Genome-scale bacterial transcriptional regulatory networks: reconstruction and integrated analysis with metabolic models

José P. Faria, Ross Overbeek, Fangfang Xia, Miguel Rocha, Isabel Rocha and Christopher S. Henry

Submitted: 9th July 2012; Received (in revised form): 9th October 2012

Abstract
Advances in sequencing technology are resulting in the rapid emergence of large numbers of complete genome sequences. High-throughput annotation and metabolic modeling of these genomes is now a reality. The high-throughput reconstruction and analysis of genome-scale transcriptional regulatory networks represent the next frontier in microbial bioinformatics. The fruition of this next frontier will depend on the integration of numerous data sources relating to mechanisms, components and behavior of the transcriptional regulatory machinery, as well as the integration of the regulatory machinery into genome-scale cellular models. Here, we review existing repositories for different types of transcriptional regulatory data, including expression data, transcription factor data and binding site locations and we explore how these data are being used for the reconstruction of new regulatory networks. From template network-based methods to de novo reverse engineering from expression data, we discuss how regulatory networks can be reconstructed and integrated with metabolic models to improve model predictions and performance. We also explore the impact these integrated models can have in simulating phenotypes, optimizing the production of compounds of interest or paving the way to a whole-cell model.

Keywords: genome-scale metabolic (GSM) model; transcriptional regulatory network (TRN); de novo reverse engineering; integrated metabolic and regulatory models

INTRODUCTION
Systems biology has provided numerous tools for modeling biological systems [1], many of which depend on the reconstruction of genome-scale metabolic (GSM) models. These models now exist for a growing number of organisms, including prokaryotic, archaeal and eukaryotic species [2]. With the advent of next-generation sequencing, the development of GSMs has become routine [2, 3] and many steps involved in the reconstruction and optimization...
of draft GSMs have been automated [4]. Algorithms and methods for GSM reconstruction have been reviewed in detail elsewhere [5–7].

However, nearly all existing GSMs fail to account for the impact of gene expression regulation on metabolic activity. In order to capture the impact of regulation on the behavior of an organism, a GSM must integrate some abstraction of regulatory mechanisms, which include the activity of RNA polymerase, transcription factors (TFs), promoters, TF binding sites (TFBSs) and sigma factors. Sigma factors allow the recognition of the enzyme by the promoter region, enabling transcription to begin. TFs bind to specific TFBSs in the promoter region and can act as activators, repressors or both (dual regulators). In eukaryotes, TFs are able to perform other tasks affecting regulation, such as chromatin-modifying activities [8]. Other elements have been identified as taking part in the control of transcription regulation in bacteria, such as riboswitches [9], RNA switches [10], antisense RNA [11] or microRNAs [12]. Here, we focus on regulation by TFs, a mechanism illustrated in Figure 1. Also displayed are some of the technologies, tools and resources necessary for reconstructing transcriptional regulatory networks (TRNs).

The integration of these regulatory mechanisms in GSMs requires methods for the reconstruction and analysis of TRNs. Once a regulatory model has been constructed for an organism, it can be integrated with GSMs to improve predictive accuracy and reveal new biological insights. For example, some cellular processes exhibit a dominance of regulatory mechanisms, affecting their behavior and leading to incorrect predictions when only metabolism is accounted for [13]. The first genome-scale integrated metabolic and regulatory model for *Escherichia coli* [14] revealed that regulation significantly affects growth phenotype predictions and these predictions improved with the addition of regulatory constraints. Simultaneously, the study of TRNs has unveiled novel interactions; in *Salmonella enterica*, 14 regulators were identified that affect the same genes leading to a systemic infection [15]. Similar studies led to the discovery of novel regulatory mechanisms in *Saccharomyces cerevisiae* [16].

Here, we review the reconstruction of TRNs and their integration with metabolic models. First, we explore the data available for TRN reconstruction, covering the most prominent databases of expression data and repositories of TF/TFBS data. Next, we examine how data availability triggered the development of a variety of TRN inference methods, including reverse engineering from expression data sets [17–21], network inference from TFBS site data [22–24] and knowledge-based template methods [25].

The integration of regulatory and metabolic networks for predictive modeling is possible only with the development of integrated phenotype simulation methods. The most widely used approach for simulating GSMs is flux balance analysis (FBA) [26]. To account for regulatory information, FBA was expanded with new methodologies, including regulatory FBA (rFBA) [13] and steady-state regulatory FBA (SR-FBA) [27]. We review these FBA-based methodologies, as well as other approaches that allow for a characterization of alternative

![Figure 1: Technologies, tools and resources for transcriptional regulatory network modeling and reconstruction.](https://academic.oup.com/bib/article-abstract/15/4/592/404002)
REGULATION DATA FOR TRN RECONSTRUCTION—FROM STANDARDS AND TECHNOLOGIES TO DATABASES

The development of microarray technologies gave rise to a revolution in biomedical research [31], also bringing new problems such as quality control of experiments [32] and selection of an appropriate level of detail [33]. To address these issues, the Functional Genomics Data Society launched a proposal to standardize the publishing and sharing of microarray data (MIAME) [34]. The majority of the community adopted the proposal, requiring authors to follow the MIAME guidelines. Publishers also required authors to store data [35] in either NCBI's Gene Expression Omnibus (GEO) [36] or EBI's ArrayExpress [37], the major public gene expression data repositories, both MIAME compliant.

