Metabolic network modeling with model organisms
L Safak Yilmaz and Albertha JM Walhout

Flux balance analysis (FBA) with genome-scale metabolic network models (GSMNM) allows systems level predictions of metabolism in a variety of organisms. Different types of predictions with different accuracy levels can be made depending on the applied experimental constraints ranging from measurement of exchange fluxes to the integration of gene expression data. Metabolic network modeling with model organisms has pioneered method development in this field. In addition, model organism GSMNMs are useful for basic understanding of metabolism, and in the case of animal models, for the study of metabolic human diseases. Here, we discuss GSMNMs of most highly used model organisms with the emphasis on recent reconstructions.

Address
Program in Systems Biology, Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA, United States

Corresponding authors: Yilmaz, L Safak (lutuf.yilmaz@umassmed.edu), Walhout, Albertha JM (marian.walhout@umassmed.edu)

Introduction
A metabolic network is an organismal system that converts carbon and energy sources and electron donors and acceptors into biomass, energy, and by-products. Deficiencies in this system cause disease when biomass production or energy generation is impaired, or when toxic by-products accumulate. On the other side, engineering of a metabolic system can produce higher yields of biomass or valuable by-products. Thus, a mechanistic understanding of metabolism is crucial for various disciplines, from biomedical to biofuels research [1].

A commonly applied powerful method of metabolic analysis is constraint-based metabolic network modeling at the whole system level [2]. In this approach, all annotated metabolic genes in an organism are first matched to enzymes and then to reactions to obtain gene-protein-reaction associations (GPRs). These GPRs are used to reconstruct a genome-scale metabolic network model (GSMNM), which is then used to calculate the flux distribution over the entire network in any defined condition for the organism (see below). For model organisms, high-quality genomic annotations allow the reconstruction of comprehensive GSMNMs that can be parameterized and validated with publically available experimental datasets. In addition, since many properties of metabolic networks are conserved across taxa, model organism GSMNMs can be used to study human disease with animal models, while plant models can instruct agriculture.

Here, we first summarize the basics of genome-scale metabolic network modeling, and then explore GSMNMs and their applications in common model organisms [3] (Figure 1). GSMNMs mentioned are listed in Table 1, together with the latest human model [4] for comparison.

Mathematical modeling with genome scale metabolic networks
The steps of mathematical network modeling are summarized in Figure 2a. After GPRs are annotated, the reaction list is complemented by necessary transport reactions that carry metabolites between different compartments, exchange reactions that define the input and output of the system, and biomass reactions that represent growth. Next, this network is converted to a mathematical model that describes the mass balance of each metabolite as the difference between the fluxes of reactions that produce and consume it. The combination of mass balance equations for all metabolites yields a linear algebraic equation (Figure 2a) where a stoichiometry matrix (S) is multiplied by the reaction flux vector (v) to obtain the production rates of compounds. Because of the large number of reactions in GSMNMs, kinetic modeling approaches that allow the prediction of metabolite concentrations are not feasible. Instead, the mass balance is solved for an assumed steady state, that is, a zero sum of fluxes at each compound node, to obtain the flux vector, which describes the metabolic state as a flux distribution across the network. Importantly, the steady state assumption refers to the internal metabolism only, and is therefore not limited to the true steady state established for cells in a continuous flow reactor. It is also valid for any stable metabolic state, as in exponential growth, homeostasis, or growth in a small time interval that can be considered a quasi-steady state.

There is no single, unique solution to the governing mass balance equation. Instead there is a distribution of possible solutions (Figure 2b). To find a biologically
Model organisms reviewed. Percentage of genes in a human genome scale metabolic network model [4][4] that have orthologs to each model organism is shown based on Ref. [31]. This number is not available for E. coli. Bold letters indicate whether the metabolic network modeling with the indicated organism is relevant to basic understanding (B), human disease (D), industrial applications (I), or agricultural (A) applications.

