Advantages and limitations of current network inference methods

Article *in* Nature Reviews Microbiology · October 2010

DOI: 10.1038/nrmicro2419 · Source: PubMed

2 authors:

Riet De Smet
Ghent University
23 PUBLICATIONS **1,239** CITATIONS
SEE PROFILE

Kathleen Marchal
Ghent University
489 PUBLICATIONS **10,041** CITATIONS
SEE PROFILE

Advantages and limitations of current network inference methods

Riet De Smet and Kathleen Marchal

Abstract | Network inference, which is the reconstruction of biological networks from high-throughput data, can provide valuable information about the regulation of gene expression in cells. However, it is an underdetermined problem, as the number of interactions that can be inferred exceeds the number of independent measurements. Different state-of-the-art tools for network inference use specific assumptions and simplifications to deal with underdetermination, and these influence the inferences. The outcome of network inference therefore varies between tools and can be highly complementary. Here we categorize the available tools according to the strategies that they use to deal with the problem of underdetermination. Such categorization allows an insight into why a certain tool is more appropriate for the specific research question or data set at hand.

Module inference
Identifying groups of co-expressed genes from gene expression data using clustering or biclustering algorithms.

Guilt-by-association principle
The assumption that genes with similar functions exhibit similar expression patterns. This allows the function of an unknown gene to be inferred from the function of annotated genes that are co-expressed with the unknown gene.

Centre of Microbial and Plant Genetics/Bioinformatics, Department of Microbial and Molecular Systems, Katholieke Universiteit Leuven, Kasteelpark Arenberg 20, 3001 Leuven, Belgium. Correspondence to K.M.
e-mail: kathleen.marchal@biw.kuleuven.be
doi:10.1038/nrmicro2419
Published online
31 August 2010

The insight that genes and proteins do not work in isolation but act together in intricate networks has launched the era of systems biology[1,2]. In bacteria, regulation at the transcriptional level is pivotal to guaranteeing metabolic flexibility and cellular integrity[1,2]. In *Escherichia coli* the transcription-regulatory network (TRN) was shown to be composed of basic modular components that contribute to the specificities of global response dynamics, for example by speeding up cellular responses or making them more robust (that is, able to respond to a wide range of environmental signals)[3,4]. Deciphering the gene co-expression network and the TRN (BOX 1) is therefore crucial to understanding bacterial cellular behaviour. The number of computational methods that are being developed to reconstruct TRNs from genome-wide expression data is rapidly increasing; here, these methods are referred to as expression-centred methods. Module inference methods, which focus on the co-expression network, rely on the guilt-by-association principle to identify functional relationships between genes, searching for gene sets or modules that exhibit a similar expression behaviour across experimental conditions (BOX 1). Methods that infer TRNs go one step beyond and infer causality relationships in the network by also identifying the transcriptional programmes of the genes or modules, to describe how transcription factors (TFs) cause the observed changes in expression of their cognate target genes (BOX 1).

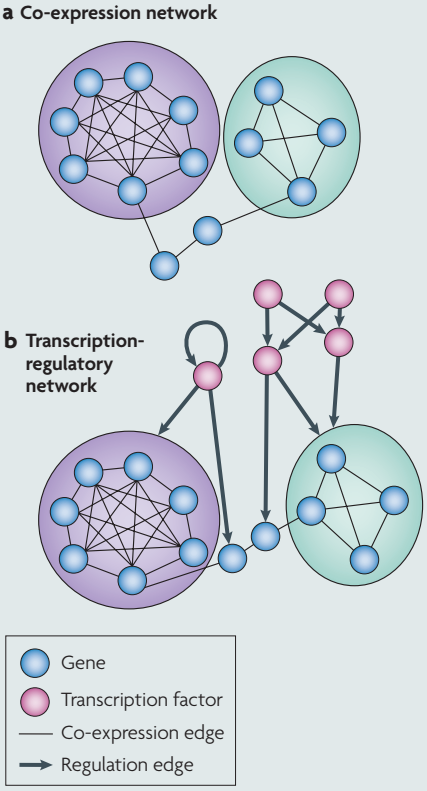
Applying these inference procedures on public data sets of well-studied model organisms has considerably improved our global understanding of TRNs. In bacteria, simple regulons that comprise only a few operons show expression modularity. The operon organization seems crucial for preserving this modular level of co-expression under some conditions, whereas under other conditions the presence of intra-operonic promoters breaks up the modularity[5–7]. In addition, complex regulation involving multiple regulators generally results in single genes showing highly specific expression behaviour that is not shared with other genes[8]. By focusing on the role of the regulatory programme in *E. coli,* it was observed that not only global TFs but also local regulators respond to a range of conditions[9]. In addition, many TFs are active in similar conditions and thus trigger similar sets of genes, suggesting either redundancy in their function or an intricate cooperation between different TFs to mediate a common response[9].

Several notable examples have set the stage for adopting inference methods in daily laboratory practice. The unprecedented link between protein mistranslation and the reaction to reactive oxygen species in response to antibiotics treatment was unveiled by combining network inference with experimental evidence in *E. coli*[10]. Similar approaches were used to unravel the complex network regulating host–pathogen interactions in *Salmonella enterica* subsp. *enterica* serovar Typhimurium[11] and to chart the transcriptional network of the archeon *Halobacterium salinarum* for the first time[12]. Computationally inferred interactions therefore offer a useful resource for putting experimental findings into a more global context, by finding novel interactions that have yet to be unveiled, by unfolding links between

Box 1 | **Co-expression networks versus transcription-regulatory networks**

Co-expression network
This is a network representation in which the nodes represent the genes and the edges represent the degree of similarity in the expression profiles of the genes (see the figure, part **a**). Cliques or highly connected subgraphs correspond to modules of co-expressed genes. The edges are undirected, indicating that they represent only a correlation or dependency relationship between the nodes and do not reveal the cause of the relationship.

Transcription-regulatory network
This is a bipartite graphical network representation in which the nodes represent either transcription factors (TFs) or target genes (or modules) (see the figure, part **b**). Edges are directed, as they reflect a causal relationship: they indicate that an observed correlation in the expression patterns of two nodes is caused by a node corresponding to a TF regulating a node corresponding to a target gene. A transcriptional programme corresponds to a set of TFs sharing the same set of target genes, ideally under a similar subset of conditions.



the pathway under investigation and other cellular processes or by identifying the conditions under which a regulator of interest is active.

To guide users in choosing the most appropriate network inference tool for their own application, we provide a scheme that allows state-of-the-art transcriptional inference methods to be classified on the basis of the strategies used to solve the inference problem, focusing mainly on top-down network inference methods. In contrast to previous categorizations, our classification uses a combination of criteria that relates directly to the biological interpretation of the outcome rather than being merely data set related[13] or computationally focused[14,15]. We use representative tools of each class to show how using different strategies results in inferring different types of interactions. We also describe how to interpret benchmark studies. Finally, we give a perspective on the future of these inference tools in light of novel data generation procedures.

Inferring TRNs is an underdetermined problem

Under the assumption that each gene is regulated by only one regulator, inferring the interaction network in *E. coli* would require the individual links between approximately 4,500 genes and each of the 300 known and predicted regulators to be assessed[16], a total of 1,350,000 (that is, 4,500 × 300) tests. When also taking into account the existence of combinatorial regulation and feedback loops, the theoretical number of combinations can no longer be exhaustively enumerated. This means that the number of possible solutions is prohibitively large, and clever algorithms or optimization strategies are needed to screen them in a time-efficient way. In addition, module inference (finding the best combination of genes and conditions that define a co-expressed gene set according to preset criteria) is prohibitively complex. The large number of possible solutions (the large search space), together with the restricted number of independent data points and the low information content of the available data[17–19], turns TRN and module inference into an underdetermined computational problem with many possible solutions, all of which explain the data equally well but only one of which can be the biologically true solution.