These databases integrate data from a variety of technologies that can help determine regulatory interactions, although expression profiling and genome binding and occupancy studies have become the most prevalent. Expression profiling techniques vary from the traditional array oligonucleotide hybridization technology for measuring gene expression level to mRNA quantification methodologies such as serial analysis of gene expression (SAGE) [38, 39] or reverse transcriptase PCR (RT-PCR). Genome binding and occupancy experiments have the advantage of identifying the spots corresponding to DNA–protein binding targets. Chromatin immunoprecipitation with array hybridization (ChIP-chip) [40, 41] is used to overcome limitations of common expression profiling. Other ChIP technologies have also been developed in combination with different expression techniques such as SAGE (ChIP-SAGE [42]) to achieve a particular level of detail, depending on the organism and tissue studied [43]. With the development of next-generation sequencing technologies, ChIP-Seq [44] and RNA-Seq emerged [45, 46]. ChIP-Seq enables whole-genome ChIP assays, while RNA-Seq provides a capacity for direct measurement of mRNA, small RNA and non-coding RNA abundances [47]. ChIP methods have been widely used to collect expression data from *E. coli* [48–50] and, more recently, RNA-Seq methods have been adjusted for studying bacterial transcriptomes [51, 52]. RNA-Seq has been also successfully used to detect transcription start sites [53] that can be used for regulon inference.

Data available for TRN inference can be categorized into two major groups: (i) databases of gene expression data (including genome binding experimental data) and (ii) databases of TF and TFBS. Table 1 shows the most notable databases of the former group.

We surveyed GEO, as the major expression database, gathering statistics on the type of studies conducted, availability of data, quantification of bacterial data and the most represented microbes (Figure 2). These statistics clearly indicate that most of the current data are from expression profiling, with 18 498 experimental series (85%). Although next-generation sequencing technologies were introduced recently [57], we can already see a change in the types of experiments being performed (Figure 2B). Examining the organisms for which expression data are available, we find that only 7% of data sets are from bacteria (Figure 2C), with *E. coli* being the most represented prokaryote (Figure 2D).

Table 1 also includes other notable databases, from which we highlight the Many Microbe Microarrays database [54] currently holding ~2000 microarrays for *E. coli*, *S. cerevisiae* and *Shewanella oneidensis*. The data available are all from Affymetrix single channel microarrays, allowing a uniform normalization procedure and higher quality data. The *E. coli* data have already been applied for TRN inference [60].

Figure 3 shows the discrepancy between the number of sequenced genomes and the number of genomes for which any type of expression data exists. In this study, we cluster bacterial genomes available in the PubSEED [61] (a large repository of genomes and annotations) at the taxonomical level of family. The set of 20 bacterial families associated with expression data in GEO are shown in the phylogenetic tree. On an average, 16.2% of the 3493 PubSEED genomes that fall into these families have expression data linked to them. Expression data are available for 55% of the genomes in the Gammaproteobacteria family, demonstrating the extensive amount of data available for this taxonomic clade. In contrast, more than half of the bacterial phyla have expression data for <10% of their species, revealing that numerous phylogenetically distinct clusters of microbes have little gene expression experimentally characterized.
Repositories with regulatory interactions also hold valuable information. Table 2 shows the most comprehensive resources available for prokaryotes. Organism-specific databases are available for well-known organisms such as E. coli, Bacillus subtilis and Mycobacterium tuberculosis, including a comprehensive collection of regulatory information. Among those, RegulonDB is the most comprehensive resource for regulatory interactions data of any single organism (E. coli). In its latest release, genetic sensory response units are introduced to better represent the biology of gene regulation [64], trying to capture all the

![Figure 2: Survey of the GEO database.](https://academic.oup.com/bib/article-abstract/15/4/592/404002/595)

**Table 1:** Gene expression repositories with bacterial transcriptional data

<table>
<thead>
<tr>
<th>Database</th>
<th>Main features</th>
</tr>
</thead>
<tbody>
<tr>
<td>ArrayExpress [37]</td>
<td>EBI's database for expression data. Data submitted by users and imported from GEO. Advanced queries and ontology-driven searches.</td>
</tr>
<tr>
<td>M3D [54]</td>
<td>Data uniformly normalized from Affymetrix microarrays for Escherichia coli, Saccharomyces cerevisiae and Shewanella oneidensis.</td>
</tr>
<tr>
<td>SMD [55]</td>
<td>Partially public database with data from ~60 organisms. Escherichia coli, Mycobacterium tuberculosis and Streptomyces coelicolor are among the most represented microbes. Data analysis framework embedded.</td>
</tr>
<tr>
<td>COLOMBOS [56]</td>
<td>Cross-platform expression compendia for E. coli, B. subtilis and S. enterica subspecies serovar Typhimurium. Provides tools for expression analysis and extraction of relevant information.</td>
</tr>
</tbody>
</table>
**Figure 3:** Comparison of bacterial genomes with expression data in GEO versus genomes with complete DNA sequences in the PubSEED [61]. The 20 bacterial families that contain genomes with expression data in GEO are arranged in a topological tree. For each family, the most abundantly sampled species in the PubSEED was picked to represent that family and the alignment of their 16S sequences was used to reconstruct the bacterial family tree. The color coding of the tree nodes denotes the phyla they belong to. Most phyla contain only one family, with the exception of Cyanobacteria (three families), Bacteroidetes (four families) and Firmicutes (three families). The last two phyla are especially overrepresented in terms of both sequenced genomes and expression data. The numbers on the right of each tree node denote the number of genomes with GEO expression data (566 in total) and the number of genomes present in the PubSEED (3493 in total). Archea organisms were removed from this study since we aim to survey only bacterial genomes. In the horizontal bar plot, we show the fraction of each bacterial family for which expression data are available. The tree was designed with the Interactive Tree of Life Tool [62, 63].