Running FBA on a draft metabolic network model can reveal gaps that prevent biomass production or other reactions from carrying flux. These are to be iteratively corrected before validating the GSNNM with experimental observations (Figure 2a). A validated GSNNM can be used for different types and levels of predictions by varying the constraints used during FBA. In the simplest case, none of the fluxes are known or only uptake or secretions of main metabolites are experimentally meaningful flux distribution, fluxes are constrained by reaction reversibility and any known flux such as uptake rates of nutrients and secretion rates of by-products. In addition, an objective is set for maximizing or minimizing a subset of fluxes and optimized as a linear programming problem (Figure 2a). A typical objective function is maximization of biomass production to represent growth-oriented metabolism. This mathematical method is referred to as constraint-based flux balance analysis (FBA). FBA should not be confused with metabolic flux analysis (MFA), which uses isotope labeling and derives fluxes generally in core pathways by best fitting the flux distribution to the pattern of labeled metabolites in the central carbon metabolism only [5].

Table 1

<table>
<thead>
<tr>
<th>Organism</th>
<th>Year</th>
<th>Name</th>
<th>Reactionsb</th>
<th>Genes</th>
<th>Metabolitesb,c</th>
<th>Subcellular compartmentsd</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Homo sapiens</em></td>
<td>2016</td>
<td>Recon 2.2</td>
<td>7785</td>
<td>1675</td>
<td>2623 (5325)</td>
<td>c,m,n,x,r,g,l,i,e</td>
<td>[4][4]</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>2011</td>
<td>iJO1366</td>
<td>2583</td>
<td>1366</td>
<td>1136 (1805)</td>
<td>c,p,e</td>
<td>[14][14]</td>
</tr>
<tr>
<td></td>
<td>2014</td>
<td>EcoCyc–18.0–GEM</td>
<td>&gt;2289a</td>
<td>1445</td>
<td>1453 (nr)</td>
<td>c,p,e</td>
<td>[15][15]</td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>2003</td>
<td>iFF708</td>
<td>1379</td>
<td>708</td>
<td>555 (796)</td>
<td>c,m,e</td>
<td>[20][20]</td>
</tr>
<tr>
<td></td>
<td>2013</td>
<td>iTO977</td>
<td>1562</td>
<td>977</td>
<td>817 (1353)</td>
<td>c,m,x,e</td>
<td>[21][21]</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
<td>2016</td>
<td>iCEL1273</td>
<td>1985</td>
<td>1273</td>
<td>887 (1401)</td>
<td>c,m,e</td>
<td>[58][58]</td>
</tr>
<tr>
<td></td>
<td>2016</td>
<td>na</td>
<td>1922</td>
<td>979</td>
<td>901 (1846)</td>
<td>c,m,n,i,e</td>
<td>[57][57]</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>2008</td>
<td>na</td>
<td>196</td>
<td>211</td>
<td>151 (224)</td>
<td>c,m,e</td>
<td>[23][23]</td>
</tr>
<tr>
<td><em>Mus musculus</em></td>
<td>2005</td>
<td>na</td>
<td>1220</td>
<td>473</td>
<td>786 (872)</td>
<td>c,m,e</td>
<td>[27][27]</td>
</tr>
<tr>
<td></td>
<td>2008</td>
<td>na</td>
<td>2027</td>
<td>1399</td>
<td>1631 (2104)</td>
<td>c,m,e</td>
<td>[35][35]</td>
</tr>
<tr>
<td></td>
<td>2009</td>
<td>na</td>
<td>1344</td>
<td>nr</td>
<td>nr (1042)</td>
<td>c,m,e</td>
<td>[26][26]</td>
</tr>
<tr>
<td></td>
<td>2010</td>
<td>na</td>
<td>1494</td>
<td>724</td>
<td>945 (1285)</td>
<td>c,m,e</td>
<td>[25][25]</td>
</tr>
<tr>
<td></td>
<td>2010</td>
<td>iMM1415</td>
<td>3725</td>
<td>1415</td>
<td>1503 (2775)</td>
<td>c,m,n,x,r,g,l,e</td>
<td>[32][32]</td>
</tr>
<tr>
<td></td>
<td>2013</td>
<td>iSS1393</td>
<td>4091</td>
<td>1393</td>
<td>1536 (2950)</td>
<td>c,m,n,x,r,g,l,e</td>
<td>[33][33]</td>
</tr>
<tr>
<td></td>
<td>2015</td>
<td>na</td>
<td>2916</td>
<td>636</td>
<td>1097 (2072)</td>
<td>c,m,n,x,r,g,l,e</td>
<td>[34][34]</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>2009</td>
<td>na</td>
<td>1406</td>
<td>na</td>
<td>1253 (nr)</td>
<td>c,m</td>
<td>[43][43]</td>
</tr>
<tr>
<td></td>
<td>2010</td>
<td>AraGEM</td>
<td>1625</td>
<td>1419</td>
<td>1515 (1748)</td>
<td>c,m,pl,v,x,e</td>
<td>[46][46]</td>
</tr>
<tr>
<td></td>
<td>2011</td>
<td>iRS1597</td>
<td>1798</td>
<td>1597</td>
<td>1684 (1820)</td>
<td>c,m,pl,v,x,e</td>
<td>[47][47]</td>
</tr>
<tr>
<td></td>
<td>2012</td>
<td>na</td>
<td>1617</td>
<td>1791</td>
<td>1188 (nr)</td>
<td>c,m,pl,v,x,r,e</td>
<td>[49][49]</td>
</tr>
<tr>
<td></td>
<td>2013</td>
<td>na</td>
<td>2769</td>
<td>2857</td>
<td>2371 (2739)</td>
<td>c,m,pl,v,x,e</td>
<td>[44][44]</td>
</tr>
<tr>
<td></td>
<td>2014</td>
<td>na</td>
<td>549</td>
<td>627</td>
<td>236 (407)</td>
<td>c,m,pl,x,i</td>
<td>[50][50]</td>
</tr>
</tbody>
</table>

