A review of methods for the reconstruction and analysis of integrated genome-scale models of metabolism and regulation

Fernando Cruz[1], José P. Faria[2], Miguel Rocha[1], Isabel Rocha[3], Oscar Dias[1*]

[1] Centre of Biological Engineering, University of Minho, Braga 4710-057, Portugal

[2] Data Science and Learning Division, Argonne National Laboratory, Argonne, IL, USA

[3] Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, 2780-157 Oeiras, Portugal

* - Correspondence: odias@ceb.uminho.pt

(3243/2500-3500)

Abstract (89/250)

The current survey aims to describe the main methodologies for extending the reconstruction and analysis of genome-scale metabolic models and phenotype simulation with Flux Balance Analysis mathematical frameworks, via the integration of Transcriptional Regulatory Networks and/or gene expression data. Although the surveyed methods are aimed at improving phenotype simulations obtained from these models, the perspective of reconstructing integrated genome-scale models of metabolism and gene expression for diverse prokaryotes is still an open challenge.

Introduction

High-throughput large-scale omics experiments are nowadays disseminated in biochemical research, supporting the study of the genomics, transcriptomics, and metabolomics layers of the cellular's molecular machinery. Currently, the volume of studies in the different omics fields has provided means for systems biology to thrive (1). This interdisciplinary field proposes differentiated approaches, such as the reconstruction of *in silico* networks and models, to provide quantitative and qualitative descriptions of biological systems as a whole.

Reconstruction of GSMMs

Nowadays, the generation of Genome-Scale Metabolic Models (GSMMs) is a common practice in systems biology. The reconstruction of these comprehensive models, through modeling techniques and genomics data, allows predicting cells' metabolic behavior (2–4).

A GSMM is an *in silico* representation of the biochemical reactions taking place within a the metabolism of a given organism (5). A genome-wide functional annotation that provides the required metabolic information over the organism of interest should be performed to assemble this representation. This information is linked to existing metabolic knowledge retrieved essentially from biochemical databases and literature. These steps help to create the reaction set, upon which the metabolic network is assembled.

The link from metabolic genes to proteins (mainly enzymes or membrane transporter proteins), as well as from proteins to reactions, is established by Gene-Protein-Reaction (GPR) associations. GPR associations must be cautiously defined during the reconstruction, taking into account isoenzymes, protein complexes and cascade reactions (3), through the use of *AND* or *OR* Boolean rules.

In the next iteration, biomass and organism-specific constraints are formulated from the retrieved knowledge to assemble a final stoichiometric model. The final GSMM may then be exported in a standard format, such as the Systems Biology Markup Language (SBML) (6). Several platforms, such as *merlin* (7), ModelSEED (8), RAVEN (9) and CarveMe (10), have been developed specifically for performing or assisting the reconstruction of these models (11).

The classic principles of chemical engineering are used to infer the dynamic mass balances of all metabolites in the metabolic network. A single ordinary differential equation (ODE) is created for each metabolite, accounting for its stoichiometry in the whole reaction set. Due to the lack of kinetic rates for all reactions in the ODE set, a steady-state approximation is used to reduce the mass balances to a set of linear equations. In a pseudo-steady-state paradigm, the concentration of a metabolite is assumed to remain constant throughout time (4).

When used to determine flux values, the set of linear equations defines a linear system, typically underdetermined, as the number of fluxes is much higher than the number of mass balance constraints, also referred to as the null space of S (12). Additional mass balance constraints can be added to the system to limit the flux that each reaction can acomodate by the imposition of both lower and upper bounds.

The system can be solved mathematically transforming it into an optimization problem,using several constraint-based approaches to predict the phenotypic behavior of the organism on a wide variety of environmental and genetic conditions. One of the most popular approaches is the Flux Balance Analysis (FBA) framework (12). FBA can compute an optimal solution, out of the feasible space determined by both mass balance and flux constraints using linear programming (LP). FBA requires the definition of an objective function, which should be relevant to the undergoing problem, which is commonly defined as the maximization or minimization of a specific metabolic flux (e.g., biomass reaction), and quantitatively determines how much each reaction contributes to a phenotype (4).

Parsimonious Flux Balance Analysis (pFBA) (13) and Flux Variability Analysis (FVA) (14), are alternative mathematical frameworks that also employ LP to allow analyzing *in silico* flux distributions. This set of tools is extremely helpful for validating a reconstructed model using experimental data of the organisms of interest. COnstraint-Based Reconstruction and Analysis (COBRA) Toolbox (15), COBRApy (16), OptFlux (17), and ReFramed (https://github.com/cdanielmachado/reframed) are prominent computational tools that have implemented these methods.

Although GSMMs have proven to be valuable throughout the years (18–24), there are limitations. Indeed, they are not yet capable of accounting for biological regulatory phenomena, such as the control of gene expression (25). The lack of this additional layer of information in these models can lead to erroneous *in silico* phenotype simulations, due to the lack of constraints that allow reaching the most accurate flux distribution according to experimental data.

Several methods have been proposed to improve phenotype simulations obtained from GSMMs, which will be herein surveyed. Most of these new methodologies are aimed at combining additional layers of omics data, namely transcriptomics, to limit the cone of allowable flux distributions. Also, these methods often resort to the integration of gene expression data and/or regulatory information obtained from Transcriptional Regulatory Networks (TRN)s being, therefore, prominent efforts made towards the reconstruction of integrated genome-scale models of metabolism and gene expression. The utilization of these integrated models can be useful to improve phenotype simulations or extend the analysis of regular GSMMs.

Reconstruction of TRNs

A TRN can be represented as a bipartite graph that comprehends vertices and edges. Vertices are usually regulators and target genes, whereas edges determine how these regulatory elements are connected and interact with each other, often under a causal relationship.

Inferring TRNs is fundamentally an underdetermined problem associated with a large search space where many solutions explain the data equally well (25,26). High-quality transcriptional information is scarce, in databases or literature and focused on a few well-studied organisms. Thus, the number of potential regulatory interactions between a Transcription Factor (TF) and target genes is considerably larger than the actual true biological interactions.

Although methods for reconstructing TRNs have been extensively reviewed in the literature (25–31), there are a panoply of classification systems and procedures. Moreover, as new methods are released each year, the complexity increases. Hence, assigning classes to these new approaches can be a complicated task. More importantly, this reveals that standard platforms and methodologies to assemble TRNs using diverse

sources of regulatory information, such as gene expression data (32,33) or transcriptional information (34–36), are still missing.

Nevertheless, an integrative workflow for reconstructing bacterial TRNs has been proposed by Faria *et al.* 2014 (25). The authors suggested that comparative-genomics approaches, namely the inference of TRNs using template curated networks and the prediction of cis-regulatory elements, can be integrated with the output of *de novo* reverse engineering tools. This workflow addresses the possibility of reconstructing TRNs for less described prokaryotic organisms using a variety of sources of regulatory information.

Template-network methodologies are based on the conservation of prokaryotic TRNs across evolution (37–39). As described by Faria *et al.* 2014 (25), template-network-based methods usually perform a search for orthologous genes in the genome of the organism of interest to propagate TRNs to strains of a well-characterized model organisms or closely related ones.

Cis-regulatory elements detection rely on the assumption that a regulatory interaction between a given TF and target gene can be inferred from the detection of the Transcription Factor Binding Sites (TFBS). The prediction of these *cis*-regulatory elements is a problem in which computational methods can assist (27,40). Although these computational tools are unable to infer a complete TRN from TFBS data, they can be integrated in the following workflow towards such a goal. The principles of this methodology were implemented by Alkema *et al.* 2004 (41) in Regulogger and the RegPredict web-based platform (42).