**Table 2:** Databases with notable bacterial transcriptional data

<table>
<thead>
<tr>
<th>Database</th>
<th>Organism(s)</th>
<th>Main features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Organism specific</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DBTBS [66]</td>
<td>B. subtilis</td>
<td>Compendium of regulatory data with promoters, TFs, TFBS, motifs and regulated operons.</td>
</tr>
<tr>
<td>RegulonDB [64]</td>
<td>E. coli</td>
<td>Compendium of regulatory data, promoters, TFs, TFBS, transcription units, operons and regulatory network interactions.</td>
</tr>
<tr>
<td>EcoCyc [65]</td>
<td>E. coli</td>
<td>Comprehensive database with gene products, transcriptional, post-transcriptional data and operon organization.</td>
</tr>
<tr>
<td>DPInteract [67]</td>
<td>E. coli</td>
<td>DNA binding proteins and binding site data.</td>
</tr>
<tr>
<td>MTBRegList [68]</td>
<td>M. tuberculosis</td>
<td>TFBS and regulatory motifs.</td>
</tr>
<tr>
<td><strong>Organism class/family</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cTFbase [70]</td>
<td>Cyanobacteria</td>
<td>Putative TFs.</td>
</tr>
<tr>
<td>TractorDB [71]</td>
<td>Gamma-proteobacteria</td>
<td>TFBS predictions.</td>
</tr>
<tr>
<td><strong>Non-organism specific</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ExtraTrain [73]</td>
<td>Bacteria and Archea</td>
<td>Transcriptional data and extragenic regions</td>
</tr>
<tr>
<td>DBD [74]</td>
<td></td>
<td>TF predictions.</td>
</tr>
<tr>
<td>RegTransBase [75]</td>
<td></td>
<td>Regulatory interactions from literature and TFBS.</td>
</tr>
<tr>
<td>PRODORIC [76]</td>
<td>Bacteria</td>
<td>TFs, TFBSs, regulon lists, promoters, expression profiles.</td>
</tr>
<tr>
<td>sRNAMap [77]</td>
<td></td>
<td>Small non-coding RNAs and regulators.</td>
</tr>
<tr>
<td>ODB [78]</td>
<td></td>
<td>Known and putative operons.</td>
</tr>
<tr>
<td>RegPrecise [79]</td>
<td></td>
<td>Regulon database.</td>
</tr>
</tbody>
</table>
phenomena involved in regulation, from the initial signal to gene response. Another major resource for *E. coli* data is EcoCyc [65], integrating RegulonDB and curated data from >21 000 publications and TRN descriptions that include genes, ligands and regulators with their targets. DBTBS [66] is the major resource for *B. subtilis* regulatory data.

Less-comprehensive databases present fewer types of different regulatory information (sometimes only TFBS predictions or TF information) but cover a wide range of bacteria (Table 2). Notable examples are ODB [78], which stores known operon data for ~10 000 operons in 56 organisms and putative operons for >1000 genomes; RegTransBase [75], which collects regulatory data from the literature and RegPrecise [79], a repository of manually curated regulons that provides tools for regulon propagation.

Reconstruction of TRNs can use different types of data and the accurate selection of data/database(s) for the method of choice is paramount in the reconstruction process. Organism-specific databases are particularly useful for reverse engineering methodologies as training data sets and essential for validation. Methodologies based on comparative genomics approaches make good use of less comprehensive databases but cover a wider range of organisms.

**TRN RECONSTRUCTION—FROM TEMPLATE NETWORKS AND INFEERENCE ALGORITHMS TO INTEGRATION WITH GSMS**

TRN reconstruction aims to make sense of gene expression and binding site data by revealing the interactions between the different elements of the cell’s regulatory machinery. Different methodologies have been proposed for TRN inference. However, there is no consensus for classification in the literature. Some reviews classify methods as bottom–up and top–down [80], others focus on inference from a specific type of data such as gene expression [81], while others present methods and computational tools [82]. Here, we review and categorize different methodologies within two major types: genomics-driven and data-driven. The first uses comparative genomics approaches, while the second refers to *de novo* reverse engineering from expression data.

**TFBSS DATA-BASED METHODS VIA PREDICTION OF CIS-REGULATORY ELEMENTS**

TRN reconstruction from binding site data can also be defined as a comparative genomics approach. Prior to the development of the first binding-site approaches, most methods relied almost entirely on functional information from expression data [19, 89]. The GRAM (genetic regulatory models) algorithm [90] was the first to combine the use of expression data and binding site data in a genomewide inference
process, enabling the inclusion of information about physical interactions between regulatory genes and their targets. Other work focused on the conservation of the regulatory machinery across different organisms.