Extracting general tendencies from inference results (for example, assessing the number of genes that exhibit a modular expression behaviour or the differences in regulon size) can be better supported, statistically, than strongly emphasizing the individually inferred interactions. However, it is exactly these individual interactions that wet-laboratory researchers are interested in.

Strategies to deal with underdetermination

The problem of underdetermination relates to the size of the search space: the larger the search space, the larger the complexity of the inference problem and the more difficult it will be to find the unique solution that approximates the biological truth. To tackle this problem of underdetermination, module and network inference methods adopt different strategies to reduce the search space and/or extend the amount of independent information (FIG. 1).

'Conceptualization by simplifying biological reality' is a commonly used strategy that renders network inference more tractable. TRNs have been shown to be modular in structure[20], which implies that the network consists of overlapping modules of functionally related genes. Genes belonging to the same module act in concert under certain environmental cues[21–23], explaining their coordinated expression behaviour. Modules are identified by methods that rely on clustering or biclustering[24]. Module-based network inference procedures, which are primarily designed to infer transcriptional programmes, assign a regulatory programme to these modules, rather than assigning an individual programme to each single gene, as is the case with direct network inference methods. This drastically lowers the number of interactions that must be evaluated during the inference process. Another simplification relates to the definition of combinatorial regulation, in which multiple regulators act together to mediate specific condition-dependent responses. Inferring a transcriptional programme that uses combinatorial regulation means that all possible combinations of regulators and binding modes (that is, cooperative, synergistic and so on) must be evaluated in order to explain the observed expression behaviour. As this is computationally intractable for large data sets, all large-scale network inference methods make an approximation of combinatorial regulation.

Expression modularity
Refers to the modular structure of the co-expression network. This network can be broken down into modules, or groups of co-expressed genes, the function of which can be separated from that of other modules.

Top-down network inference
Reverse engineering or *de novo* reconstruction of the structure of biological networks on a genome-wide scale by exploiting high-throughput data. By contrast, bottom-up regulatory network inference is the construction of a quantitative model from the data using a known, mathematically formalized connectivity network as input; estimating the kinetic parameters of this model from the data allows the dynamic behaviour of the network to be modelled.

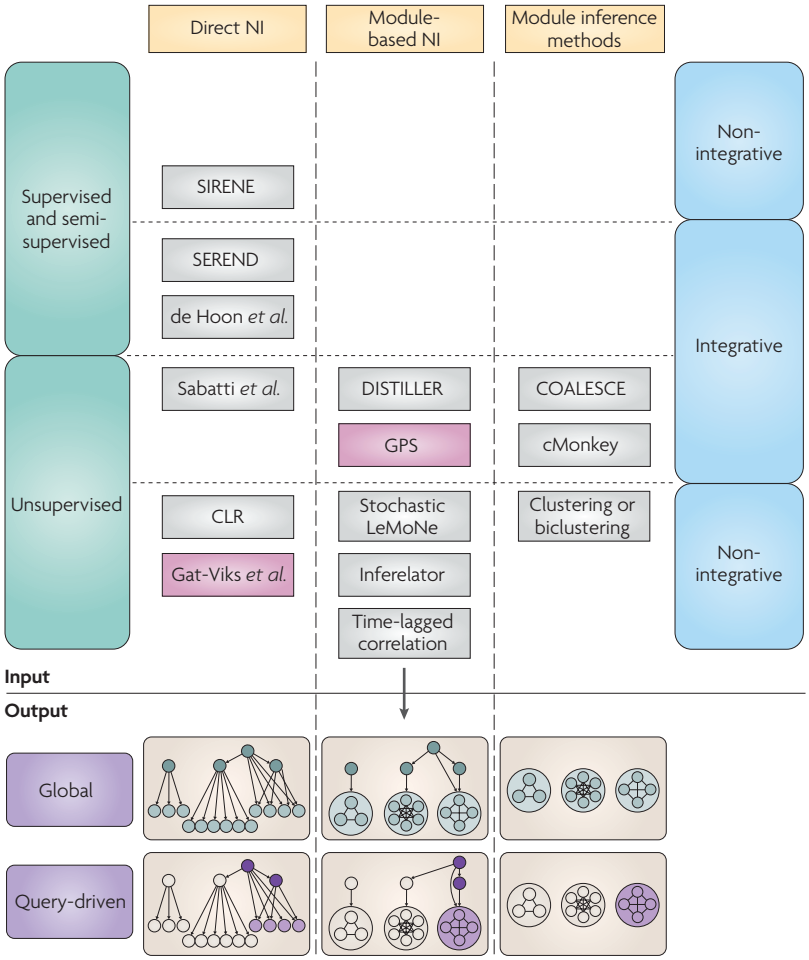


Figure 1 | **Categorization of different state-of-the-art methods for module and network inference.** Module inference methods search for sets of co-expressed genes. The major goal of network inference (NI) methods, on the other hand, is to search for a regulatory programme that explains an observed expression behaviour. NI methods can be categorized according to the strategies that they use to cope with the problem of underdetermination. Direct NI methods consider all genes on an individual basis, whereas module-based NI methods conceptualize the network by treating sets of co-expressed genes as single entities (modules). NI and module interference methods can be further divided according to whether they complement expression data with additional data sources (integrative methods) or use expression data only (non-integrative methods). Supervised and semi-supervised methods treat the inference problem as a classification problem, whereas unsupervised methods do not. The output of the methods can be global, indicating that they search for global patterns in the data, or query-driven, starting from a predefined set of core genes or core pathways and expanding on those. Most of the available programs can be used in either a query-driven or a global mode. The methods indicated in pink are specifically designed to be query driven. CLR, context likelihood of relatedness; COALESCE, combinatorial algorithm for expression- and sequence-based cluster extraction; DISTILLER, data integration system to identify links in expression regulation; GPS, gene promoter scan; LeMoNe, learning module networks; SEREND, semi-supervised regulatory-network discoverer; SIRENE, supervised inference of regulatory networks.

Optimization strategy
A strategy used to screen the search space so that the optimal (or almost optimal) solution can be found without having to evaluate all possible solutions.

A second strategy relates to extending the expression data with other available information. Integrative methods combine the expression data with complementary data describing the TRN from a different angle, such as chromatin immunoprecipitation-on-chip (ChIP-chip) data or motif data, and these methods often obtain more reliable interactions and a more complete picture of the network. Moreover, during the search, prioritizing predictions for which the independent data sources agree allows the search space to be traversed more efficiently.

As a third strategy, query-driven methods reduce the search space by intentionally restricting the number of interactions that needs to be evaluated: instead of searching for a global pattern, as global inference methods do, query-driven methods concentrate their search on a predefined set of core genes or on a subnetwork of interest, and they then expand on this core gene set or subnetwork given the available data.

A fourth strategy is to use supervised (and semi-supervised) methods, which treat network inference as a classification problem and can be considered to be a specific way of exploiting known information in a query-driven manner.

As each strategy uses different assumptions and poses different constraints, the specific strategy or combination of strategies that are adopted determine the type of interactions that can be found. This is shown below, using results obtained from state-of-the-art inference tools that have successfully been applied to microbial data sets. For an algorithmic description of the inference tools mentioned below, see BOXES 2,3.

Module-based versus direct network inference. Usually, a biclustering method is used for module inference[24]. Most module-based network inference methods also use module inference based on biclustering as a first step, before the assignment of the transcriptional programme. Exploiting the concept of modularity offers advantages from both the biological and the statistical points of view. Most module-based approaches not only predict regulatory interactions, but also identify the experimental conditions under which the predicted interactions take place. This information can be helpful for designing the appropriate conditions under which experimental validation of the predicted interactions should be performed[8,25]. Assuming that modularity exists also contributes to the statistical robustness of the inferred interactions: each of the co-expressed genes in a module confirm the data for the other genes in that module by providing evidence for a certain regulatory programme, whereas for direct methods the evidence for a particular regulator–target interaction is based on only a single-gene observation.