De novo reverse engineering tools are widely used for inferring TRNs from gene expression data. Indeed, a vast repertoire of computational tools based on the *de novo* reverse engineering approach can be found in the literature, and consequently numerous ways to classify these tools (28–30). Nevertheless, *de novo* reverse engineering methods are usually classified by mathematical formulation. Hence, to highlight the most used mathematical formulations, data-driven methods are usually based on the following:

- Correlation (e.g. COREGNET (43)),
- Information-theoretic (e.g. (44)),
- Boolean algebra (making use of the widely known binary operators *AND*, *OR,* and *NOT* to describe regulatory interactions (45), e.g. ModEnt (46)),
- Regression-based (e.g. GENIE3 (47)),
- ODEs (e.g. Inferelator (48)),
- Bayesian models (e.g. Gat-Viks *et al* 2007 (49)),

Available state-of-the-art TRNs' reconstructions for prokaryotic organisms include well-known prokaryotic organisms such as *Escherichia coli* (50,51), *Bacillus subtilis* (52,53) and *Mycobacterium tuberculosis* (54), having hundreds of regulators and thousands of target genes. These TRNs can, therefore, be used as gold-standards in the template-network-based approach or supervised methods. Interestingly, other TRNs for

prokaryotic organisms less described in the literature or having less amount of gene expression data (44,55–61), are also available, though in less number.

The TRNs available in the literature are usually the result of a specific gene expression data-driven analysis or the collection of regulatory information from literature and databases. Although these TRNs can be used in the comparative genomics or data-driven approaches, not all of them can be easily integrated in GSMMs, as only genome-scale TRNs are actually useful for the integration and simulation of integrated models.

Integrated models

Combining regulatory elements with information on metabolic stoichiometry is a complex task. There are many ways for controlling metabolism (62), which are well represented in the large diversity of methods proposed to quantify such influence (25,63–75). Nevertheless, the common denominator is that most methods start with GSMMs.

In detail, several of these methods integrate complete functional TRNs (76–84) or gene expression data (85–95) into GSMMs, whereas others impose additional constraints using information on allosteric and post-translational modifications (66,67,73). A different strategy is the combination of multiple layers of regulation (63,65,72,74). For higher eukaryotes such as humans, the control of gene expression also plays an essential role in the differentiation between different tissues or cell-types. Thus, algorithms for tailoring a GSMM according to a specific cell-line or tissue, commonly referred to as context-specific models, have been proposed (75,96–104). These principles and their main implementations are depicted in the Figure 1.

Figure 1: Overview of several methods for integrating additional constraints into GSMMs based on the regulation of metabolism. Whereas some methods integrate complete functional TRNs or gene expression data into GSMMs, others impose further constraints based on allosteric and post-translational modifications. Additionally, other methods integrate multiple omics layers of regulation of metabolism. For higher eukaryotes such as humans, context-specific models have also been based on tailoring the flux cone of solutions.

Surveying all approaches is out of the scope of this review. The following sections will cover the integration of TRNs or gene expression data into GSMMs, focusing on the control of gene expression at the transcriptional level. Figure 2 highlights both approaches, namely the integration of TRNs (Figure 2 A) and gene expression data (Figure 2 B) into GSMMs.

The differences between the integration of TRNs and gene expression data into GSMMs are associated with the type and amount of data that these sources can offer to the metabolic landscape of GSMMs.

Methods capable of integrating TRNs into GSMMs provide comprehensive knowledge regarding the metabolic and regulatory events occurring inside the cell at the genome-scale. As a result, both regulatory and metabolic networks can be analyzed together at the genome-scale, extending the range of applications of a regular GSMM.

On the other hand, gene expression data comprehend a set of snapshots of the transcriptome for several experimental conditions. Thus, a gene expression dataset can solely offer gene expression levels at a given experimental condition.

The group of methods aimed at integrating gene expression data with GSMM's comprises methods using only transcriptomics data for tailoring the flux distributions, so no structure or rules describing the regulatory interactions are observed in this class of methods. Thus, the integration of gene expression data focuses on improving the prediction of flux distributions, rather than the study and analysis of an additional biochemical network at the genome-scale.

Methods have also been classified according to the main formulations, as previously suggested by Machado *et al.* 2014 (70). Organizing methods into containers, according to their main formulations and features, facilitates the decision process when selecting an adequate method for the existing constraints and data sources.

Hence methods were classified into discrete (Figure 2 C) or continuous (Figure 2 D), according to whether phenotype simulations were performed with discrete, namely Boolean logic ("ON/OFF"), or continuous constraints.

Accordingly, a method is systematically classified as discrete if the result of the integration is a Boolean value (e.g., 1 for "ON" and 0 for "OFF"), imposing additional constraints on the system. These methods are also referred to as *switch,* since TRN or gene expression data switch reactions on or off. The state of a given metabolic gene is determined by evaluating either the Boolean regulatory rule or thresholding the gene expression level. Then, metabolic reactions mapped to metabolic genes are accessed according to the GPR rules to determine the resulting states. Thus, reactions having a one-to-one direct GPR rule are active/inactive according to the state of the metabolic gene. Reactions catalyzed by enzyme complexes, encoded by multiple yet mandatory genes, are considered inactive if at least one metabolic gene is not available. In contrast, reactions catalyzed by isoenzymes, namely multiple enzymes catalyzing the same reaction, are considered active if at least one metabolic gene is active.

Alternatively, there are methods aimed at circumventing the rigid Boolean logic, called *valve* methods, which impose continuous constraints to adjust a given flux distribution gradually and according to penalties, expression scores, or normalized expression levels obtained from the gene expression data. Typically, continuous integration is performed through the implementation of slack variables in the constraints' formulations, altering the reactions' bounds. The slack variable represents penalties, expression scores, or normalized expression levels retrieved from gene expression data for the metabolic genes associated with a given reaction. As before, the state of the metabolic reactions mapped to metabolic genes is assessed through GPR

rules, through selecting the best penalty, expression score, or normalized expression level for the slack variable.

The methodology for assigning a value to the slack variable, when a set of isozymes catalyzes a given reaction, comprises several distinct approaches. These include: methods in which the slack variable assumes the maximum expression score of the associated genes; methods where the slack variable takes the sum of expression scores of all genes encoding the isozymes catalyzing a single reaction; and, methods in which the reaction is replicated, according to the number of isozymes, and each new reaction is associated with one, and one only, gene.

Regarding reactions catalyzed by an enzyme complex, a group of methods establishes that the minimum expression score of all encoding genes is assigned to the slack variable. In contrast, other methods define the utilization of either the geometric or arithmetic mean of the expression score of all genes associated with an enzyme complex or isozymes, respectively.

Furthermore, methods capable of integrating gene expression data into GSMMs were also divided into single-condition (Figure 2 A) or multi-condition (Figure 2 B). Notice that this classification was not used to classify those methods aimed at integrating TRNs into GSM models, as it will be explained next.

Methods were classified as single-condition (Figure 1 A) or multi-condition (Figure 1 B) according to whether phenotype simulations were performed for one or more conditions/states in the gene expression dataset, respectively. For instance, a given method is considered multi-condition if it adjusts the flux cone of solutions by considering all conditions in the gene expression dataset or the gene differential expression between two conditions. Otherwise, the methods are classified as single-condition.

Notice that the latter classification was not used to classify those methods aimed at integrating TRNs into GSMMs. Methods capable of integrating TRNs into GSMMs do not require a gene expression dataset, thus classifying them into single- or multi-condition would be meaningless. Other methods capable of assembling and integrating TRNs into GSM models GSMMs often use the whole dataset and can then perform condition-specific phenotype simulations. Hence, classifying these methods as single-condition would be misleading.

Figure 2: Two examples of the integration of TRNs (A and C) or gene expression data (B and D) into GSMMs. The integration of TRNs (A) does not require gene expression data, while methods that integrate gene expression data (B) are capable of tailoring the flux cone of solutions by accounting for one (single-condition) or more (multi-condition) conditions in the gene expression dataset. Both types of integration can be mediated by discrete (C) or continuous (D) variables.