a na, not available; nr, not reported.
b Number of reactions and metabolites were obtained from both papers and model files (SBML or XLS), if available. When reported and model-based values did not match, model values were used.
c Number of unique metabolites followed by the total number of compartmentalized metabolites in parenthesis.
d Abbreviations: c, cytosol; m, mitochondria; n, nucleus; x, peroxisome; r, endoplasmic reticulum; g, golgi apparatus; l, lysosome; i, intermembrane space; e, extracellular space; pl, plastid; v, vacuole.

Since the model is not available, compartmentalization was inferred from the publication based on indirect information.
obtained. Then network properties can be analyzed using FBA and flux variability analysis (FVA) for hypothesis-driven discovery [2]. In more advanced applications, large experimental datasets are integrated with the GSMNM, such that, flux predictions are constrained by global gene expression [6] or metabolomics [7] to get an accurate picture of the metabolic state of the organism (Figure 2b).

Reconstructions and applications of model organism metabolic network models

*Escherichia coli*, the model bacterium

Applications of GSMNMs have been most widespread and successful with microorganisms, which have simple life styles accountable by FBA. This is because a bacterial culture in a bioreactor can be modeled as an open system that takes provided nutrients as the input and yields biomass and by-products as the output. The pioneering work on constraint-based metabolic network modeling in microorganisms was conducted with early *E. coli* reconstructions [8,9]. Later, *E. coli* GSMNMs served for the development of new methods such as dynamic FBA [10], parsimonious enzyme usage FBA (pFBA) [11], and regulatory FBA [12]. A review of most *E. coli* GSMNMs have recently been provided [13] and will not be repeated here. An example from the previous review [14] and the most recent reconstruction [15] are included in Table 1.

*Saccharomyces cerevisiae*, the budding yeast as a model eukaryote

There are more than two dozen GSMNMs reconstructed for the budding yeast *S. cerevisiae* to date [16]. This scale of effort reflects the fact that *S. cerevisiae* is the most studied unicellular eukaryote serving as a general eukaryotic model, and that it is directly used as a workhorse for the production of valuable metabolites such as ethanol [17]. In addition, *S. cerevisiae* can also be easily grown in controlled bioreactors and modeled as a unicellular open system, although, in contrast to bacteria, the metabolism is compartmentalized to organelles (Table 1). As with *E. coli*, GSMNMs of the yeast have already been reviewed [16,18,19] and we show only two representative models in Table 1 [20,21].
Drosophila melanogaster, the fly as a model animal
Unfortunately, a genome-scale metabolic network model is not yet available for the fruit fly D. melanogaster, although a core metabolic model to study the central metabolic properties of the muscle tissues of hypoxia resistant flies was published nearly a decade ago [22]. The model was reconstructed using metabolomics data obtained from dissected thoraxes, and by linking detected metabolites with GPR annotations. Later, the same model was expanded based on highly expressed enzymes in the thorax [23] (Table 1). These modeling efforts showed that FBA is a suitable method to study disease-related metabolic phenotypes in Drosophila. However, a genome-scale reconstruction will be needed to realize the full potential of metabolic network modeling with this organism.