A comparison of the results from the direct network inference method CLR (context likelihood of relatedness) and the module-based method Stochastic LeMoNe (learning module networks) shows how adopting the concept of modularity determines the interactions that can be inferred (BOX 2; FIG. 2). By exploiting modularity, LeMoNe and related methods[26] can assign regulators with expression profiles that are less similar to those of their target genes than is the case with CLR or similar methods[27,28]. Indeed, LeMoNe performs better than CLR at inferring regulatory programmes for genes that are grouped in coarse-grained modules which correspond to larger pathways (for example, Fis, RNA polymerase σ-factor S (RpoS) and PurR) and for which the genes show an overall low degree of co-expression

Box 2 | **Expression-based and integrative network inference methods**

CLR (context likelihood of relatedness)[32] is an unsupervised, direct, expression-based network inference method that reconstructs an interaction between a transcription factor (TF) and a target gene based on a correlation in their expression behaviour, as assessed by the mutual information measure.

Stochastic LeMoNe[34] is an unsupervised, module-based method that infers the transcription-regulatory network (TRN) from expression data. It uses a fuzzy, two-way clustering approach to assign genes and conditions to modules and subsequently assigns a regulatory programme to these pregrouped gene sets. Each module contains the genes for which the expression profiles best fit the same multivariate normal distribution, which simultaneously partitions the conditions within the module according to overexpression or underexpression. The transcriptional programme assigned to each module consists of the set of regulators for which the expression profiles best explain all or part of the condition partitions in the module.

Inferelator[25] assigns a transcriptional programme to either individual genes or predefined modules of co-expressed genes that are obtained by the integrative module inference method cMonkey[47]. Multiparametric logistic regression is used to search for tightly co-expressed modules that are enriched for genes that make up highly connected subgraphs in metabolic and functional association networks and/or that contain statistically over-represented *de novo*-detected motifs. Inferelator uses standard regression with model shrinkage to build a parsimonious, predictive model for the expression behaviour of the module or the gene, using changes in environmental influences and TF expression levels as predictors. The design matrix can capture binary interactions (AND, OR or XOR interactions) between TFs.

DISTILLER (data integration system to identify links in expression regulation)[8] is an integrative, module-based network inference method that combines expression data with interaction data (for example, motif or chromatin immunoprecipitation-on-chip (ChIP-chip) data) to search for co-regulated modules. It uses an unsupervised strategy based on itemset mining to exhaustively enumerate all gene sets that are co-expressed under a subset of conditions and that share the same motifs. A probabilistic filtering step is used to identify the most relevant set of non-redundant modules from this exhaustive list.

The method described by Sabatti *et al.*[43] is a hidden-component model[9] that is related to the original network component analysis (NCA)[112,113] strategy. This method uses a linear model to decompose E, which is the measured expression data in a product of a sparse connectivity matrix (A) that contains the interactions between all TFs and their targets as well as P, the hidden condition-dependent activities of the TFs[112]. Methods differ in the way they use constraints to uniquely identify A and P. Liao *et al.*[112] constrain A using the known network, whereas Sabbatti *et al.*[43] use motif information as a prior in a Bayesian framework to guide the reconstruction of the unobserved TF activities and interactions. As these methods exploit known information to constrain their search space, they can be considered direct, integrative, unsupervised network inference methods.

COALESCE (combinatorial algorithm for expression- and sequence-based cluster extraction)[48] is an integrative, non-supervised module inference procedure that uses a Bayesian framework to integrate sequence and expression data. *De novo* motif detection occurs concurrently with the biclustering of the genes and conditions. Motifs, represented by probabilistic suffix trees, are assigned to a developing bicluster if their occurrence in the module is sufficiently enriched compared with their presence in the genomic background. Additional information on sequence conservation or nucleosome placement can be used to guide the motif and module inference.

Methods that explicitly use time series gene expression data to infer causal relationships are known as time-lagged correlation analysis methods. They generally consist of two steps[37,38]. In the first step, genes with similar expression profiles across multiple time points (by Pearson correlation) are grouped in a module or cluster. In the second step, causal effectors such as the regulators, the modules that contain the regulators[37], or environmental inputs[38] are related to the target modules using time-lagged correlation, a measure that is related to the Pearson correlation coefficient but that takes into account shifts in time between the expression of the causal effector and the target module.

with each other or with their transcriptional regulator[29]. Conversely, CLR has a higher precision than LeMoNe for identifying targets of those bacterial regulators that are dedicated to one or, at most, a few operons, because in bacteria such operonic regulators are tightly co-expressed with their targets (for example, glucitol operon repressor (GutR), IscR, BetI and arabinose operon regulatory protein (AraC)). A direct method such as CLR also covers interactions for a larger range of regulators than a module-based method such as LeMoNe, as module-based inference methods lose interactions with target genes that are not co-expressed with a sufficient number of other target genes[29].

Modelling combinatorial regulation. Inferring combinatorial regulation in its full complexity is also computationally intensive. Most direct methods, both supervised (such as SEREND (semi-supervised regulatory-network discoverer)[30] and SIRENE (supervised inference of regulatory networks)[31]; see below and BOX 3) and unsupervised (such as CLR[32]), simplify the problem by assigning regulators to their target genes one by one and composing the combinatorial regulatory programmes in a post-processing step that finds sets of regulators which control the same target genes. Although this substantially reduces the complexity of the network inference problem, such a stepwise approach renders it impossible to distinguish between truly complex combinatorial regulation, in which the signals of multiple TFs are integrated to trigger the observed gene expression pattern, and condition-dependent regulation, in which different TFs act independently to mediate expression of their target genes under different subsets of conditions. For example, applying CLR to data from *E. coli* resulted in the correct assignment of the regulators GadE, GadW and GadX to several genes involved in the acid response[32]. However, the true, more complex relationship of GadW and GadX with GadE, which is the main regulator of the acid response and is under the control of both GadW and GadX[33], could not be unveiled.

Module-based inference methods such as Stochastic LeMoNe[34] and DISTILLER (data integration system to identify links in expression regulation)[8] (BOX 2) automatically take into account the condition dependency of the inferred transcriptional programmes: regulators that are assigned to the same genes under different subsets of conditions are assumed to act independently, as each of them is responsible for triggering a different condition-dependent response. Regulators that are predicted to regulate the same target genes in similar conditions, on the other hand, are presumed to act combinatorially, as they are needed simultaneously to trigger the observed co-expression response. For example, using DISTILLER, it was predicted that the *E. coli* global regulator cyclic AMP regulatory protein (Crp) interacts, depending on the conditions, with different specific regulators[8]. Neither DISTILLER nor Stochastic LeMoNe can infer the mode of the combinatorial interactions between the assigned regulators — that is, whether the assigned TFs act together in an additive or synergistic way (AND), in a combinatorial interaction, such that the presence of one of the regulators is sufficient to trigger expression of the target gene (OR), or in a mutually exclusive manner (XOR). By combining the expression profiles of the regulators according to these different possible interactions

Box 3 | **Query-driven and supervised network inference methods**

The method proposed by Gat-Viks *et al.*[64] is a query-driven, expression-based inference method. Qualitative knowledge of a pathway of interest is formalized as a Bayesian network, in which the nodes represent different molecular entities (genes, proteins or metabolites) and the edges represent the interactions between them, with their corresponding connection logics. Such a probabilistic formulation of the network allows uncertainty to be included in the model. In a first model refinement step, possible model improvements (changes in topology and interaction logics) are evaluated. Refinements resulting in a model that better predicts the observed expression values are withheld. In a second expansion step, transcription factor (TF) activities are predicted from the network model, and a likelihood score is used to assign additional target genes for which the expression can be predicted by the TF activity profile. Thus, the method identifies sets of genes that are regulated by the same set of regulators and according to a common logic.