An analysis of these methods, encompassing the year of publication, availability of a tool with a user-friendly interface (namely a Graphical User Interface (GUI) without the requirement of coding skills), type of reaction constraint formulation, as well as the organism used for proof of concept has also been conducted. This information is available at the supplementary material 1. Figure 3 provides, on the other hand, a complete understanding of the methods described next, as well as their categorization according to the classification axes described above.

Integrating TRNs

For simulation purposes, the first attempts to integrate TRNs within GSMMs, namely Regulatory Flux Balance Analysis (rFBA) (76,105–107), Steady-state Regulatory Flux Balance (SR-FBA) (77) and the method proposed by Herrgård *et al.* 2006 (79), are based on the *switch* approach, to complement the metabolic system with additional constrains outlining which genes are activated or silenced in the network for specific *stimuli*.

As proof of concept, rFBA was successfully used to create the very first integrated genome-scale model of metabolism and gene expression for *E. coli* (106,107). In this reconstruction, as well as in the integrated network of *S. cerevisiae* provided by Herrgård *et al.* 2006 (79), a Boolean network collected from literature was integrated through a set of GPR rules with the GSMM imposing regulatory events as additional time-dependent constraints. On the other hand, SR-FBA performs steady-state simulations by including all valid metabolic and regulatory constraints in the system in a single step, through a Mixed-Integer Linear Programming (MILP) formulation. For that, nested Boolean expressions are formulated as a set of linear constraints, by recursively iterating over the structure of the regulatory layer and GPR rules, to add auxiliary variables representing intermediate Boolean terms (77). As shown in Figure 3, these methods have been classified as discrete, and none provides a user-friendly interface without the requirement of coding skills.

Two platforms, namely Toolbox for Integrating Genome-scale Metabolism (TIGER) (84) and FlexFlux (83), have been developed for integrating Boolean-based TRNs into GSMMs. TIGER can convert a series of logic Boolean rules, which can be thought of as a Boolean TRN, into a set of mixed-integer inequalities. Then, several algorithms for integrating gene expression data into the metabolic model and simulating phenotypic behavior can be implemented in the toolbox. Other implementations already available in this toolbox, such as Metabolic Adjustment by Differential Expression (MADE) (95), can be used for simulations.

FlexFlux differs from TIGER insofar as it is the only tool that provides a user-friendly interface for the integration of TRNs into GSMMs. This computational tool developed in Java® allows the input of Systems Biology Markup Language (SBML) (6) with the SBML Qualitative (SBML-qual) extension. SBML-qual is the standard file format extension for storing and sharing qualitative multi-state TRNs (108). In this way, a regular SBML file can hold a computer representation of qualitative models of biological networks.

Qualitative multi-state regulatory networks can then be used to determine multi-state qualitative constraints for metabolic flux analyses using FBA. Furthermore, FlexFlux allows the translation of the discrete qualitative states into continuous intervals, thereby constraining a reaction flux continuously or discretely (83).

Probabilistic Regulation of Metabolism (PROM) (78), PROM2.0 (109), and Integrated Deduced REgulation And Metabolism (IDREAM) (82) are all based on a probabilistic model for TRNs, which are integrated with a constraint-based model using a continuous method. PROM and PROM2.0 were the first attempts to circumvent the previous rigid discrete constraints added to a GSMM by setting the reactions' flux bounds proportional to the probabilities of their associated metabolic genes. In turn, the probability of a metabolic gene being activated in the whole set of conditions is defined together from the TRN and gene expression dataset provided as input. In short, PROM approaches can determine the probability of a given gene being or not activated, when the set of regulating TFs is either activated or silenced. The probability is calculated according to the frequency that each gene is active in the dataset (of either perturbed or over/under-expressed TFs). Likewise, the effect of perturbations on the regulatory network can also be robustly predicted.

Although PROM-based approaches are probably the best examples for integrating both TRNs and gene expression data into a GSMM, the gene expression dataset must have a large number of measurements per condition. PROM and PROM2.0 have been validated with *E. coli* and *M. tuberculosis* experimental gene expression data and the respective TRNs.

The IDREAM method resulted from the combination of Environment and Gene Regulatory Influence Network (EGRIN) (55,110) and PROM frameworks to create an enhanced genome-scale model of metabolism and gene expression for *Saccharomyces cerevisiae* (82). Contrariwise to the previous approaches, this methodology has used a *de novo* reverse engineering method called EGRIN to complement the yeast TRN collected from the database YEAst Search for Transcriptional Regulators And Consensus Tracking (YEASTRACT) (111). Then, the phenotype simulations are conducted in a similar way as in the PROM-based approaches.

Transcriptional Regulated Flux Balance Analysis (TRFBA) (81) and CoRegFlux (80) also provide a framework for the integration of gene expression data and TRNs in a continuous manner. Whereas the former requires a TRN for the organism of interest, the latter provides tools for inferring the regulatory network from gene expression data using CoRegNet (43). Nevertheless, CoRegFlux allows us to use a curated TRN rather than using the provided data-driven method.

Regarding the TRFBA methodology, this FBA-based approach considers gene expression levels as two additional types of continuous constraints. The first is represented by a constant parameter that converts the gene expression levels to the upper bounds of the reactions. The second type of linear constraints to be added to the

system can be thought of as the linear regression of each target gene from the regulating TFs.

CoRegFlux differs from TRFBA in that it uses a statistical reverse engineering method to infer targets of a given set of regulators at the genome-scale. Then, the influence score (similar to correlation scores for activation or repression) of each regulator in the set of target genes is calculated with CoRegNet from a large gene expression training dataset. Influence scores are used to train a linear model capable of predicting the gene expression of metabolic genes using a new gene expression dataset. These predicted levels of expression are then translated into flux bounds for the phenotype simulations using FBA or Dynamic Flux Balance Analysis (dFBA) (112).

All methods surveyed here are listed in the supplementary material 1.

Integrating gene expression data

The method proposed by Åkesson *et al.* 2004 (87), followed by MADE (95), were the earliest approaches for tailoring the flux cone of solutions using discrete variables obtained solely from gene expression data. In the case of the method developed by Åkesson *et al.* 2004 (87), a reaction is simply switched "off" with a zero flux bound if the associated genes are found to be under-expressed in the corresponding condition (single-condition method). MADE, on the other hand, tries to surpass the problem of arbitrary thresholding under-expression by considering multiple conditions (multi-condition method). Statistical significance between changes in gene expression levels across sequential conditions is calculated to infer whether a gene is activated (95).

E-Flux (113) and the method proposed by Lee *et al.* 2012 (94) have introduced several novelties when compared with the previous methodologies. These methods were the first attempts to constraint an FBA-based model using continuous variables. Nevertheless, these approaches are radically different. E-Flux directly maps gene expression levels into flux bound constraints, assuming the maximum flux of a given reaction to be a linear function of the expression of the associated genes in the same condition (single-condition method). Lee and coworkers (94) do not introduce or alter flux bound constraints directly into the GSMM. An alternative objective function that minimizes the distance between flux distributions and gene expression data is applied for each phenotype simulation (single-condition method).

The Transcriptional-controlled Flux Balance Analysis (TFBA) method, proposed by van Berlo *et al.* 2011 (93), is aimed at overcoming the problem of setting an arbitrary threshold to determine whether a gene is activated or not. The TFBA assumption is that differential gene expression between two conditions should also be reflected in the flux of the reactions associated with this gene. For that, the authors formulated constraints defining upper and lower limits for fluxes according to the gene expression, though assuming their transgression to be possible. The optimization problem (MILP

formulation) consists of finding the flux distribution that minimizes the number of transgressions.