Regulogger [91] was introduced to generate ‘regulogs’ or sets of genes that are co-regulated and have their regulation processes conserved across several organisms. Using Staphylococcus aureus, regulogs were produced for well-known sets of genes and provide clues about the functions of unannotated genes. Studies of δ-proteobacteria [23] revealed that very diverse species of proteobacteria have similar regulatory mechanisms.

The principles behind this methodology were reviewed by Rodionov [92]. Figure 4B describes one of the two strategies proposed. The first step is to gather all available information related to TFs and TFBSs in a selected model organism. These data are then used as a training set for the TFBS model. The accuracy of the methodology is closely connected to the quality and quantity of sequences used for training. E. coli is usually used as a model species for Gram-negative bacteria and B. subtilis for Gram-positive bacteria. If the TFBSs corresponding to a particular TF are unknown, all genes regulated by the TF in the model species are identified and then orthologs for these genes in closely related genomes are found. With a TFBS training set built by this process or experimentally determined (see Table 2), positional weight matrices (PWMs) are constructed for the collection of binding sites. Several algorithms are available that perform motif pattern recognition [93] to construct PWMs. One of the first algorithms developed for this task was AlignACE [94]. This algorithm was recently upgraded to W-AlignACE [95] incorporating a new learning approach [96] and showing increased accuracy in obtaining PWMs for gene sequences, gene expression data and ChIP-chip data [95].

Using the PWMs, one can perform a genomewide search for putative TFBSs on the target genomes. This comparative-genomics-based approach requires a high-quality training set; using genomes that are not closely related can lead to generation of false positive TFBS predictions. Even for a set of closely related genomes, selecting a threshold for binding site detection can be difficult. The final step of the TFBS prediction involves the verification of site consistency. Early studies on E. coli and Haemophilus influenzae regulon predictions showed conservation of co-regulated genes by orthologous
TFs [97]. Based on this principle, a search is conducted for binding sites upstream from the operons regulated by each TF. If the site is conserved, the TFBS prediction is assumed to be correct. On the other hand, if matches to the predicted TFBS motif are found dispersed across the genome, the prediction is assumed to be a false positive. By accounting for changes in the operon structure, further consistency checks are possible. This method showed improved results in binding site detection in several studies such as nitrate and nitrite respiration in \( \gamma \)-Proteobacteria [98] and nitrogen metabolism in Gram-positive bacteria [99].

These methodologies have been implemented in the RegPredict web resource [100], a state-of-the-art tool for TRN reconstruction with TFBS data. The webserver comprises a large set of comparative genomics tools available in two reconstruction frameworks; the first reconstructs regulons for known PWMs and the second performs de novo regulon inference for unknown binding sites using analysis of regulon orthologs across closely related genomes. One of the novelties of RegPredict is the concept of clusters of co-regulated orthologous operons to facilitate and improve consistency check. This semi-automated approach provides the community with a more swift reconstruction, curation and storage of regulons. RegPredict was used for TRN reconstruction of the central metabolism of the Shewanella genus [101], for the analysis of the regulation of the hexunurate metabolism in Gammaproteobacteria [102] and for the elucidation of control mechanisms for proteobacterial central carbon metabolism by the HexR regulator [103]. FITBAR [104] is another web tool for prokaryotic regulon prediction that aims to fill the gap of the lack of statistical comparison for calculating the significance of the predictions.

Techniques also exist for predicting TFBSs when the available regulatory information is not sufficient for regulon-based approaches. Phylogenetic footprinting [105] identifies highly conserved untranslated regions upstream from the genes of interest, since these are prime regulatory site candidates. An orthologous search for these regions is performed across closely related genomes; candidate binding sites are identified and these sites are used to perform a regulatory motif search across all analyzed genomes. This technique successfully identified the FabR regulon in E. coli and regulon members in several cyanobacteria genomes [106]. Another approach has been described as subsystem oriented [92] based on the hypothesis that one TF regulates the genes on the same metabolic pathway. A search for orthologous genes on the same metabolic pathway of closely related genomes is conducted. Using the orthologous operons from the same subsystem, one can perform a motif search to build the PWM and search for TFBS.

Concepts of this approach were also implemented in RegPredict with the introduction of the SEED subsystems [61] for regulon reconstruction and curation.

**DE NOVO REVERSE ENGINEERING**

As gene expression data became available through microarray technologies, development began on methods for inference of regulatory networks from expression data [107]. Early reviews describe several mathematical formalisms such as Bayesian networks, Boolean networks and differential equations to represent regulatory networks [108], together with appropriate algorithms to support network inference.

The development of these methodologies led to the creation of the dialogue for reverse engineering assessments and methods (DREAM) project in 2007 [109], bringing together experts from different areas and aiming to provide tools to enable the unbiased evaluation of various methods [110], hosting annual challenges. The lessons gained from the results obtained in those challenges have provided improved methods for network inference [111]. Each year different methods are ranked as top performers on specific sub challenges that differ in either the type of data or network size.