Mus musculus, the model mammal
Several GSMMNs are available for the mouse (Table 1). These models are useful to study human disease, but also have industrial relevance since mouse hybridoma cell lines are employed for the production of biopharmaceuticals such as monoclonal antibodies (MAb) and vaccines [24]. Indeed, initial mouse GSMMNs targeted the modeling of hybridoma cell lines for a rational engineering approach to improve the production yields of MAb [25,26,27]. These models were tested using batch or continuous cultures of hybridoma cell lines. FBA-based prediction of growth rates and by-products, constrained by metabolomics measurements, were validated and improved over time. These models have further served as knowledge bases for other studies that used the provided biomass composition [28], maintenance energy [29], and other network properties [30].

Another lineage of mouse models was based on a human reconstruction [31]. Reactions in the human network associated with genes that had mouse orthologs were first extracted, which was followed by the arrangement of transport reactions and filling of the created gaps [32]. The final model, called iMM1415, included eight compartments (Table 1). In a later study [33], iMM1415 was not only updated but also extended to include intestine-specific transport and exchange reactions, and was combined with a Bacteroides reconstruction to develop a unified host-microbiota model. This interspecies model was useful in the analysis of synergistic and competitive interactions between the host and bacterial metabolism. The most recent mouse model was developed using a similar approach to iMM1415 reconstruction and was subsequently converted to a germ cell-specific model by integration of transcription levels in germ cells during spermatogenesis [34]. Using FBA constrained by gene expression levels throughout a spermatogenesis period, the authors determined metabolic genes and reactions that are critical for the germ cell differentiation for commitment to meiosis.

In addition to the above GSMMNs, an independent mouse model was semi-automatically reconstructed using the information stored in publicly available databases [35] (Table 1). Although only half of the reactions in this model are able to carry flux because of network gaps, it has been useful as a resource [36*,37].

Arabidopsis thaliana, the model plant
Flux analysis using FBA and MFA has been particularly popular in plant research, which is not surprising given the broad interest in engineering plant metabolism to improve yields of agriculturally and industrially valuable products [5,38,39]. A relatively large number of GSMMNs are available for the model plant A. thaliana (Table 1) [40*,41,42]. The first reconstruction was targeted at and validated by heterotrophic plant cells grown in suspension [43]. This model was later expanded by GPR annotations and compartmentalized into plant organs [44]. With these modifications the model was able to represent both heterotrophic and photoautotrophic metabolism, which was recently exploited to develop a model capable of simulating metabolism in day-night cycles with two separate compartments [45].

Another lineage of plant GSMMNs was started with the first photosynthetic A. thaliana model, AraGEM [46]. This reconstruction was updated in a study that compared A. thaliana and Zea mays metabolism [47] (Table 1). AraGEM was recently used to create a multi-tissue network that represents a whole plant [48**]. Metabolic networks for multiple plant tissues were derived from the generic model and combined in an organism framework. FBA was done by minimizing photon usage as the objective function for the entire network.

Two additional Arabidopsis GSMMNs have been developed. The most compartmentalized plant model [49] (Table 1) was used to derive the metabolic states of multiple plant organs using organ-specific protein expression datasets. The most recent reconstruction of A. thaliana focused on central metabolism (Table 1) [50]. This small but robust model of leaf cells consisted of manually curated GPRs, and was used to estimate the energetic cost of amino acid and enzyme synthesis during photoautotrophic growth conditions that employed carbon fixation by Rubisco, the most abundant protein in the world.