Likewise, the method developed by Fang *et al.* 2012 (92) is based on the differential gene expression between two conditions, namely reference and perturbed conditions. This method assumes that the flux distribution of a reference condition can be determined using the FBA or FVA frameworks, while the differential gene expression between the reference and perturbed conditions is used for tailoring the flux distribution of the perturbed one. Also, this method considers the variation of the biomass composition between reference and perturbed conditions.

Similarly to TFBA and the method proposed by Fang *et al.* 2012 (92), the Gene Expression Flux Balance Analysis (GX-FBA) method (91) also determines the flux distribution for the reference condition using FBA. Then, GX-FBA employs a new objective function and new constraints derived from the difference between reference and perturbed states to perform the *in-silico* phenotype simulation of the latter state. A wide range of phenomena associated with temperature and known to induce virulence in the gram-negative bacterium *Yersinia pestis* was used as proof of concept.

Temporal Expression-based Analysis of Metabolism (TEAM) (90) and Adaptation of Metabolism (AdaM) (89) are the only methods developed for integrating time-series gene expression data into constraint-based models. The former uses dFBA (112) to predict time-series flux distributions based on temporal gene expression profiles. Using a cost minimization scheme similar to the strategy proposed in the context-specific Gene Inactivity Moderated by Metabolism and Expression (GIMME) method (98), TEAM is capable of determining the flux distribution of a GSMM, constrained with gene expression levels of each time step in the dataset. TEAM was tested with time-series gene expression data from *Shewanella oneidensis*.

AdaM consists of a flux-based bilevel optimization problem that extracts minimal operating networks from a given GSMM (89). This algorithm infers the minimal operating networks in agreement with the differential gene expression pattern between time-steps. Then, Elementary Flux Modes (EFM)s (114) are computed with these minimal operating networks rather than computing the flux distributions at each time step. Reactions are weighted according to the number of EFMs in which these are present. The optimization problem consists of finding the minimal network having the largest weight.

Angione *et al.* 2015 & 2016 (86,88) formulated methods, for example, the Metabolic and Transcriptomics Adaptation Estimator (METRADE), aimed at measuring the adaptability to a changing environmental condition over time. These approaches have provided equally valid methodologies for integrating gene expression data in metabolic networks. In short, these methods have modeled both upper and lower bounds of each reaction as a continuous logarithmic function of the associated gene expression levels.

Reaction Inclusion by Parsimony and Transcript Distribution (RIPTiDe) (85) is aimed at circumventing the assumption that reaction fluxes are directly related to the gene

expression levels for a given condition. The authors have proposed an unsupervised method that assigns weights (continuous variable) to reactions according to the normalized expression levels of associated genes over the entire dataset. Then, a pFBA simulation considering these linear coefficients is performed. The novelty of this method consists of its validation with precise transcript abundance obtained with RNA-sequencing (RNA-seq).

The methods capable of integrating gene expression data into GSMMs addressed herein are available in the supplementary material 1.

Synopsis

The reconstruction of GSMMs is common practice in systems biology nowadays. The advent of the GSMM reconstruction for many organisms was facilitated by the adoption of standard protocols (3), as well as the existence of user-friendly computational tools (7,8), capable of assembling these models from different genomic, enzymatic and stoichiometric data. Nevertheless, the simulation of GSMMs still presents today false-positive phenotypes for several environmental conditions.

The reconstruction of TRNs is a well-known strategy in systems biology for understanding the regulatory machinery of a given organism (26,28,30). Although there are many methodologies for assembling a TRN, standard protocols and computational platforms are yet missing to support the reconstruction of TRNs for less described organisms using different data sources. The workflow suggested by Faria *et al.* 2014 (25) highlighted several methodologies that can be combined to extend the reconstruction of TRNs to more bacterial species. To the best of our knowledge little progress has been made to provide a user-friendly platform capable of achieving such a goal. More importantly, the reconstruction of genome-scale TRNs using such integrative workflow, would be pivotal for the reconstruction and simulation of integrated models.

The integration of the control of gene expression into GSMMs has been surveyed in this work. A systematic classification that grasps the difference between the several methodologies, capable of integrating and simulating regulatory events into GSMMs was proposed herein. Although part of the reviewed methods have already been surveyed before (25,64,68–71,115), TIGER, FlexFlux, METRADE, IDREAM, TRFBA, CoRegFlux, RIPTiDe and the method proposed by Angione *et al.* (2016) have never been addressed elsewhere in reviews, to the best of our knowledge. Moreover, a detailed categorization that highlights the methodologies used to perform the integration of the regulatory layer into GSMMs has not been provided. This systematic categorization can guide the decision process of selecting the most adequate method of integration and simulation.

As shown in Figure 3, there are several methods and toolboxes capable of integrating and simulating TRNs into GSMMs using a discrete approach (76,77,79,83,84). The TRNs used by these methods and toolboxes were mainly reconstructed from literature, which

might be a time-consuming approach. The remaining methods allow to assemble TRNs from gene expression data using *de novo* reverse engineering methods. The resulting TRNs can be integrated and simulated with a given GSMM. FlexFlux is the prominent exception as it can perform the integration of the TRN in the GSMM using either discrete or continuous variables.

To date, only two prokaryotic organisms, *E. coli* (76–78,81,83) and *M. tuberculosis* (78,109), and the yeast *S. cerevisiae* (79–82,84), have integrated genome-scale models as a result of the integration of complete TRNs into a metabolic network. Nevertheless, some of these reconstructions still require gene expression datasets, namely several methods in the continuous sub-group.

Regarding the methods for integrating gene expression data, most of these have provided means for integrating transcriptomics data as continuous constraints from one or more conditions (Figure 3). Only the method proposed by Åkesson *et al.* 2004 (87), as well as MADE (95), use discrete variables to simulate integrated models of metabolism and gene expression.

Besides *E. coli*, *M. tuberculosis,* and *S. cerevisiae*, methods for integrating gene expression data have also provided integrated models for *S. oneidensis* (90) and *Y. pestis* (91).

Figure 3: Classification of methods aimed at the reconstruction of integrated genome-scale models of metabolism and gene expression. These methods have been divided according to the integration of TRNs (white boxes) or solely gene expression data into GSMMs. Discrete and continuous categories were used to classify these methods according to the usage of discrete, namely Boolean logic ("on/off"), or continuous constraints. Methods capable of integrating gene expression data into GSMMs have been further divided into single-condition (orange circles) and multi-condition (blue ellipses) whether phenotype simulations were performed for one or more conditions/states in the gene expression dataset, respectively. Each inner circle stands for a prokaryotic organism, while the outer circle stands for the baker's yeast *Saccharomyces cerevisiae*.

A vast diversity of methods for the integration of gene expression data in GSMMs has been found. Yet, most methods require large gene expression datasets to be robust, which might not be the case for all organisms. Other methods resort to mapping levels of gene expression directly with the reactions bounds, which again might not be the best approach (70,115,116).

Furthermore, the methods for integrating gene expression data with metabolic models previously evaluated by Machado and coworkers (70), namely E-Flux (113), MADE (95), GX-FBA (91) and the method developed by Lee *et al.* 2012 (94) have shown to perform poorly in the designed benchmark. None of the methods have outperformed

each other in the phenotype simulations nor pFBA, which indicates that the promising results reported by these methods seem to be mere artifacts related to rigid constraints created around the nature of the gene expression dataset.

The reconstruction of integrated models using TRNs is, in theory, more useful than merely integrating gene expression data into GSMMs. Integrated models that result from the integration of TRNs provide comprehensive knowledge regarding the metabolic and regulatory events happening inside the cell, thus leading to a broader range of applications when compared to a regular GSMM (117,118).

Moreover, the diversity of methods for reconstructing TRNs using different data sources, such as gene expression, transcription factor binding site, or comparative genomics analysis, eases the reconstruction of TRNs for most prokaryotic organisms having a sequenced genome (25). However, the absence of a user-friendly computational tool based on the ensemble of these different approaches is missing. In contrast, the same strategy has yielded results in the reconstruction of GSMMs (7–9,11,15,119).