GPS (gene promoter scan)[62] is a query-driven, integrative network inference method that starts from a set of genes regulated by a common TF. Each gene is represented by a list of features, consisting of its expression profile and a detailed description of its promoter elements. The set of query genes is separated into distinct clusters according to these features, resulting in these genes being grouped according to their specific regulation patterns. A fuzzy *k*-nearest-neighbour classifier is used to extend the obtained clusters with new targets on the basis of the similarity between the feature profile of the new gene and that of a cluster.

SIRENE (supervised inference of regulatory networks)[31] is a supervised, expression-based, direct network inference method that splits network inference into multiple binary classification problems for each TF. One SVM (support vector machine)-based classifier is trained per TF, according to similarities in the expression profiles of target and non-target genes: genes regulated by a TF are likely to be co-expressed, whereas non-targets are not. This TF-specific classifier is then used to predict which genes are regulated by the TF, resulting in a ranked list of potential targets.

SEREND (semi-supervised regulatory-network discoverer)[30] and the method described by de Hoon *et al.*[68] are supervised (or semi-supervised), integrative network inference methods. A training set of known targets and non-targets is used to determine the parameters of two separate logistic regression functions that map the expression values and motif scores in the training set to their predictor variables (which determine whether the gene is activated, repressed or not regulated by the TF). The targets of a TF are thus expected to have a similar motif and expression profile. Motif and expression data are treated separately to guarantee proper balancing of the unequal number of features in each data set. A metaclassifier, also based on logistic regression, combines the outcomes of these separate expression-based and motif-based classifiers. The complete classifier is subsequently used to predict the probability that genes belong to the same regulon.

Search space
All possible solutions that need to be evaluated to find the one that is the most optimal according to preset criteria. In most inference problems, the number of possible solutions is prohibitively large and cannot be enumerated exhaustively.

Clustering
Grouping of genes that have similar expression patterns across all conditions.

Biclustering
Combining the selection of co-expressed gene sets with a condition selection step to infer the set of conditions that is relevant to the clustered genes.

(AND, OR or XOR) before assessing how well they explain the target's expression behaviour, Inferelator[25] can infer these more complex modes of transcriptional interactions. Recently, CLR was also extended to account for synergistic relationships (synergy-augmented CLR) — that is, when the expression value of a third gene can be better explained by two genes together than by each of them separately[35,36]. Using this approach, novel links were uncovered in the original *E. coli* CLR network, such as the fact that the expression of *fecA*, which encodes an Fe^{3+} dicitrate transport protein, depends on both *fecI* and *aceK* (encoding isocitrate dehydrogenase kinase/phosphatase), with *aceK* presumably acting as an indirect inhibitor of ferric citrate transport[36].

Integrative versus expression-based approaches. Non-integrative expression-based network inference methods extract information about regulator–target interactions from the expression data itself. Except for those supervised expression-based methods that exploit the observed co-expression behaviour of known targets of a particular TF, such as SIRENE[31] (see below), most non-integrative methods assume that the expression profile of the regulator is a proxy for its activity; for example, this assumption is made by Stochastic LeMoNe[34], CLR[32], Inferelator[25] and correlation-based methods[37,38]. This assumption disregards the important role of regulation mechanisms acting at levels other than transcription[39] and restricts the interactions that can be inferred to those of regulators that are either co-expressed or inversely correlated with their targets[40] (FIG. 3). As a result, expression-based inference methods such as CLR, Stochastic LeMoNe and other related methods[26–28] are biased towards inferring interactions of autoregulators or operonic regulators that are tightly co-expressed with their targets[29]. Moreover, most expression-based inference methods cannot distinguish between regulators that actually regulate a gene (that is, that have a direct causal effect) and regulators that are simply co-expressed with a gene (that is, mere correlation). This problem can be partially alleviated by inferring networks from dynamic data instead of from static data, as time series inherently contain information about causal effects, if one assumes that the expression of the TF needs to be altered before it can affect its targets (in a direct way or through a regulatory cascade). Inferring networks from dynamic data requires special techniques that capture the expression dynamics (for example, the lag in expression profiles between genes). Time-lagged correlation analysis (BOX 2) was used to infer the regulatory network that mediates the response to alternating light conditions in the cyanobacterium *Synechocystis*[38], and the *Bacillus subtilis* regulatory network was inferred using the same technique[37]. In practice, inference of networks from dynamic data is restricted by the insufficient time resolution of the available samples, which complicates the matter of distinguishing true from noisy signals and results in fast responses being missed.

By complementing gene expression with additional transcriptional information (such as motif data or DNA–protein interaction data), integrative network inference methods[8,30,41–45] can extend the scope of their predictions beyond interactions that can be inferred from co-expression behaviour and usually result in more reliable predictions (FIG. 3). Sabatti *et al.*[43] proposed a direct integrative approach based on hidden component analysis (BOX 2) that overlays a network topology derived from known and motif-based interactions with expression data. This method was used to infer the transcriptionally active edges in the *E. coli* network. By exploiting the known information on regulatory motifs and transcriptional interactions in the EcoCyc database[46] in a supervised way, the direct integrative method SEREND inferred novel interactions for previously characterized regulators of *E. coli* (see below).

Module-based network inference methods such as DISTILLER[8] (BOX 2) rely on an integrative module detection step to derive regulatory programmes. Integrative module inference (DISTILLER, cMonkey[47] and COALESCE (combinatorial algorithm for expression- and sequence-based cluster extraction)[48] (BOX 2)) searches for genes that not only show

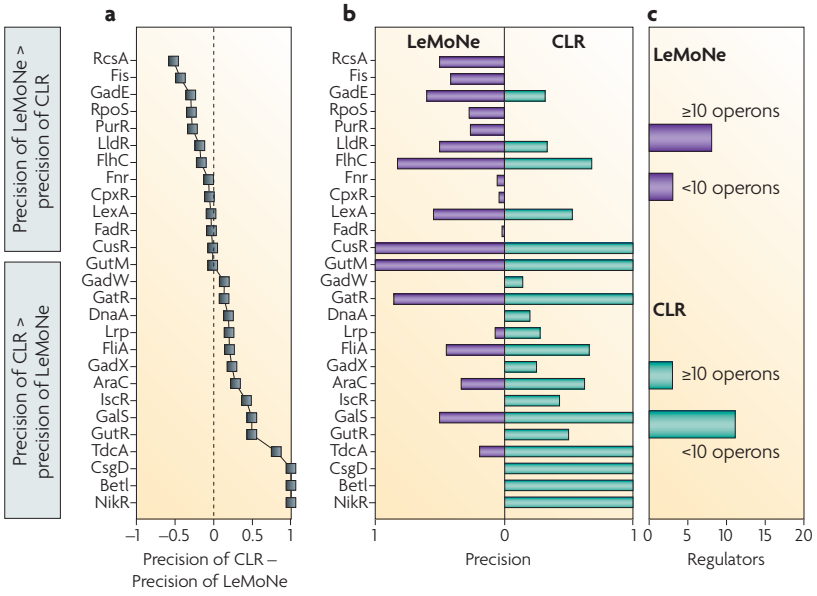


Figure 2 | **Complementarity in the type of interactions inferred by direct and module-based inference methods.** CLR (context likelihood of relatedness) and Stochastic LeMoNe (learning module networks), as representatives of direct and module-based inference methods, respectively, were applied to the same *Escherichia coli* compendium[32]. The precision of the inferred interactions was calculated as described in Faith *et al.*[32], using experimentally documented interactions in RegulonDB[69] as a standard. **a** | A comparison of the precision with which true interactions were inferred for both methods; the difference in the precision obtained with CLR and with LeMoNe was calculated for each regulator. Regulators are ranked according to this difference in precision. A high negative value indicates a higher precision for LeMoNe than for CLR, and high positive values indicate the opposite. **b** | The values of the regulator-specific precision for LeMoNe and CLR. **c** | The size distribution of the the known regulon membership, according to RegulonDB, for the regulators for which either LeMoNe or CLR show a higher precision. Parts **a** and **b** illustrate the complementarity between both methods in retrieving interactions for different regulators. Part **c** shows that LeMoNe predicts, on average, correct targets for more global regulators (with a larger regulon size), whereas CLR typically predicts targets for regulators with fewer known targets. Note that predictions for regulators that are not documented in RegulonDB are not included in this plot.