Past reviews have categorized reverse engineering network inference methods according to: (i) mathematical modeling approach [81, 112], (ii) module-based or direct inference methods [80, 113] and (iii) unsupervised and (semi)-supervised methodologies [80, 114, 115].

In the first category [81, 116], the differential equation (ODEs)-based [117, 118], mutual information-based [119, 120] and Bayesian network-based methods [121, 122] are the most popular approaches. Other notable approaches are based on Boolean networks [123], neural networks [124, 125], correlation analysis [125] and relevance networks [127].

The second category divides methods into those based on a modular view of regulatory networks that infer regulatory programs for sets of co-expressed genes and those able to infer the regulatory behavior...
of individual genes (direct inference) [95]. Module-based inference is inspired by evidence that regulatory networks exhibit a modular structure of co-expressed genes [128, 129], using a separate algorithm for the module inference step, typically based on clustering or biclustering algorithms, such as cMonkey [130]. Direct inference methods search for single interactions between targets and their regulators [60, 131] (Figure 5). A comparison between representative methods of both approaches showed that none can be defined as the best solution [113]: the module-based method LeMoNe [132, 133] is able to retrieve more efficiently targets for regulators with a high number of targets and the direct-inference method, context likelihood of relatedness (CLR) [60] is preferable for detecting regulators with one or few targets. Thus, these methods can be seen as complementary when handling genome-scale regulatory model reconstruction.

The third category divides methods into supervised [134, 135] and unsupervised [136, 137]. The former use a training set of known interactions creating classification problems (e.g. to infer whether a given gene is regulated by a TF) (Figure 5). Some supervised methods are known as semi-supervised [138, 139]. Supervised methods have shown to provide more accurate predictions than unsupervised methods [140], with successes in expanding the compendium of TF-gene interactions in E. coli [138]. At the same time, when inferring interactions for an organism that is not well known, the lack of a proper training set can lead to a better performance by unsupervised methods.

A detailed review of the mathematical formalisms and detailed inference algorithms is out of the scope of this review. From the overwhelming number of methods available, we chose to briefly describe 10 methods, including the most widely used, the most recent [80] and the best performing from the DREAM challenges [110, 111, 141–143]. We focus our review on methods that produce genome-scale regulatory network reconstructions in the form of regulatory models that may be integrated with GSMs. While no method currently exists that completely satisfies these criteria, several algorithms, given in Table 3, can provide important results in the route to achieve the goal of fully integrated genome-scale models.

ARACNE [131] is one of the most widely used methods, first applied to infer regulatory interactions on human B cells [149]. Also, it has shown capacity for genome-wide inference in bacterial species such as Streptomyces coelicor [150]. CLR introduced the use of data from different experimental conditions for the same organism to infer regulatory interactions and enabled the identification of >700 novel interactions in E. coli [60]. Being one of the most cited methods with an ability to predict edges in the RegulonDB, CLR is the method of choice for regulatory interactions studies [151]. It was recently used to unveil virulence factors in Salmonella [152]. A newer algorithm based on CLR, called synergy augmented-CLR [144], was the best-performing method in the DREAM2 genome-scale inference challenge, exploiting the concept of synergy among multiple interacting genes [153], where a pair of genes is used to infer the expression of a third to increase prediction accuracy.

The Inferelator [146] was applied for genomewide reconstruction of Halobacterium. A mixed approach combining this method with CLR was one of the top performers in the DREAM3 in silico network challenge [145], using a modified version of CLR to compute mutual information values that are subsequently used by Inferelator to produce an ODE model. This method, called tlCLR (time-lagged CLR), takes advantage of two types of data: steady-state data from knockout experiments and

Figure 5: Network inference methods classification. (A) Network node module based versus direct inference. (B) Supervised versus unsupervised. Supervised methods require a training set of previous known interactions.
time series gene expression data. Another method using different types of data was introduced by Yip et al. [147] gathering steady-state data from a noise model and time series data from an ODE model; this method was the top performer of the DREAM3 in silico challenge. Most algorithms in Table 3 can use steady-state or time series data, thus showing the benefits of integrating both types of data.

DREAM5 featured a genome-scale network inference challenge with a large data set from a compendium of microarray data for *E. coli* comprising 805 chips, 334 TFs and 4511 genes. Large data sets were also provided for network inference on *S. cerevisiae* and *S. aureus*. Gene network inference with ensemble of trees (GENIE3) [135] uses tree-based methods decomposing the inference problem of $p$ size into $p$ distinct regression models. This method was the best performer overall and the top performer in the in silico network. GENIE3 had already been the best performer in the DREAM4 in silico inference for the 100-gene-multifactorial sub-challenge, where only multifactorial data were provided and showed equal capacity in successfully inferring networks from real data when compared was widely used methods such as CLR and ARACNE [135].

Several methods integrate multiple data types (e.g. inference from expression, binding site data) to facilitate TRN reconstruction. semi-supervised regulatory network discoverer (SEREND) [138] uses a semi-supervised and iterative approach to unveil regulatory interactions. SEREND depends on a curated set of TF-gene interactions and TF-gene motif scores as a training set to construct a logistic regression model. The known predictions are then expanded and the predictions validated with ChiP-chip and time-series expression data. This approach was used to better predict and to give new insights into the factors involved in activation and repression in the aerobic/anaerobic regulation mechanism in *E. coli* [138].