In short, the existence of standardized protocols and easy to use computational tools for the generation of GSMMs has eased its practice in systems biology to study the metabolism of many organisms. In contrast, the absence of the computational tools that ease the reconstruction of TRNs from different sources of regulatory data hindered a similar approach.

The integration and analysis of regulatory events into GSMMs has been surveyed herein. A systematic classification has been created to grasp the difference between the several methodologies capable of integrating and simulating regulatory events into GSMMs. As a result, two primary approaches have been determined, namely the integration of TRNs and/or the integration of gene expression data.

The major obstacle when using the methods described in this survey to simulate integrated genome-scale models of metabolism and gene expression is not reproducing their results, but rather extending their implementations to other organisms and case studies. This hurdle poses a stiff challenge for using these methods out of the scope they were aimed at during development.

The requirement for large gene expression datasets with specific experimental conditions, the usage of TRNs reconstructed solely from literature, and the output of biased results strictly related to rigid constraints, are specific indicators of issues preventing the scaling-up of the reconstruction and analysis of integrated models. In short, there is a vast diversity of methods capable of integrating and simulating the effect of regulation into the metabolism, though few approaches that ease the reconstruction of these integrated models.

Hence, the perspective of reconstructing integrated genome-scale models of metabolism and gene expression for diverse prokaryotes is still a complex endeavor.

The implementation of a user-friendly computational framework that does not require coding skills, is capable of running a semi-automated pipeline for reconstructing TRNs or analyzing gene expression data, and performs its integration into standard GSMMs, would be a clear breakthrough towards the reconstruction and simulation of integrated genome-scale models of metabolism and gene expression. This hypothetical computational tool should be able to combine different sources of regulatory information which are seldom combined.

Perspectives

- The advent of high-throughput large-scale omics experiments has been supporting the study of the genomics, transcriptomics, and metabolomics layers of the cellular's molecular machinery. Systems biology can take advantage of the sheer volume of studies in these different omics fields by proposing differentiated approaches, such as the reconstruction of *in silico* networks and models
- The reconstruction and analysis of integrated models based on the integration of TRNs into GSMMs has not been conventional for non-model prokaryotic organisms. Usually, these lack large gene expression datasets, or have few sources of regulatory data. In addition to the absence of an established methodology and of easy to use tools and algorithms, the reconstruction and integration of TRNs into GSMMs is almost impracticable.
- Hence, a user-friendly computational framework that facilitates the reconstruction of TRNs and allows to integrate these into GSMMs would be a step towards facilitating the extension of integrated models to other prokaryotic organisms.

Acknowledgements

This study was supported by the Portuguese Foundation for Science and Technology (FCT) under the scope of the strategic funding of UIDB/04469/2020 unit and BioTecNorte operation (NORTE-01-0145-FEDER-000004) funded by the European Regional Development Fund under the scope of Norte2020 - Programa Operacional Regional do Norte. Fernando Cruz holds a doctoral fellowship (SFRH/BD/139198/2018) funded by the FCT. This study was supported by the European Commission through project SHIKIFACTORY100 - Modular cell factories for the production of 100 compounds from the shikimate pathway (Reference 814408). The submitted manuscript has been created by UChicago Argonne, LLC as Operator of Argonne National Laboratory ("Argonne") under Contract No. DE-AC02-06CH11357 with the U.S. Department of Energy. The U.S. Government retains for itself, and others acting on its behalf, a paid-up, nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government. The Department of Energy will provide

public access to these results of federally sponsored research in accordance with the DOE Public Access Plan.

Competing Interests

The Authors declare that there are no competing interests associated with the manuscript.

Contribution

FC and OD conceived and designed the study. FR assessed the methods and drafted the manuscript. OD managed its coordination and helped to draft the manuscript. JPF, MR and IR participated in the coordination of the study and helped to draft the manuscript. All authors read and approved the final manuscript.

Bibliography

1. Gray AN, Koo B-M, Shiver AL, Peters JM, Osadnik H, Gross CA. High-throughput bacterial functional genomics in the sequencing era. Curr Opin Microbiol. 2015 Oct 1;27:86–95.

2. Rocha I, Förster J, Nielsen J. Design and Application of Genome-Scale Reconstructed Metabolic Models. Methods Mol Biol vol 416 Microb Gene Essentiality. 2007;416:409–31.

3. Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc. 2010;5(1):93–121.

4. Dias O, Rocha I. Systems Biology in Fungi. In: Paterson R, editor. Molecular Biology of Food and Water Borne Mycotoxigenic and Mycotic Fungi. Boca Raton: CRC Press; 2015. p. 69–92.

5. Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ. Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol. 2009 Feb;7(2):129–43.

6. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics. 2003 Mar 1;19(4):524–31.

7. Dias O, Rocha M, Ferreira EC, Rocha I. Reconstructing genome-scale metabolic models with merlin. Nucleic Acids Res. 2015;43(8):3899–910.

8. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol. 2010 Sep 29;28(9):977–82.

9. Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, Nielsen J. The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for Penicillium chrysogenum. Maranas CD, editor. PLoS Comput Biol. 2013 Mar 21;9(3):e1002980.

10. Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Res. 2018 Sep 6;46(15):7542–53.

11. Faria JP, Rocha M, Rocha I, Henry CS. Methods for automated genome-scale metabolic model reconstruction. Vol. 46, Biochemical Society Transactions. Portland Press Ltd; 2018. p. 931–6.

12. Orth JD, Thiele I, Palsson BO. What is flux balance analysis? Nat Publ Gr. 2010;28(3):245–8.

13. Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, et al. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Mol Syst Biol. 2010 Jul 1;6(1):390.

14. Gudmundsson S, Thiele I. Computationally efficient flux variability analysis. BMC Bioinformatics. 2010 Dec 29;11(1):489.

15. Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nat Protoc. 2019 Mar 1;14(3):639–702.

16. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Syst Biol. 2013 Aug 8;7(1):74.

17. Vilaça P, Rocha I, Rocha M. A computational tool for the simulation and optimization of microbial strains accounting integrated metabolic/regulatory information. Biosystems. 2011 Mar 1;103(3):435–41.

18. Edwards JS, Palsson BO. Systems properties of the Haemophilus influenzae Rd metabolic genotype. J Biol Chem. 1999 Jun 18;274(25):17410–6.

19. Bro C, Regenberg B, Förster J, Nielsen J. In silico aided metabolic engineering of Saccharomyces cerevisiae for improved bioethanol production. Metab Eng. 2006 Mar;8(2):102–11.

20. Henry CS, Zinner JF, Cohoon MP, Stevens RL. iBsu1103: A new genome-scale metabolic model of Bacillus subtilis based on SEED annotations. Genome Biol. 2009 Jun 25;10(6).

21. Bordbar A, Palsson BO. Using the reconstructed genome-scale human metabolic network to study physiology and pathology. J Intern Med. 2012 Feb;271(2):131–41.

22. Dias O, Pereira R, Gombert AK, Ferreira EC, Rocha I. iOD907, the first genome-scale metabolic model for the milk yeast Kluyveromyces lactis. Biotechnol J. 2014;9(6):776–90.

23. Dias O, Saraiva J, Faria C, Ramirez M, Pinto F, Rocha I. IDS372, a phenotypically reconciled model for the metabolism of streptococcus pneumoniae strain R6. Front Microbiol. 2019;10(JUN).

24. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, et al. A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011. Mol Syst

Biol. 2011 Jan 11;7(1):535.

25. Faria JP, Overbeek R, Xia F, Rocha M, Rocha I, Henry CS. Genome-scale bacterial transcriptional regulatory networks: reconstruction and integrated analysis with metabolic models. Brief Bioinform. 2014 Jul 1;15(4):592–611.