co-expression, but also share a common regulatory binding site (identified by *de novo* motif detection or ChIP analysis). Exploiting complementary data sources to confirm expression-based module assignments reduces the assignment of false members to true modules and the detection of spurious modules. As the observed co-expression in a module also implies true co-regulation when using integrative module inference methods, the module inherently contains information that infers the transcriptional programme: for example, each module is assigned the regulator that is known to recognize the motif or binding site associated with the module. Applying DISTILLER to a cross-platform *E. coli* expression compendium and motif data for 67 known regulators resulted in the prediction of 278 new interactions for 29 different regulators[8]. Of the 11 new interactions for fumarate and nitrate reduction regulatory protein (Fnr) that were experimentally verified by ChIP–quantitative real-time PCR, none were predicted by the non-integrative methods CLR[32] and Stochastic LeMoNe[29]. When using integrative approaches in combination with *de novo*-detected motifs, the assignment of a cognate regulator will be based on additional, computationally derived criteria (for example, the genomic proximity of the genes encoding the regulator and its targets)[5] or on a concomitant expression-based inference of the regulatory programme[25]. In the future, mapping of cognate regulators to novel motifs will be further facilitated by integration with data resulting from protocols that globally survey an organism's proteome for sequence-specific interactions with putative DNA regulatory elements[49,50].

So, inference methods that use only expression data are useful for organisms for which there is little additional information available. Integrative methods, on the other hand, provide a more complete view of the network and are more likely to predict true positive interactions. However, the additive value of integrative methods depends on the quality of both the additional data[51] and the algorithm used.

Global versus query-driven inference. Global module inference methods[22,52–59] search for the modules that explain most of the data. This usually corresponds to identifying large pathways that consist of many genes and that are usually responsible for the general responses to major metabolic or condition shifts, such as the pathways that regulate flagellar synthesis, amino acids biosynthesis and the DNA damage response. As such, global approaches provide a general view of the active TRN and the resulting physiological state in the cell. Query-driven module detection methods, on the other hand[60,61], search for genes that are co-expressed, in a condition-dependent way, with a predefined set of genes (also called query genes). These algorithms are deliberately biased towards finding a specific local solution in the search space according to the particular interests of the user. This solution is usually not easy to find using a global approach, as either the expression signals of the query genes are too low to be significant or the local solution is obscured by a more global one. For example, searching an *E. coli* compendium for a PurR-related module using a known PurR target as a query returns a module that is indeed significantly enriched for previously known PurR targets ($P < 1 \times 10^{-15}$), whereas with a global approach the module that contains the most PurR-related genes (under default conditions) is much larger and enriched for more general functions related to amino acid biosynthesis and translation (R.D.S., unpublished observations). Query-driven approaches are thus typically used to expand or curate a particular pathway or process either by searching for additional genes that are co-expressed with genes known to be involved in the pathway or by filtering out genes that are not co-expressed with the majority of the so-called pathway genes. For instance, the query-driven Signature Algorithm (SA) refined the gene set involved in the tricarboxylic acid (TCA) cycle in *Saccharomyces cerevisiae* using the homologues of 37 *E. coli* TCA cycle genes as queries[61].

Most of the global network inference methods described above can be applied in a query-driven setting by restricting their input data sets. In some cases

Motif
TF-binding site or specific sequence tag that is recognized by a TF and is located in the promoter region of a gene.