Gene promoter scan (GPS) [148] is also able to integrate other types of data; but as a module-based method, it follows a different approach. GPS is a machine learning method that builds promoter models and their relationships computed from a data set. In the next step, characterized profiles (groups of promoters) are generated. The best profiles are used as candidates for genomewide predictions. Studies with *E. coli* and *S. enterica* using GPS unveiled previously unknown interactions and novel members of the PhoP protein controlled regulon [148].

DISTILLER [137] is another method that exploits the concept of regulation modularity integrating other sources of data for network inference. This framework can be applied to any organism and incorporate motif and ChiP-Chip data. The integrated approach was used to study the ‘FNR’ regulon in *E. coli* identifying novel predictions that were experimentally validated. These studies provided insights on modularity dynamics pointing to the existence of polycistronic transcription [155].

A search for the best inference method usually turns to benchmarking studies; but the choice of benchmark data sets presents a problem, with different studies showing very sparse results [156, 157]. Lessons from all the DREAM challenges show that there is no individual best method. Results from community predictions, a combination of several reverse engineering methods, are closer to a state-of-art/best method, outperforming results from

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Modeling approach</th>
<th>Inference approach</th>
<th>Semi/supervised</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARACNE [131]</td>
<td>Mutual information (MI)</td>
<td>X</td>
<td>Yes</td>
</tr>
<tr>
<td>CLR [60]</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SA-CRL [144]</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>tCLR [145]</td>
<td>+MI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inferelator [146]</td>
<td>ODE model</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Yip et al. [147]</td>
<td>+Noise model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GENIE3 [135]</td>
<td>Regression tree</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEREND [138]</td>
<td>Logistic regression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPS [148]</td>
<td>Fuzzy clustering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DISTILLER [137]</td>
<td>Association rules (itemsets)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DI, direct inference; MB, module-based.

Table 3: Methods for reverse engineering of gene regulatory networks from expression data.
individual algorithms. The determination of error profiles enables the advantages and limitations of each inference method to be assessed in order to determine which method is ‘the best’ for a specific inference problem.

The methods described above show recent advances, providing a good summary of the huge number of approaches that have been put forward. However, the underlying problem is complex, given the large search spaces involved and the still restricted availability of data that leads to an undetermined problem where many solutions can explain the data equally well. Hence, most of the methods rely on heuristic methods using different strategies to simplify the problem. The most important simplification is to reduce the search for a network or model explaining the data, with a huge number of possible interactions between the different entities involved, to the search of individual interactions or to small clusters or modules. This allows in some cases for distinct methods to be integrated to better support the results and, in the most elaborate methods, being followed by steps of determining regulatory programs based on these individual interactions.

**PHENOTYPE SIMULATION BY INTEGRATED METABOLIC AND REGULATORY NETWORKS**

The simulation of phenotype from genotype using reconstructed models has been one of the major goals and challenges of systems biology [158–160]. Early work on the integration of metabolic networks with gene expression data revealed that some cellular phenotypes cannot be described by the metabolic flux distribution alone [13]. Whole-cell modeling is required to capture many phenotypes and while this has been one of the great challenges of the century [161], integration of regulatory networks is one key milestone toward achieving this goal [162]. Significant advances have been made in the reconstruction of metabolic, regulatory and signaling networks [7, 163], as well as in the integrated simulation of these three network types [164, 165]. Here, we focus on the potential for the simulation of integrated metabolic and regulatory networks and the challenges that arise in this integrated approach [166].

Several mathematical formalisms have been applied to model different types of biochemical networks (e.g. Boolean and Bayesian networks, constraints-based optimizations, ODEs). The many types of approaches for integrated network reconstruction and analysis have been reviewed recently [82, 167, 168]. Here, we focus on the methods that can be applied at genome scale, mainly stoichiometric models using the constraints-based approach [169, 170].

Constraints-based stoichiometric models do not account for intercellular dynamics. Instead, they assume a pseudo-steady-state for the cell, in which metabolite accumulation does not occur. This is described mathematically by a set of linear constraints on the flux through each metabolic reaction, defined by the mass balance for each internal metabolite (Figure 6):

\[ S \cdot v = 0, \]

where \( S \) represents the stoichiometric matrix and \( v \) the vector of fluxes through all metabolic reactions. The set of fluxes that satisfy these constraints define the feasible space for reaction fluxes (Figure 5). Constraints can be imposed on reaction reversibility and directionality (\( v > 0 \)), on enzyme capacity (\( v < v_{\text{max}} \)) and on nutrient availability and uptake.

Extensions have been made to these basic mass–balance and flux boundary constraints to capture the additional constraints imposed by regulatory interactions. Figure 7 shows existing methods for analysis and simulation of integrated metabolic and regulatory networks. Global network analysis methods such as extreme pathway analysis [171] were developed to analyze specific pathway properties, such as length and redundancy. These methods were used successfully to characterize changes in the solution space with the addition of regulatory constraints [28].