Caenorhabditis elegans, the nematode as a model animal
The nematode C. elegans (the worm) is a self-reproducing hermaphrodite with a relatively short life cycle. Although the worm has been widely used to study development, neurobiology and aging, it has recently also emerged as a powerful model for studying the effects of diet on metabolism and growth [51]. The laboratory diet of C. elegans typically consists of a pure bacterial culture. Different
bacterial species can be fed to the worm, and can be combined with easy genetic screening in both the animal and its diet [52–54]. In addition, C. elegans can be uniquely used to model some human diseases or specific components of the human diet. For instance, the two vitamin B12-dependent enzymes, methionine synthase and methylmalonyl-CoA mutase, are present in both humans and the nematode, but not in flies or yeast (Figure 3). We have recently discovered an alternative pathway to propionic acid breakdown that does not depend on vitamin B12, illustrating the power of this model [55].

The overall function of C. elegans metabolism can be seen as the conversion of bacterial biomass into worm biomass. Unlike any other animal model, C. elegans can be easily grown in liquid cultures as a dense population (hundreds of animals per ml) with both the bacterial and a chemically well-defined diet called axenic medium [56]. Thus, the nematode is a unique model animal in its suitability for metabolic network modeling and FBA, and a candidate for repeating the success of GSMMNs with microorganisms in an animal model. Two GSMMNs of C. elegans were recently published [57,58] (Table 1). Both models are able to represent growth with the bacterial and axenic diet and involve peculiar traits of C. elegans metabolism known so far, such as the presence of a glyoxylate shunt in an animal and the absence of a de novo NAD biosynthesis pathway. However, the focus of reconstruction and validation was different for each model.

The first model in Table 1, iCEL1273 [58], was shown to quantitatively account for the growth of C. elegans at pseudo-steady states in two different stages of life, the L4 growing larvae and egg-laying adults. Gene essentiality for growth at any life stage was systematically analyzed based on assumptions on animal physiology, such as the non-redundant functioning of paralogs in different tissues and the minimization of enzyme usage for an optimal metabolic state (implemented by pFBA). In addition, specific phenotypes were shown to be predictable, including the slow-down of growth in the absence of methionine synthase or the lack of vitamin B12 (Figure 3). The utility of this model was shown by analyzing the metabolic state of dormant dauer larvae in comparison to growing larvae. Integration of gene expression data was sufficient for iCEL1273 to predict the observed differences in these states such as low energetic activity, lack of growth, and dependence on stored carbon resources in the dauer state, but reverse in the other. iCEL1273 is available at a dedicated website named WormFlux (http://wormflux.umassmed.edu). This webtool is relatively unique because it shows the systematic annotation of all genes in the organism with a pipeline that uses multiple resources [58], and all metabolic reactions used in the reconstruction. Genes, enzymes, metabolites, and pathways are interlinked, searchable, and linked to other databases.

The second C. elegans GSMMN in Table 1 [57] was also quantitatively challenged, this time using measured amino acid concentrations in wild-type versus perturbed worms. The effect of knocking out heat-1, which codes an enzyme responsible for the initial step of branched chain amino acid breakdown, was simulated to indirectly deduce the changes in amino acid levels using FBA, and verified by experimental observations. The utility of this model was shown by focusing on aging. Integration of gene expression in a time series dataset from younger to older worms successfully predicted the decreasing overall activity of metabolism, and revealed the specific metabolic advantages of long-lived mutants such as the improved activity of the TCA-cycle and ubiquinone biosynthesis at later stages of life. The effective start of metabolic network modeling with the nematode opens up an exciting new field for studying the mechanistic relationships between diet, genotypes, phenotypes and aging at the systems level.

Conclusions and future perspectives
We derived four important conclusions from our review to provide perspective for future GSMMN reconstructions of model organisms and their applications. First, although most of the older GSMMNs in Table 1 are highly cited, the direct use of these models by other groups is rare. The lack of usage may be attributed to the required expertise to run GSMMNs. We believe this can be changed by building devoted web-based tools that allow user-friendly and interactive FBA with GSMMNs. WormFlux [58] has been developed for C. elegans with this purpose and efforts to implement FBA, FVA, and gene expression integration on this webtool are underway.