26. De Smet R, Marchal K. Advantages and limitations of current network inference methods. Nature Reviews Microbiology. 2010.

27. Thompson D, Regev A, Roy S. Comparative Analysis of Gene Regulatory Networks: From Network Reconstruction to Evolution. Annu Rev Cell Dev Biol. 2015 Nov 13;31(1):399–428.

28. Barbosa S, Niebel B, Wolf S, Mauch K, Takors R. A guide to gene regulatory network inference for obtaining predictive solutions: Underlying assumptions and fundamental biological and data constraints. Biosystems. 2018 Dec 1;174:37–48.

29. Huynh-Thu VA, Sanguinetti G. Gene Regulatory Network Inference: An Introductory Survey. In: Methods in Molecular Biology. Humana Press Inc.; 2019. p. 1–23.

30. Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. Nat Rev Mol Cell Biol. 2008 Oct 17;9(10):770–80.

31. Rodionov DA. Comparative Genomic Reconstruction of Transcriptional Regulatory Networks in Bacteria. 2007;

32. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2012 Nov 26;41(D1):D991–5.

33. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress update—simplifying data submissions. Nucleic Acids Res. 2015 Jan 28;43(D1):D1113–6.

34. Novichkov PS, Brettin TS, Novichkova ES, Dehal PS, Arkin AP, Dubchak I, et al. RegPrecise web services interface: programmatic access to the transcriptional regulatory interactions in bacteria reconstructed by comparative genomics. Nucleic Acids Res. 2012 Jul;40(W1):W604–8.

35. Santos-Zavaleta A, Sánchez-Pérez M, Salgado H, Velázquez-Ramírez DA, Gama-Castro S, Tierrafría VH, et al. A unified resource for transcriptional regulation in Escherichia coli K-12 incorporating high-throughput-generated binding data into RegulonDB version 10.0. BMC Biol. 2018 Dec 16;16(1):91.

36. Sierro N, Makita Y, de Hoon M, Nakai K. DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. Nucleic Acids Res. 2008 Jan 1;36(suppl_1):D93–6.

37. Lozada-Chavez I, Janga SC, Collado-Vides J. Bacterial regulatory networks are extremely flexible in evolution. Nucleic Acids Res. 2006 Jul 19;34(12):3434–45.

38. Gelfand MS. Evolution of transcriptional regulatory networks in microbial

genomes. Vol. 16, Current Opinion in Structural Biology. 2006. p. 420–9.

39. Madan Babu M, Teichmann SA, Aravind L. Evolutionary Dynamics of Prokaryotic Transcriptional Regulatory Networks. J Mol Biol. 2006 Apr 28;358(2):614–33.

40. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, et al. Assessing computational tools for the discovery of transcription factor binding sites. Nature Biotechnology. 2005.

41. Alkema WBL, Lenhard B, Wasserman WW. Regulog analysis: Detection of conserved regulatory networks across bacteria: Application to Staphylococcus aureus. Genome Res. 2004;

42. Novichkov PS, Rodionov DA, Stavrovskaya ED, Novichkova ES, Kazakov AE, Gelfand MS, et al. RegPredict: An integrated system for regulon inference in prokaryotes by comparative genomics approach. Nucleic Acids Res. 2010;

43. Nicolle R, Radvanyi F, Elati M. COREGNET: reconstruction and integrated analysis of co-regulatory networks. Bioinformatics. 2015 Sep 15;31(18):3066–8.

44. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, et al. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. 2007;

45. Karlebach G, Shamir R. Constructing logical models of gene regulatory networks by integrating transcription factor-DNA interactions with expression data: An entropy-based approach. J Comput Biol. 2012 Jan 1;19(1):30–41.

46. Karlebach G, Shamir R. Constructing logical models of gene regulatory networks by integrating transcription factor-DNA interactions with expression data: An entropy-based approach. J Comput Biol. 2012 Jan 1;19(1):30–41.

47. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. PLoS One. 2010;

48. Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, et al. The inferelator: An algorithn for learning parsimonious regulatory networks from systems-biology data sets de novo. Genome Biol. 2006;

49. Gat-Viks I, Shamir R. Refinement and expansion of signaling pathways: The osmotic response network in yeast. Genome Res. 2007 Mar;17(3):358–67.

50. Fang X, Sastry A, Mih N, Kim D, Tan J, Yurkovich JT, et al. Global transcriptional regulatory network for Escherichia coli robustly connects gene expression to transcription factor activities. Proc Natl Acad Sci U S A. 2017 Sep 19;114(38):10286–91.

51. Gao Y, Yurkovich JT, Seo SW, Kabimoldayev I, Dräger A, Chen K, et al. Systematic discovery of uncharacterized transcription factors in Escherichia coli K-12 MG1655. Nucleic Acids Res. 2018 Aug 23;46(20):10682–96.

52. Faria JP, Overbeek R, Taylor RC, Conrad N, Vonstein V, Goelzer A, et al. Reconstruction of the Regulatory Network for Bacillus subtilis and Reconciliation with Gene Expression Data. Front Microbiol. 2016 Mar 18;7:275.

53. Arrieta-Ortiz ML, Hafemeister C, Bate AR, Chu T, Greenfield A, Shuster B, et al. An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network. Mol Syst Biol. 2015 Nov 17;11(11):839.

54. Turkarslan S, Peterson EJR, Rustad TR, Minch KJ, Reiss DJ, Morrison R, et al. A comprehensive map of genome-wide gene regulation in Mycobacterium tuberculosis. Sci Data. 2015 Mar 31;2:150010.

55. Bonneau R, Facciotti MT, Reiss DJ, Schmid AK, Pan M, Kaur A, et al. A Predictive Model for Transcriptional Control of Physiology in a Free Living Cell. Cell. 2007 Dec 28;131(7):1354–65.

56. Galán-Vásquez E, Luna B, Martínez-Antonio A. The Regulatory Network of Pseudomonas aeruginosa. Microb Inform Exp. 2011;1(1):3.

57. Marbach D, Costello JC, Küffner R, Vega NMN, Prill RJ, Camacho DM, et al. Wisdom of crowds for robust gene network inference. Nat Methods. 2012 Jul;9(8):796–804.

58. Fredrickson JK, Romine MF, Beliaev AS, Auchtung JM, Driscoll ME, Gardner TS, et al. Towards environmental systems biology of Shewanella. Vol. 6, Nature Reviews Microbiology. 2008. p. 592–603.

59. Rodionov DA, Rodionova IA, Li X, Ravcheev DA, Tarasova Y, Portnoy VA, et al. Transcriptional regulation of the carbohydrate utilization network in Thermotoga maritima. Front Microbiol. 2013;4(AUG).

60. de Jong A, Hansen ME, Kuipers OP, Kilstrup M, Kok J. The Transcriptional and Gene Regulatory Network of Lactococcus lactis MG1363 during Growth in Milk. PLoS One. 2013 Jan 17;8(1).

61. Schmitt WA, Raab RM, Stephanopoulos G. Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data. Genome Res. 2004 Aug;14(8):1654–63.

62. Nelson DL, Cox MM. Lehninger Principles of Biochemistry. 6th edit. W.H. Freeman; 2008. 1328 p.

63. Lerman JA, Hyduke DR, Latif H, Portnoy VA, Lewis NE, Orth JD, et al. In silico method for modelling metabolism and gene product expression at genome scale. Nat Commun. 2012 Jan 3;3(1):929.

64. Hao T, Wu D, Zhao L, Wang Q, Wang E, Sun J. The genome-scale integrated networks in microorganisms. Vol. 9, Frontiers in Microbiology. Frontiers Media S.A.; 2018.

