al. \( (62) \) compared in vivo proteomic data with the in silico protein list associated with a genome-scale metabolic \textit{H. influenzae} model. Approximately 38\% of the proteins associated with the metabolic model were identified with high confidence, and approximately 90\% of the proteins in the model were experimentally identified as “candidate” proteins under microaerobic or anaerobic growth conditions. This study demonstrates that metabolic models can be used to predict the essentiality of proteins for cell growth and to address numerous unresolved
questions about intermediary metabolism of H. influenzae. In addition, a framework for incorporating metabolic flux data was developed and applied to the central metabolic network of E. coli.[80]

7. Biotechnological applications of genome-scale in silico models

To date, traditional metabolic pathway analysis (such as glycolysis, pentose phosphate pathway and TCA cycle) has served largely as conceptual frameworks for research and teaching. These pathways provide an important means of effective communication regarding the metabolism of various organisms, and will continue to be an important reference and basis for network reconstruction. Although useful, these traditional pathway definitions do not provide for quantitative, systemic evaluations of biological reaction networks because they focus on subsets of networks without consideration of network-wide interactions.[54]. In comparison, the constraint-based pathway analysis of genome-scale models can be used to assess all possible functional states (phenotypes) of reconstructed networks by enumerating all possible solutions defined by the stoichiometry[69], and can generate hypotheses about how to engineer the gene content of an organism for a desired phenotype.

Recently, bi-level optimization algorithms have been adopted to identify gene manipulation strategies for generating microbial strains that achieve specific metabolic engineering objectives.[58]. These metabolic engineering strategies use genome-scale metabolic models and a dual-level, nested optimization structure to predict which gene deletion(s) and/or addition(s) will lead to desired biochemical production while retaining viable growth characteristics. These techniques have established computational frameworks for the rational design of a microbial production system, and were applied in silico to design succinate, lactate, 1,3-propanediol,[10], amino acids,[58], and hydrogen[59] overproduction strains.

The combination of OptKnock strain design and adaptive evolution strategies was recently applied to genetically engineer strains of E. coli with enhanced lactic acid production capabilities.[25]. Adaptive evolution in the laboratory can be used to enhance the cellular growth characteristics by subjecting strains to specific selective pressures.[32,24]. In this case, strains were designed and constructed to optimize the coupled objectives of enhanced growth and lactate secretion rates. The engineered strains were then subjected to exponential growth phase serial passage for approximately 1000 generations to select the fastest growing variants which coordinately secrete the most lactate. This example demonstrates the power of the combined use of constraint-based modeling and analysis to guide strain design and evolutionary pressure to meet practical metabolic engineering goals.
The constraint-based modeling approach also has potential for medical applications, and can be used for identifying proteins that are essential for growth. In pathogenic microbial models, each identified essential gene could serve as a potential drug target for the development of novel, effective therapeutics in the future.

8. Perspective

Fig. 2 illustrates the iterative development of a constraint-based genome-scale model and some of the potential uses of in silico reconstructions. Genome-scale models provide a framework in which diverse high-throughput data types can be incorporated. Currently, significant reconstruction efforts are underway for two types of intercellular biochemical reaction networks—metabolic and transcriptional reactions. Ultimately, all biochemical events (metabolism, transcription, translation, regulation, signaling, and cell replication) must be integrated to generate a comprehensive cellular model of an organism, and to further broaden our capability to interpret and predict cellular phenotypes. This type of comprehensive model will allow for the simultaneous reconciliation of multiple high-throughput data (such as transcriptomic, proteomic, metabolomic, fluxomic, and phenomic data). In addition, the incorporation of additional constraints, such as osmolarity, electroneutrality, and thermodynamics, should be included to further reduce and more precisely define the allowable solution space, resulting in more accurate predictions. Potential biotechnological applications of genome-scale network reconstructions will become more apparent and practical as more models are developed and existing models enhanced.

Acknowledgements

Oh Y-K was supported by Korea Science and Engineering Foundation. We would like to thank Dr. Sang Yup Lee for inviting us to write this review of our work.
Reference

definition, characteristics, and capabilities. Proc Natl Acad Sci USA 97: 5528-33.
22. Fong SS et al. (2005) In silico design and adaptive evolution of Escherichia coli for production of lactic acid.
23. Frerer J et al. (2005) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network.
Genome Res 15: 244-51.
26. Guardia MJ et al. (2000) Cybemetic modeling and regulation of metabolic pathways in multiple steady states of
29. Ibarra RU et al. (2002) Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal
66. Reed JL and Palsson BO (2004) Genome-scale in silico models of E. coli have multiple equivalent phenotypic states.