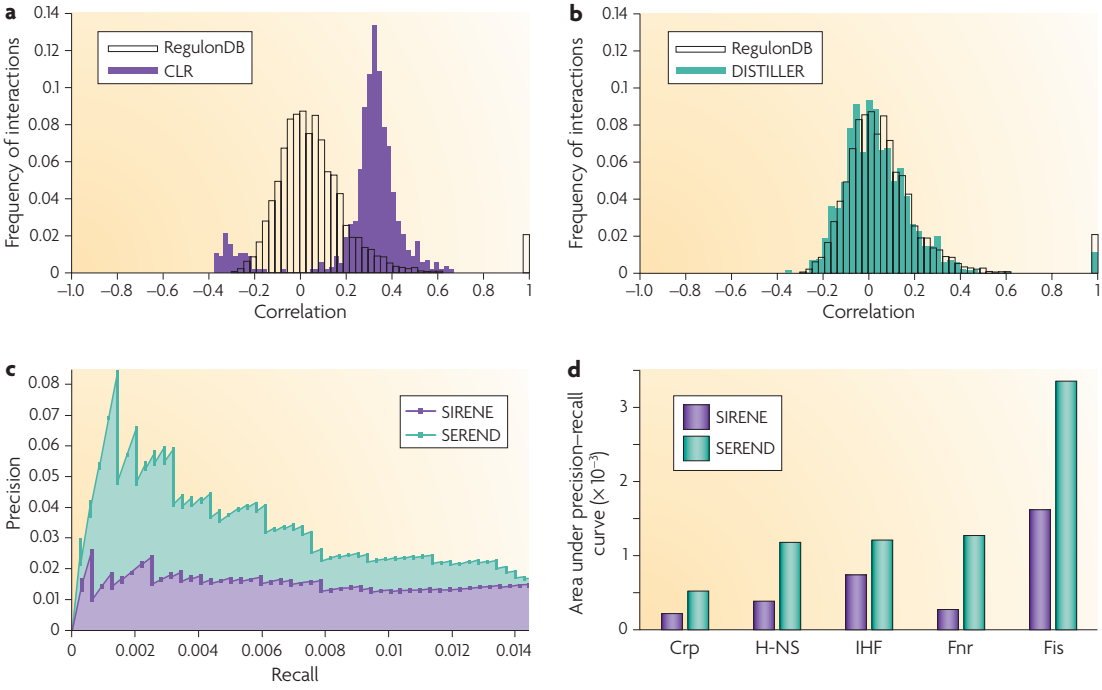


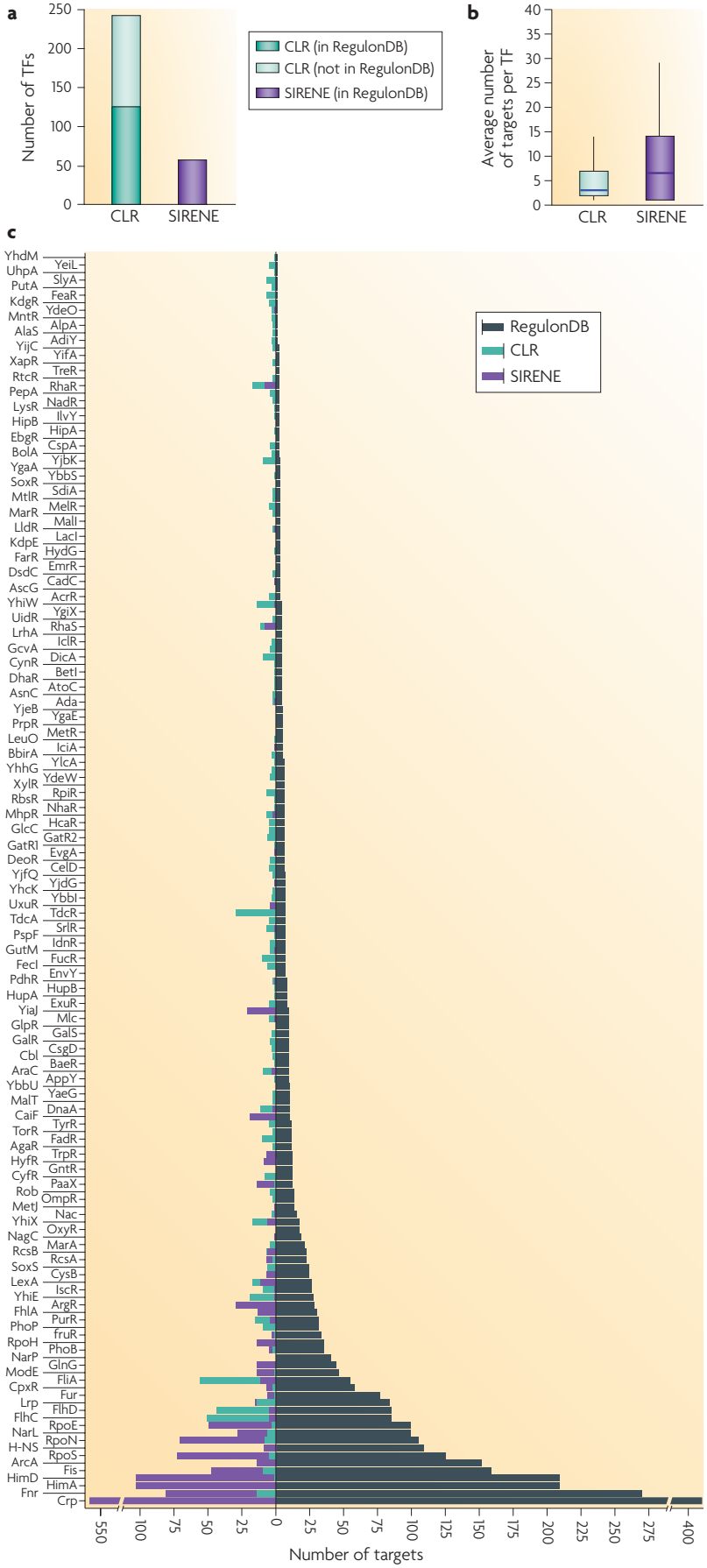
Figure 3 | **The different characteristics of interactions inferred by expression-based and integrative network inference methods. a**,**b** | Expression-based methods that estimate the activity levels of the regulators from their expression profiles are biased towards predicting interactions for regulators that are tightly positively or negatively correlated with their targets. For methods that infer regulatory programmes from complementary data sources, this is not the case. The expression-based method CLR (context likelihood of relatedness; part **a**) and the integrative method DISTILLER (data integration system to identify links in expression regulation; part **b**) were applied to the same *Escherichia coli* expression compendium (results were taken from Lemmens *et al.*[8]). The histograms display the number of predicted pairwise TF–target interactions as a function of their mutual co-expression. As a reference, the same distribution is shown for all interactions documented in RegulonDB[69]. A correlation coefficient of 1 corresponds to the situation in which the profiles of the regulator and the target gene are exactly the same, which is the case for autoregulators. **c**,**d** | Integrative methods result in more reliable predictions than methods that use only expression information. The performances of an expression-based network inference method (SIRENE; supervised inference of regulatory networks) and an integrative (SEREND; semi-supervised regulatory-network discoverer) network inference method are compared using chromatin immunoprecipitation-on-chip (ChIP-chip) data as an external standard. Part **c** displays the precision–recall curve for SEREND and SIRENE predictions made for cyclic AMP regulatory protein (Crp). The area under the precision–recall curve, indicated by shading, is used as an estimate of the overall performance of the network inference method. Part **d** compares the areas under the precision–recall curves for SIRENE and SEREND for five different regulators for which ChIP-chip data[114–116] is available: Crp, H-NS, integration host factor (IHF), fumarate and nitrate reduction regulatory protein (Fnr) and Fis. SEREND, the integrative method, outperforms SIRENE in retrieving ChIP-chip targets for each of the regulators considered.

Classification problem
A problem that can be solved by a system whereby properties or features of known targets and non-targets of a regulator are derived from high-throughput data and used to construct a classifier function — that is, a mathematical function that describes the relationship between the class labels (being a target versus being a non -target) and the corresponding properties of the high-throughput data. These classifier functions can then be used to predict whether or not a gene of interest is a target of the studied TF on the basis of its data properties.

Operonic regulator
Regulator dedicated to one specific operon.

De novo motif detection
Computational strategy to identify TF binding sites without any prior information on the sequence of the site. Such a strategy relies on certain subsequences being statistically over-represented in a set of co-regulated genes.

this can be advantageous; for example, methods such as CLR, Stochastic LeMoNe and Inferelator perform better if the transcriptional programme can be inferred from a prespecified list of regulators rather than from a full gene list, because erroneous interactions with non-regulators will be eliminated a priori. Algorithms specifically designed for query-driven network searches focus on one or a few core pathways[62–65]. By constraining the search space to only those solutions that contain the query, these methods can make more detailed network models than would be possible in a global setting. Gat-Viks *et al.*[64] (BOX 3) formalized the existing qualitative knowledge about the yeast osmotic response as a probabilistic model. Interrogating this model with expression data allows both refinement of the model, by correcting erroneous interactions, and extension of the original network with novel targets that are affected by components of the original network. Alternatively, kinetic approaches for modelling the dynamics between TFs and target genes from time series expression data, which are still intractable on a genome-wide scale, have been successfully applied in a query-driven mode to validate the outcome of a ChIP-chip experiment. So far, these approaches have only been applied to higher eukaryotes[66]. The GPS (gene promoter scan) algorithm[62] (BOX 3) is another query-driven network inference method that takes advantage of detailed promoter descriptions in combination with expression data from mutants to extend the regulon of a predefined regulator. More specifically, GPS identified four additional PhoP targets in *S.* Typhimurium that were previously thought to be only indirectly PhoP dependent. Furthermore, the identified PhoP targets in *E. coli* were assigned to different modules, one of which primarily contained genes that are



involved in acid resistance. This allowed a novel link between PhoP regulation and bacterial acid resistance to be established[62,67].

Supervised versus unsupervised inference of the regulatory programme. Supervised methods treat inference as a classification problem. They start from a set of known TF–target interactions and, on the basis of this predefined training set, characteristic features are derived, such as TF binding sites (SEREND[30] and de Hoon *et al.*[68]) or the degree of co-expression between TF targets (SIRENE[31], SEREND[30] and de Hoon *et al.*[68]). These characteristics are subsequently used to evaluate a new candidate gene as a potential target of a TF. Genes that share many characteristics with the known targets of the TF are classified as true targets, and the others as non-targets. Such a classification strategy depends on the quality of the training set of true-positive and true-negative interactions. It is straightforward to extract examples of positive interactions from curated databases, such as RegulonDB[69] (*E. coli*), EcoCyc[46] (*E. coli*) and DBTBS[70] (*B. subtilis*) (see Supplementary information S1 (table) for further information about databases), but the definition of true-negative interactions is much less trivial. Genes that are not known to interact with a specific regulator — that is, 'unknowns' — are often treated as negatives. However, our knowledge of TRNs is still limited and there is therefore a good chance that such a set of 'unknowns' contains as-yet-uncharacterized true-positive interactors for a given TF, in which case the classification results will be deteriorated.

By extrapolating from previously known information, interactions that are predicted with supervised methods are generally reliable but are restricted to regulators with sufficient previously known targets, such as global regulators and σ-factors from well-characterized model organisms (such as *E. coli*[30,31] and *B. subtilis*[68]) (FIG. 4). SEREND was shown to be very useful for extending the

Figure 4 | **Complementarity in the type of interactions inferred by supervised versus unsupervised network inference methods.** SIRENE and CLR (context likelihood of relatedness), as representative supervised and unsupervised network inference methods, respectively, were applied to the same *Escherichia coli* compendium[32]. For both methods, the top 1,422 interactions were considered. **a** | The number of transcription factors (TFs) for which interactions could be inferred by each method. **b** | The average number of targets inferred per TF by each method. As they exploit known information, supervised methods are more comprehensive than unsupervised methods for predicting targets for a specific regulator. **c** | The number of documented targets for all of the regulators reported in RegulonDB[69], ranked accordingly, is shown on the right side of the graph. The regulators for which most targets have been described to date correspond to global regulators and σ-factors. For each inference method, the number of inferred interactions per regulator is indicated on the left side of the graph. Supervised methods are biased towards predicting targets for those regulators that have a sufficiently high number of previously known targets.