65. Lloyd CJ, Ebrahim A, Yang L, King ZA, Catoiu E, O'Brien EJ, et al. COBRAme: A computational framework for genome-scale models of metabolism and gene expression. Darling AE, editor. PLOS Comput Biol. 2018 Jul 5;14(7):e1006302.

66. Brunk E, Chang RL, Xia J, Hefzi H, Yurkovich JT, Kim D, et al. Characterizing posttranslational modifications in prokaryotic metabolism using a multiscale workflow. Proc Natl Acad Sci. 2018 Oct 23;115(43):11096–101.

67. Gonçalves E, Sciacovelli M, Costa ASH, Tran MGB, Johnson TI, Machado D, et al. Post-translational regulation of metabolism in fumarate hydratase deficient cancer cells. Metab Eng. 2018 Jan 1;45:149–57.

68. Blazier AS, Papin JA. Integration of expression data in genome-scale metabolic network reconstructions. Front Physiol. 2012 Aug 6;3:299.

69. Kim J, Reed JL. Refining metabolic models and accounting for regulatory effects. Curr Opin Biotechnol. 2014 Mar;29C:34–8.

70. Machado D, Herrgård M. Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. Maranas CD, editor. PLoS Comput Biol. 2014 Apr 24;10(4):e1003580.

71. Imam S, Schäuble S, Brooks AN, Baliga NS, Price ND. Data-driven integration of genome-scale regulatory and metabolic network models. Front Microbiol. 2015;6:409.

72. O'Brien EJ, Palsson BO. Computing the functional proteome: recent progress and future prospects for genome-scale models. Curr Opin Biotechnol. 2015 Aug 1;34:125–34.

73. Machado D, Herrgård MJ, Rocha I. Modeling the Contribution of Allosteric Regulation for Flux Control in the Central Carbon Metabolism of E. coli. Front Bioeng Biotechnol. 2015 Oct 8;3:154.

74. O'Brien EJ, Utrilla J, Palsson BO. Quantification and Classification of E. coli Proteome Utilization and Unused Protein Costs across Environments. Maranas CD, editor. PLOS Comput Biol. 2016 Jun 28;12(6):e1004998.

75. Opdam S, Richelle A, Kellman B, Li S, Zielinski DC, Lewis NE. A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models. Cell Syst. 2017 Mar 22;4(3):318-329.e6.

76. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO. Integrating high-throughput and computational data elucidates bacterial networks. Nature. 2004 May 6;429(6987):92–6.

77. Shlomi T, Eisenberg Y, Sharan R, Ruppin E. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. Mol Syst Biol. 2007 Apr;3:101.

78. Chandrasekaran S, Price ND. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. Proc Natl Acad Sci U S A. 2010 Oct;107(41):17845–50.

79. Herrgård MJ, Lee B-S, Portnoy V, Palsson BØ. Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae. Genome Res. 2006 May;16(5):627–35.

80. Banos DT, Trébulle P, Elati M. Integrating transcriptional activity in genome-scale models of metabolism. BMC Syst Biol. 2017 Dec 21;11(S7):134.

81. Motamedian E, Mohammadi M, Shojaosadati SA, Heydari M. TRFBA: an

algorithm to integrate genome-scale metabolic and transcriptional regulatory networks with incorporation of expression data. Bioinformatics. 2017 Jan 8;33(7):btw772.

82. Wang Z, Danziger SA, Heavner BD, Ma S, Smith JJ, Li S, et al. Combining inferred regulatory and reconstructed metabolic networks enhances phenotype prediction in yeast. Nielsen J, editor. PLOS Comput Biol. 2017 May;13(5):e1005489.

83. Marmiesse L, Peyraud R, Cottret L. FlexFlux: combining metabolic flux and regulatory network analyses. BMC Syst Biol. 2015 Dec 15;9(1):93.

84. Jensen PA, Lutz KA, Papin JA. TIGER: Toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks. BMC Syst Biol. 2011 Sep 23;5(1):147.

85. Jenior ML, Moutinho TJ, Papin JA. Parsimonious transcript data integration improves context-specific predictions of bacterial metabolism in complex environments. bioRxiv. 2019;

86. Angione C, Conway M, Lió P. Multiplex methods provide effective integration of multi-omic data in genome-scale models. BMC Bioinformatics. 2016 Feb 2;17(S4):83.

87. Åkesson M, Förster J, Nielsen J. Integration of gene expression data into genome-scale metabolic models. Metab Eng. 2004 Oct 1;6(4):285–93.

88. Angione C, Lió P. Predictive analytics of environmental adaptability in multi-omic network models. Sci Rep. 2015 Oct 20;5.

89. Töpfer N, Jozefczuk S, Nikoloski Z. Integration of time-resolved transcriptomics data with flux-based methods reveals stress-induced metabolic adaptation in Escherichia coli. BMC Syst Biol. 2012 Nov 30;6.

90. Collins SB, Reznik E, Segrè D. Temporal Expression-based Analysis of Metabolism. PLoS Comput Biol. 2012 Nov;8(11).

91. Navid A, Almaas E. Genome-level transcription data of Yersinia pestis analyzed with a New metabolic constraint-based approach. BMC Syst Biol. 2012 Dec 6;6.

92. Fang X, Wallqvist A, Reifman J. Modeling Phenotypic Metabolic Adaptations of Mycobacterium tuberculosis H37Rv under Hypoxia. PLoS Comput Biol. 2012 Sep;8(9).

93. van Berlo RJP, de Ridder D, Daran J-M, Daran-Lapujade PAS, Teusink B, Reinders MJT. Predicting Metabolic Fluxes Using Gene Expression Differences As Constraints. IEEE/ACM Trans Comput Biol Bioinforma. 2011 Jan;8(1):206–16.

94. Lee D, Smallbone K, Dunn WB, Murabito E, Winder CL, Kell DB, et al. Improving metabolic flux predictions using absolute gene expression data. BMC Syst Biol. 2012 Jun 19;6(1):73.

95. Jensen PA, Papin JA. Functional integration of a metabolic network model and expression data without arbitrary thresholding. Bioinformatics. 2011 Feb

15;27(4):541–7.

96. Jerby L, Shlomi T, Ruppin E. Computational reconstruction of tissue-specific metabolic models: Application to human liver metabolism. Mol Syst Biol. 2010;6.

97. Shlomi T, Cabili MN, Herrgård MJ, Palsson B, Ruppin E. Network-based prediction of human tissue-specific metabolism. Vol. 26, Nature Biotechnology. 2008. p. 1003–10.

98. Becker SA, Palsson BO. Context-specific metabolic networks are consistent with experiments. PLoS Comput Biol. 2008 May;4(5).

99. Zur H, Ruppin E, Shlomi T. iMAT: An integrative metabolic analysis tool. Bioinformatics. 2010 Dec;26(24):3140–2.

100. Agren R, Bordel S, Mardinoglu A, Pornputtapong N, Nookaew I, Nielsen J. Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. PLoS Comput Biol. 2012 May;8(5).

101. Wang Y, Eddy JA, Price ND. Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. BMC Syst Biol. 2012 Dec 13;6.

102. Schmidt BJ, Ebrahim A, Metz TO, Adkins JN, Palsson BØ, Hyduke DR. GIM3E: condition-specific models of cellular metabolism developed from metabolomics and expression data. Bioinformatics. 2013 Nov 15;29(22):2900–8.

103. Rossell S, Huynen MA, Notebaart RA. Inferring Metabolic States in Uncharacterized Environments Using Gene-Expression Measurements. PLoS Comput Biol. 2013;9(3).

104. Schultz A, Qutub AA. Reconstruction of Tissue-Specific Metabolic Networks Using CORDA. PLoS Comput Biol. 2016 Mar 1;12(3).

105. Covert MW, Schilling CH, Palsson B. Regulation of Gene Expression in Flux Balance Models of Metabolism. J Theor Biol. 2001 Nov 7;213(1):73–88.