Second, the repeated use of GSMMNs for a particular organism generally involves a modification, or even a brand new reconstruction, as is clear from the timeline of mouse and Arabidopsis reconstructions (Table 1). We think that modeling modifications are often related to the feedback loops from validation and application stages in Figure 2a. Each time a model is challenged with new research questions or experimental data, potential reconstruction handicaps or false predictions lead to repairs, which improves not only the models but also our understanding of the metabolic network. It is reasonable to expect the same trend for other model organisms for which GSMMNs have recently been reconstructed.

Third, organism-level compartmentalization is needed for multi-cellular organisms and there are advancements towards this direction [59]. The multi-organ framework developed for A. thaliana [48] is the closest we have to a whole organism model. An additional, more focused multi-compartment model for the mouse has become
C. elegans is an adequate model to study vitamin B12 metabolism but S. cerevisiae and D. melanogaster are not. Left panel shows two vitamin B12-dependent pathways. C. elegans genes encoding the enzymes are indicated. Recently discovered propionate shunt is also drawn without the details. For the genes in bold font, phylogenetic protein sequence trees are provided on the right panel. Trees were obtained from WormFlux (http://wormflux.umassmed.edu) and edited for clarity. All reasonable best hits (B) and reciprocal best hits (R) (if available) from human (HSA), S. cerevisiae (SCE), and D. melanogaster (DME) were included in these trees [58]. Thus, S. cerevisiae (SCE) and D. melanogaster (DME) do not have the vitamin B12 enzyme homologs but they have an ortholog for methionine adenosyltransferase. Hits from A. thaliana (ATH) and C. elegans paralogs (P) are also shown. Identities of other organism genes are hidden except for the taxonomy.

Available to simultaneously represent metabolism in liver, muscle and adipose tissue cells [60], although it is important to note that this model was based on a human GSNMN [31]. We advocate the development of compartmentalized models for other organisms, notably C. elegans, because these model organisms are more suitable to high-throughput and large-scale genetic perturbations. Another adequate model system for tissue frameworks would be Drosophila, should GSNMNMs become available for this organism.

Finally, other compartmentalized modeling efforts need to consider interspecies metabolic interactions, as exemplified by the host-microbiota model in the mouse [61]. To this end, C. elegans, together with its bacterial diet, provide an ideal model system [51], which has already
been used with genetic screening approaches [54]. In the future, our understanding of metabolism will likely continue to be transformed by GSMMNs that characterize multi-cellular model organisms from cellular to whole-organism level and their communal interactions with other species.

Acknowledgement
This work was funded by the National Institutes of Health grants R21GM108045, DK068429, and GM082971.

References and recommended reading
Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest


5. The most recent version of the human GSMMN with many updates that improved the model functionality. For instance, the authors introduced the intramembrane space compartment by bypass thermodynamically unfavorable loops generating infinite ATP in the older models. This addition allowed the accurate calculation of ATP yields from various carbon sources.


8. Topfer N, Kleessen S, Nikoloski Z: Integration of metabolomics data into metabolic networks. Front Plant Sci 2015, 6:49. Regular FBA does not allow prediction of concentrations, and therefore, making sense of metabolomics data with GSMMNs is challenging. This review paper summarizes emerging methods to indirectly integrate metabolomics data with GSMMNs.


19. The authors compared four yeast models with respect to predictions of fluxes in the central metabolism. Inconsistencies and errors were corrected using proper constraints, mostly on reactions depending on NADPH and NADH. Notably, the early yeast model IFF708 was the best predictor of central metabolic fluxes without these corrections.


A good example of how GSNMNs of model organisms serve as resources. The central metabolic network of CHO cells were derived from a mouse GSNM and used for performing MFA. In addition, the authors introduced a method to apply MFA to a time series in a batch culture rather than being limited to steady state or mid-exponential growth.


Three Arabidopsis models are compared with respect to their predictive power and it is shown that experimental observations of efficient growth in A. thaliana strains with engineered photosynthetic systems were qualitatively captured by all models when proper constraints were used during FBA.


The authors present a multi-tissue plant model that not only represents leaf, stem, and root systems in a framework of GSNMNs, but is also compartmentalized further into two parts to simulate differential metabolism during day and night cycles.