Precision–recall curve
Customary method of comparing the precision and recall of a network inference method in order to evaluate the performance of inference algorithm. The precision is the proportion of correctly inferred interactions, according to an external standard, out of the total number of predictions made. The recall is the degree to which the total number of existing interactions in the real network has been covered by the predictions.

Cross-validation
Statistical technique that assesses the extent to which a model fitted on a certain data set can also predict the observations made on an independent data set.

repertoire of interactions of the *E. coli* global regulators integration host factor (IHF), H-NS, Crp, Fnr and Fis[30].

To infer interactions in less studied organisms, unsupervised approaches are more suitable (such as Stochastic LeMoNe, CLR, DISTILLER and Inferelator), as they do not necessarily depend on previously known information and they can also infer interactions for regulators for which there is little or no prior knowledge (FIG. 4). Unsupervised methods that can infer transcriptional programmes from only expression data, such as CLR and Inferelator, have been shown to be useful for providing a first, global view of the TRNs of, for example, *S.* Typhimurium[71,72], *Shewanella oneidensis*[73], *Halobacterium salinarum*[12] and Cyanobacteria[74].

Choosing benchmark data sets

Benchmarking is important for being able to understand the reliability of the reconstructed network. It is based on the precision–recall curve, which is calculated according to a predefined external standard. This standard is generated by collecting all curated interactions for a particular organism and treating them as true positives, and treating as false positives all predicted interactions between a gene and a TF that are not documented in the curated database. Using such a standard tends to overestimate the false-positive prediction rate, as most genes probably interact with many more TFs than is currently documented. Moreover, the assessment ignores all new interactions with those TFs for which no interactions are documented yet. As a result, use of an external standard rewards methods that merely reproduce current knowledge but penalizes those that perform well in finding new results. To compensate for this, most current studies combine validation based on an external standard with medium-throughput experiments to also validate the new results[8,9,32].

Medium-throughput experiments avoid the unfeasible task of testing all new predictions by sampling a set of predicted interactions that is representative for the whole analysis. In practice this set usually consists of both high-confidence and low-confidence interactions for one or a subset of the assessed TFs. For *E. coli*, mainly global regulators were chosen, such as Fnr[8] and leucine-responsive regulatory protein (Lrp)[9,32], as for these regulators there is a good balance between undiscovered and already known interactions, which favours benchmarking. For example, by combining performance analysis using RegulonDB with a ChIP-based medium-throughput experiment, three groups showed that their respective methods each had a good sensitivity for detecting known interactions but also that high-scoring new predictions usually corresponded to true interactions[8,9,32].

For network inference methods that use predictive models, cross-validation can be used to validate the reliability of the inferred model; this method assesses the ability of the model to predict the expression behaviour of genes in experiments that were not used to build the model[25,34].

In several studies, ChIP-chip-derived interactions have also been used as an alternative standard to benchmark algorithms but, like any high-throughput data source, they contain many false-positive (or non-functional) and false-negative interactions. This explains the low performances that are often observed in benchmark studies using ChIP-chip data (FIG. 5).

Obtaining insight into the behaviour of the algorithm requires a more objective validation strategy that uses perfect standards, made *in silico* by simulating data that mimic real data[75,76]. Simulated data are very useful for unveiling the qualitative properties of the algorithm under all kinds of test conditions that can never be obtained with real experimental data (for example, they can be used to test noise robustness, the sensitivity of the parameter settings and the optimality of the proposed solution)[77]. Their drawback is that they can never grasp the full biological complexity of real data (such as the exact properties of the experimental noise or the multilayered aspect of gene regulation[78]). To further bridge the gap between *in silico* and real data, the use of synthetic gene networks has been proposed[79]. These are engineered circuits with well-defined network topologies and interaction structures. The dynamic behaviour of such circuits is fully characterized using real measurements, and the resulting models are used to simulate data on which inference methods can be tested.

Benchmark studies are extremely useful for guiding both users and developers. However, relying on a benchmark study to find out which algorithm is 'the best' is difficult, as the choice of an appropriate inference tool depends on the research question posed. Fair benchmark studies should describe not only in what respect an algorithm is the best, but also where it fails. The quality of a benchmark study also depends on the extent to which parameter tuning is performed to guarantee that each of the applied tools performs optimally in the setting in which they are used. In this regard, the DREAM (Dialogue on Reverse Engineering Assessments and Methods) initiative[78,80] offers a platform for the unbiased assessment of network inference methods. They organize a yearly competition in which developers can participate with their own method to infer networks from blinded data sets.

Exploiting the complementarity

The overlap between inferred results from different methods can be very low, as illustrated in FIG. 5. This, together with the observation that the results of each of the tested methods show a similar degree of overlap with an external validation standard (RegulonDB[69] or ChIP data), indicates that this discrepancy in predicted interactions is not due to the failure of one of the methods to infer biologically relevant interactions, but is rather due to the complementarity of the different methods.

It is likely that no single best method exists, and different methods highlight different interaction types, so aggregating the outcomes of complementary methods offers a means of improving the breadth and the accuracy of the predictions. This idea of combining the outcomes of different methods has already been suggested in various contexts[81], and a 'reverse-engineering by consensus' approach has been advocated recently[80,82], as a result of the outcomes of the DREAM2 and DREAM3

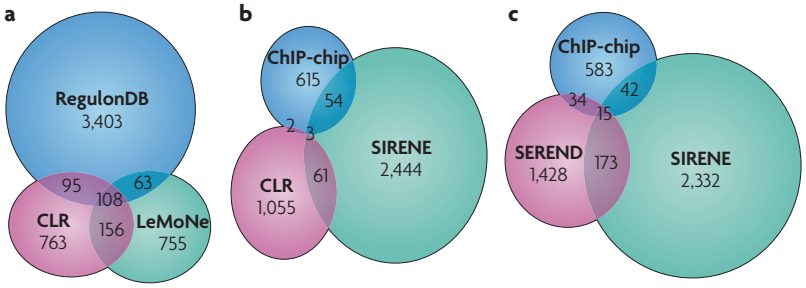


Figure 5 | **The low overlap of the predictions made by different network inference methods that rely on different strategies.** Various network inference methods were run on the same *Escherichia coli* gene expression compendium[32] and their results were compared. The proportion of shared predictions out of the total number of predictions ranges from 5.7% to, at most, 24%. The overlap with RegulonDB ((number of interactions in common with the external standard / total number of predicted interactions) x 100) ranges from 15% to 18%, and the overlap with chromatin immunoprecipitation-on-chip (ChIP-chip) data ranges from 2% to 3%, with a very low performance for CLR (context likelihood of relatedness) predictions compared with ChIP-chip data (<1%). **a** | A mutual comparison between the results of the module-based approach Stochastic LeMoNe (learning module networks) and the direct method CLR, both of which are non-integrative and unsupervised, using the known network data in RegulonDB[69] as an external standard. **b** | A comparison between the results obtained using CLR and the supervised method SIRENE (supervised inference of regulatory networks; both methods are non-integrative and direct). Available ChIP-chip data for several *E. coli* regulators was used as an external validation standard[114–116], as SIRENE uses the information in RegulonDB to make its predictions. **c** | A comparison between the results of the non-integrative method SIRENE and the integrative method SEREND (semi-supervised regulatory-network discoverer), which combines expression data with motif data (both methods are supervised and direct). Available ChIP-chip data was used as an external standard, as in part **b**.

conferences. At these meetings, it was shown that an ensemble of the predictions made by the best performing methods of the DREAM contest more closely approximated the true interaction network than did the predictions made by each method separately.