The FBA approach uses linear programming to identify the specific flux distributions that satisfy problem constraints and best reflect the state of the cell or represent target states for metabolic engineering [26, 172]. FBA was expanded to account for regulatory information with the introduction of rFBA [13], which uses a Boolean logic formalism to define additional constraints specifying which genes in the network are ON or OFF, based on specified stimuli (e.g. stress). This approach was successfully applied with the first genome-scale integrated model of metabolism and regulation in *E. coli*, resulting in the correction of several phenotype predictions compared with the use of mass balance and flux boundary constraints alone [14]. However, this approach requires the integrated model to be initialized at a relevant state for the regulatory components of the system. The Boolean
regulatory constraints are then applied to determine how the state of the regulatory components will change over time in response to stimuli. Selection of a relevant initial condition for the model remains a challenge for this methodology, since many equally consistent states exist for a set of stimuli, with equally valid associated flux distributions.

To address some of the limitations of rFBA, SR-FBA [27] was introduced, differing from rFBA in that it accounts for metabolic and regulatory constraints in a single step and quantifies the impact of these constraints on the flux distribution. This methodology enables the rapid exploration of feasible combined regulatory and metabolic states and it

Figure 6: Stoichiometric modeling. The metabolic network is used to construct the stoichiometric matrix using mass balances of the metabolites. The constraints-based approach is used to impose constraints to the stoichiometric model. $S.v = 0$, pseudo steady-state assumption; $v > 0$, reversibility constraint; $v < v_{max}$, capacity constraint.

Figure 7: Pathway-based and constraints-based methods for the analysis and simulation of integrated metabolic and regulatory networks. FBA, flux balance analysis; rFBA, regulatory FBA; SR-FBA, steady-state regulatory FBA; idFBA, integrated dynamic FBA; iFBA, integrated FBA; PROM, probabilistic regulation of metabolism; IOMA, integrative omics-metabolic analysis; tFBA, transcriptional controlled FBA.
rapidly identifies constraints that are internally inconsistent, preventing their simultaneous enforcement in a single steady-state. Yet, therein lies the substantial limitation of this approach, since inconsistent regulatory constraints often arise, because regulatory mechanisms exist to manage transitions between states of the cell in response to stimuli. Some of these transitions involve a cascade of intermediate unstable states that cannot be captured by the SR-FBA formalism. The constraints that manage these cascade transitions are not designed to be simultaneously enforced with all other constraints in the cell, meaning they appear to be internally inconsistent.

The quest for a whole-cell model led to the development of methods that also integrate signaling networks. Two methods have been proposed: integrated FBA (iFBA [164]) and idFBA (integrated dynamic FBA) [165]. iFBA is an expansion of the rFBA approach that aims to integrate signaling models, when available, for an organism or pathway being studied. An rFBA model for the central metabolism of *E. coli* [173] was combined with an ODE kinetic model for the phosphotransferase system, showing improved predictions compared with both rFBA and ODE models. The novelty of idFBA is the incorporation of slow and fast reactions in the stoichiometric framework. Slow reactions are incorporated directly into the stoichiometric matrix with a time delay; fast reactions rely on the pseudo steady-state assumption of the FBA approach. idFBA was applied to the analysis of yeast metabolism [174], demonstrating an approximation for the time-course prediction of time-delayed reactions, with the advantage of requiring fewer measured parameters than with full kinetic modeling.

Before methods such as rFBA, srFBA, iFBA or idFBA can be applied, TRNs must be translated into Boolean network models that connect external stimuli to internal metabolic reaction activity. The probabilistic regulation of metabolism (PROM) [175] approach was introduced to avoid the translation to Boolean constraints by enabling the generation of integrated models directly from high-throughput TRN data. PROM aims to circumvent the Boolean approaches that would consider a gene as either ON or OFF, with results outperforming rFBA. The differences in the predictions are attributed to the Boolean formalism of rFBA, which establishes a set of ‘rigid’ flux restrictions, where PROM presents a more continuous flux restriction.

The reconstruction of an integrated model for *M. tuberculosis* showed a potential use of PROM for drug target prediction. PROM can be seen as the closest methodology for semi-automated reconstruction of integrated metabolic and regulatory networks.

Transcription-controlled FBA (tFBA) [30] is another method that uses experimental expression data for the assessment of the regulatory state. Like PROM, tFBA aims to surpass the rigid ON/OFF gene states of a purely Boolean formulation by introducing the concept of more relaxed up/down constraints. As more experimental data are available, the level of expression of a gene can be observed to change under specific conditions without being entirely shut off. This method shows how the addition of large quantities of expression data can provide a way to improve FBA-based methods in the absence of kinetic parameters for metabolites and reactions.

**DISCUSSION**

In this survey, we begin with an overview of the data currently available for TRN reconstruction, revealing the limited number of data sets available for bacterial organisms, despite the massive amount of existing microarray data (Figures 1 and 2). We demonstrate through a phylogenetic analysis of the available expression data that large numbers of diverse organisms for which reference genomes are now available have never been examined using transcriptomic techniques. In order to fully understand bacterial regulation, expression data must be collected under a variety of conditions for as many diverse genomes as possible. We also show how next-generation sequencing technologies are beginning to dominate the latest submissions to the gene expression data repositories. While these new technologies enable the community to collect more data at a faster and cheaper price, they face the familiar problem of data standardization. Recent studies show how widespread batch effects, such as laboratory conditions, technicians and reagent brands lead to incorrect analysis of data and different results across different laboratories [176].