106. Covert MW, Palsson BO. Constraints-based models: Regulation of Gene Expression Reduces the Steady-state Solution Space. J Theor Biol. 2003 Apr 7;221(3):309–25.

107. Covert MW, Palsson BØ. Transcriptional regulation in constraints-based metabolic models of Escherichia coli. J Biol Chem. 2002 Aug 2;277(31):28058–64.

108. Chaouiya C, Bérenguier D, Keating SM, Naldi A, van Iersel MP, Rodriguez N, et al. SBML qualitative models: a model representation format and infrastructure to foster interactions between qualitative modelling formalisms and tools. BMC Syst Biol. 2013 Dec 10;7(1):135.

109. Ma S, Minch KJ, Rustad TR, Hobbs S, Zhou S-L, Sherman DR, et al. Integrated Modeling of Gene Regulatory and Metabolic Networks in Mycobacterium tuberculosis. Stelling J, editor. PLOS Comput Biol. 2015 Nov 30;11(11):e1004543.

110. Brooks AN, Reiss DJ, Allard A, Wu W, Salvanha DM, Plaisier CL, et al. A system-level model for the microbial regulatory genome. Mol Syst Biol. 2014 Jul;10(7):740.

111. Teixeira MC, Monteiro PT, Palma M, Costa C, Godinho CP, Pais P, et al. YEASTRACT: An upgraded database for the analysis of transcription regulatory networks in Saccharomyces cerevisiae. Nucleic Acids Res. 2018 Jan 1;46(D1):D348–53.

112. Mahadevan R, Edwards JS, Doyle FJ. Dynamic Flux Balance Analysis of diauxic growth in Escherichia coli. Biophys J. 2002;83(3):1331–40.

113. Colijn C, Brandes A, Zucker J, Lun DS, Weiner B, Farhat MR, et al. Interpreting Expression Data with Metabolic Flux Models: Predicting Mycobacterium tuberculosis Mycolic Acid Production. Papin JA, editor. PLoS Comput Biol. 2009 Aug 28;5(8):e1000489.

114. Schuster S, Fell DA, Dandekar T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nat Biotechnol. 2000;18(3):326–32.

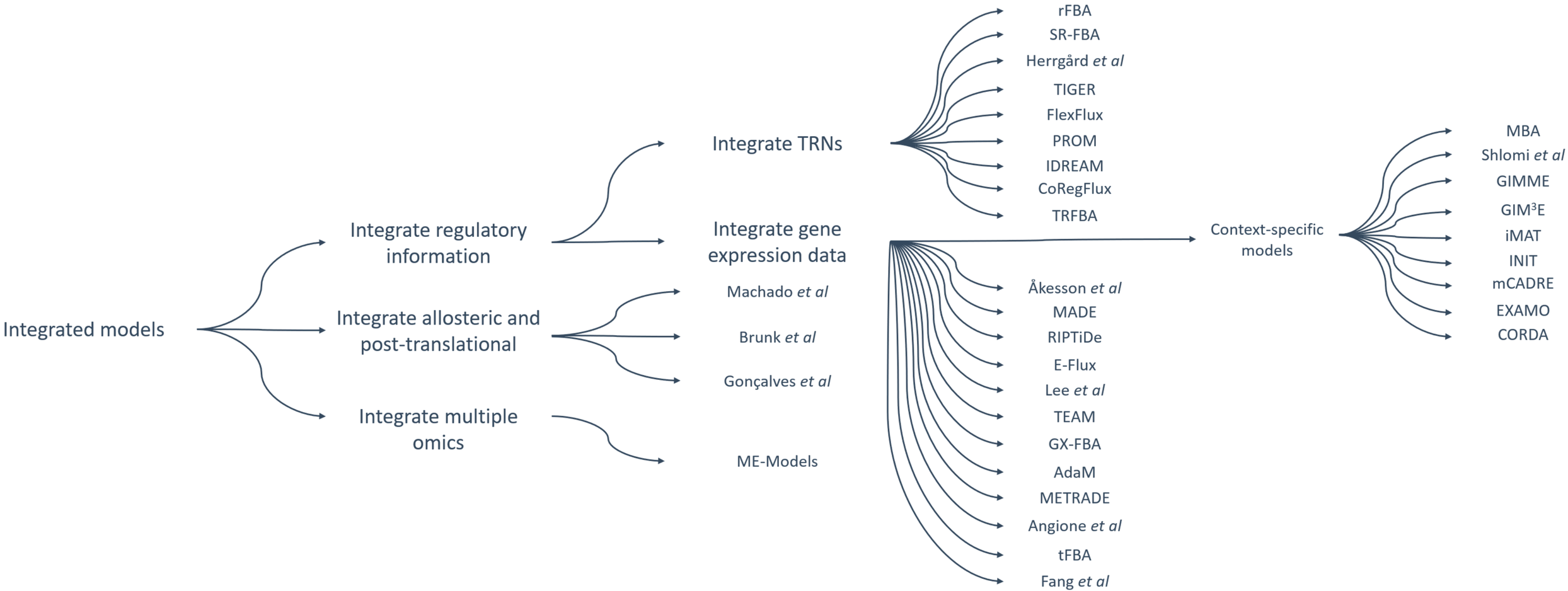
115. Vivek-Ananth RP, Samal A. Advances in the integration of transcriptional regulatory information into genome-scale metabolic models. Biosystems. 2016 Sep 1;147:1–10.

116. Kochanowski K, Sauer U, Chubukov V. Somewhat in control-the role of transcription in regulating microbial metabolic fluxes. Vol. 24, Current Opinion in Biotechnology. Curr Opin Biotechnol; 2013. p. 987–93.

117. Kim J, Reed JL. OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains. BMC Syst Biol. 2010 Apr;4(1):53.

118. Shen F, Sun R, Yao J, Li J, Liu Q, Price ND, et al. OptRAM: In-silico strain design via integrative regulatory-metabolic network modeling. Ouzounis CA, editor. PLOS Comput Biol. 2019 Mar 8;15(3):e1006835.

119. Monk J, Nogales J, Palsson BO. Optimizing genome-scale network reconstructions. Vol. 32, Nature Biotechnology. Nature Publishing Group; 2014. p. 447–52.


Integrated models
Integrate regulatory information
Integrate allosteric and post-translational
Integrate multiple omics
Integrate TRNs
Integrate gene expression data
Machado et al
Brunk et al
Gonçalves et al
ME-Models
rFBA
SR-FBA
Herrgård et al
TIGER
FlexFlux
PROM
IDREAM
CoRegFlux
TRFBA
Åkesson et al
MADE
RIPTiDe
E-Flux
Lee et al
TEAM
GX-FBA
AdaM
METRADE
Angione et al
tFBA
Fang et al
Context-specific models
MBA
Shlomi et al
GIMME
GIM3E
iMAT
INIT
mCADRE
EXAMO
CORDA

Regulatory Network

(A)

TF1

TF2

G1

G2

G3

Absolute

G1 and G2

G3

Discrete

Metabolic Network

Gene Expression Data

(B)

single-condition

multi-condition

G1

G2

G3

G1 and G2

G3

Continuous

Metabolic Network

Regulator
Under expressed

Regulator
Over expressed

Gene
Over expressed

Gene
Under expressed

Limited flux

Maximum flux

Discrete TRN

Continuous TRN

S. cerevisiae

E. coli

M. tuberculosis

S. oneidensis

Y. pestis

Herrgård *et al*

FlexFlux

CoRegFlux

rFBA

TRFBA

TIGER

IDREAM

SR-FBA

PROM

GX-FBA

RIPTiDe

E-Flux

TEAM

Lee *et al*

Åkesson *et al*

AdaM

Fang *et al*

METRADE

Angione *et al*

tFBA

MADE

Discrete gene expression data

Continuous gene expression data

Single-Condition

Multi-condition