As for data relating to the regulons, TFs, binding sites and stimuli that comprise the TRN itself, comprehensive databases are available for a few specific organisms. Multiorganism databases do exist, but
these typically focus on one type of regulatory information, lacking the information needed to fully capture the regulatory effects. The latest version of RegulonDB makes an effort in the direction of representing the complete regulation by introducing genetic sensory response units.

Next, we examine how the methods applied for the reconstruction of TRN have progressed over the past decade. As the number of available reference genomes with expression data has increased, we see a corresponding increase in the number and power of approaches based on comparative genomics. With the increasing amounts of consistent high-quality expression data, we are also seeing increasing success with methods based on the reverse engineering of TRN from expression data. As these two examples amply demonstrate, the best method for TRN reconstruction depends on the amount and type of data available. Although the size of the desired TRN to be inferred is also an important factor, we suggest that genome-scale networks will always be desired in the near future. We also note the success of community efforts that combine the advantages of several reconstruction approaches, showing that hybrid approaches are the most successful given the present knowledge, where the complementary nature of the approaches helps to improve accuracy.

In the final portion of our review, we examine several approaches for the reconstruction and analysis of integrated metabolic and regulatory models. These approaches have been successfully applied to improve our ability to accurately predict phenotype from genotype, to explore the impact of regulation on the metabolic pathways and to simulate regulatory interactions that are continuous rather than discrete. Industrial successes in fields such as bioethanol production show the potential of current models and importance of improving these models [177, 178]. Adding a ‘layer’ of regulation can help unveil and predict unobserved phenotypes. Strain optimization has been one of the main objectives of metabolic engineering and the potential for improvements integrating regulatory information recently led to the development of methods that account for this type of information [179, 180]. Yet, we still lack a full understanding of the interplay between regulation and metabolism. Several studies have shown how major transcriptional changes are not always followed by changes in the metabolic flux [181, 182].

Several unknowns remain in the analysis and reconstruction of integrated biochemical networks, mostly because we do not yet possess a full understanding of regulation. For example, some efforts have been made to develop methods to account for metabolic activity effects regulated by post-transcriptional effects [183]. Methods such as PROM and tFBA allow the relaxation of constraints to try to account for regulatory effects. Even with transcriptional regulation, there are biological effects that these network models fail to reproduce. For example, chromosome structure can physically constrain bacterial transcriptional regulation [184]. Epigenetics of transcriptional regulation are also difficult to account for and some chemical marks have been described to be linked to this type of mechanism in bacteria [185].

Some of the methods described rely on basic assumptions such as that the same TF regulates orthologous genes or that the same TF may regulate genes in the same pathway. These assumptions may fail to represent reality, however, since TRNs show considerable plasticity in their structure. Orthologous regulators have been shown to control different pathways across different species [186] and global regulators have been shown to regulate different mechanisms [187]. Incorporation of models in evolutionary processes such as duplication and horizontal gene transfer has been proposed to deal with TRN plasticity [92, 188]. TRN also showed stochasticity [189], which can be an issue, especially when these networks are modeled by using a Boolean formalism that further propagates these stochastic effects [190].

Most recently Karr et al. [191] introduced the first whole-cell model for Mycoplasma genitalium. Integrating 28 submodels, the authors managed to validate the model across a wide set of experimental data, pointing out its potential for novel biological discovery in M. genitalium. It must be noted however, that M. genitalium is the smallest bacterial genome, with only 525 genes. Thus, while this methodology does represent a large step forward toward the goal of true whole-cell models, much more work must be done before similar models can be constructed for larger and more complex organisms.

As the pursuit of a whole-cell model continues, we expect novel regulatory interactions will be discovered in our drive to build a full understanding of the cell regulatory machinery.
Key Points

- Large numbers of phylogeny for which genome sequences are
  available still lack any gene expression data.
- Repositories of data on TRN tend to be comprehensive organism
  specific or narrowly focused multimammalian.
- The best methods for reconstruction of TRN from data depend
  on the size of the desired network and the types/amount of
data available; but, in general, hybrid methods that combine
many approaches produce the best results.
- Methods for integrating regulatory and metabolic models must
  include both steady state and dynamic components and they
must accommodate more than just Boolean regulation in order
fully capture the behavior of transcriptional regulation.
- Integration of regulatory constraints in GSM models results in
  substantial improvements in accuracy of phenotype predictions,
particularly since many phenotypes cannot be fully explained
without accounting for regulation. Yet, some regulatory mechan-
isms still exist that are poorly understood and require further
study.

FUNDING

J.P.F. acknowledges funding from [FRH/BD/70824/2010] of the FCT (Portuguese Foundation for Science and Technology) PhD program. The work was supported in part by the ERDF—European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness), National Funds through the FCT/COMPETE Programme, and by FCOMP-01-0124-FEDER-009707 (HeliSysBio—molecular Systems Biology in integrated Models) and FCOMP-01-0124-FEDER-009707 (HeliSysBio—molecular Systems Biology in Helicobacter pylori), the U.S. Department of Energy under contract [DE-AC02-06CH11357] and the National Science Foundation under [0850546].

References


Faria et al.