To construct an ensemble solution that reflects an overall statistical confidence in each of the predicted interactions, inference methods are required that provide an explicit ranking of the predicted interactions according to the scoring scheme they use; such methods include Stochastic LeMoNe, CLR, DISTILLER, SEREND and SIRENE. These individual rankings can then be combined into a ranked ensemble solution that assigns a higher confidence to interactions that are repeatedly retrieved by the different methods.

As well as being useful for combining the outcomes of different methods, an ensemble solution can be used to integrate different results from a single method. Because of the large search space, finding the most optimal solution to a network inference problem is non-trivial, and optimization algorithms often result in suboptimal solutions that all approximate the true global optimal solution but differ slightly from each other. For methods that can capture different possible solutions, a consensus solution from interactions that are repeatedly inferred from the data[34,83] allows the accuracy of the predicted interactions to be increased by better approximating the global solution.

At this stage, only tentative steps have been taken to improve on TRN reconstruction through ensemble methods. Much more work is needed to assess whether ensemble solutions will succeed in simultaneously increasing precision and recall of the predicted interactions.

Conclusions and future directions

To make sense of the flood of high-throughput data that is being generated, it is necessary to integrate the use of inference methods into daily laboratory practice to assist researchers in grasping higher-level biological insights or in prediction-based hypothesis testing. State-of-the-art inference tools rely on a unique combination of strategies to solve the inference problem. Because each strategy applies different assumptions, they each have different strengths and limitations and highlight complementary aspects of the network. Categorizing the tools according to their strategies allows users to gain insights into the settings under which they can most optimally be applied. The tool that is most appropriate for a certain researcher depends on the available data and the research purpose.

The nature of the expression data generally determines whether a direct or module-based inference method will be more appropriate. When the set of available expression data is large and/or heterogeneous in the assessed conditions, module-based inference methods are to be preferred over direct inference methods. When aiming to reconstruct the complete TRN, global inference methods are more suitable than query-driven approaches. For less studied microorganisms for which only expression data is available, expression-based network inference methods are ideal for making a first-draft reconstruction of the TRN. Integrating high-throughput data on TF–target interactions along with the expression data will generally allow for a more accurate (that is, with fewer false-positive interactions) and more complete picture of the TRN — including the prediction of combinatorial control, for example. But this method might become restrictive, inferring interactions for only those TFs for which the additional information is available. This is a disadvantage if one wants to derive global network properties. When a researcher is interested in expanding our knowledge of a particular part of the regulatory network rather than gaining a complete network view, query-driven methods are to be chosen over global approaches. When a reconstructed network is to be used as a starting point for the generation of further biological hypotheses, methods that provide an explicit ranking of the inferred interactions are advantageous, as this allows the researcher to prioritize candidates for further experimental work. Moreover, in such cases researchers benefit from using an integrative or supervised approach that exploits the properties of existing interactions to infer highly reliable new interactions. However, the more the method is biased towards existing knowledge, the more it will be blind to novelty. To take full advantage of the complementarity between the different methods, a 'reverse-engineering by consensus' approach seems to be the ideal option, combining the knowledge gained from multiple inference methods or from multiple outcomes from a single computational approach[80,82].

The advent of novel technologies such as tiling arrays and, more recently, deep-sequencing techniques[84,85] gives further importance to network inference. Although most inference methods can be readily applied to these new types of expression data, as they are insensitive to the type of technology used to generate the data, they will have to be adapted to account for the more detailed level of information that results from these novel technologies, including the presence of *trans*-acting small RNAs[86] and riboswitches[87], the non-static structures of operons with multiple intra-operonic transcription sites[6,7] and so on. As well as the increased level of detail, these novel technologies provide information that was not accessible before: re-sequencing the genomes of individual bacterial strains pinpoints strain-specific mutations and copy number variations in both coding and non-coding regions, and ChIP-seq (ChIP followed by sequencing) or ChIP-tiling (ChIP followed by microarray analysis) provides more detailed mapping of the genomic regions in which *cis*-acting regulators or nucleoid proteins bind[88]. The regulation of transcription can be described from multiple angles using this new data, and so integrative methods are now further challenged to provide a more accurate and detailed picture of the TRN and to consider the full dynamics of the system[89].

Although most inference studies carried out to date have focused on understanding the condition-dependent behaviour of a TRN in one specific model bacterial strain, these new types of information that are available have opened a new application field, called 'individualized, expression-centred' network inference. Expression-centred inference uses the premise that most of the mutations or changes occurring in the regulatory network at levels other than transcription will eventually lead to an altered expression profile. This assumption allows the expression profiles of individual strains to be considered as specific phenotypes or traits[90–95]. Additional sequence-derived genomic information can then be used to explain individually observed variations in expression behaviour, similarly to the identification of eQTLs (expression quantitative trait loci) in higher eukaryotes. Inference methods that generate an explicit explanatory model for the observed expression profiles (for example, Inferelator and Stochastic LeMoNe) can easily be extended for this purpose[96–99]. Linking adaptive changes of microbial genomes[100–102] to altered expression behaviour will unveil fundamental insights into microbial evolution and will identify the multifactorial changes that underlie industrially relevant properties of naturally occurring bacterial or yeast strains[103]. Moreover, a better fundamental understanding of how expression behaviour is encoded in the genome will help further rationalize synthetic biology[104,105]. Most of the inferences from such an expression-centred approach will provide only an indirect link between the observed genomic or epigenetic alteration and the observed strain-specific expression profiles. Future inference tools should focus on finding the hidden path between a genomic change and an alteration in gene expression, by exploiting information that is available about all levels of regulation, such as the transcriptional, post-transcriptional, signalling and metabolic levels[97,106–111].

Individualized expression-centred inference studies will not only complete, but also revolutionize our understanding of bacterial gene regulation.

1. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356 (1961).
2. Ptashne, M. & Gilbert, W. Genetic repressors. *Sci. Am.* **222**, 36–44 (1970).
3. Alon, U. Network motifs: theory and experimental approaches. *Nature Rev. Genet.* **8**, 450–461 (2007).
4. Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genet.* **31**, 64–68 (2002).
5. Fadda, A. *et al.* Inferring the transcriptional network of *Bacillus subtilis*. *Mol. Biosyst.* **5**, 1840–1852 (2009).
6. Cho, B. K. *et al.* The transcription unit architecture of the *Escherichia coli* genome. *Nature Biotech.* **27**, 1043–1049 (2009).
7. Mendoza-Vargas, A. *et al.* Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in *E. coli*. *PLoS ONE* **4**, e7526 (2009).
8. Lemmens, K. *et al.* DISTILLER: a data integration framework to reveal condition dependency of complex regulons in *Escherichia coli*. *Genome Biol.* **10**, R27 (2009).
A description of the integrative reconstruction of the *E. coli* TRN using a cross-platform expression compendium and motif information, followed by experimental validation of the predicted network.
9. Zare, H., Sangurdekar, D., Srivastava, P., Kaveh, M. & Khodursky, A. Reconstruction of *Escherichia coli* transcriptional regulatory networks via regulon-based associations. *BMC Syst. Biol.* **3**, 39 (2009).
10. Kohanski, M. A., Dwyer, D. J., Wierzbowski, J., Cottarel, G. & Collins, J. J. Mistranslation of membrane proteins and two-component system activation trigger antibiotic-mediated cell death. *Cell* **135**, 679–690 (2008).
11. Yoon, H., McDermott, J. E., Porwollik, S., McClelland, M. & Heffron, F. Coordinated regulation of virulence during systemic infection of *Salmonella enterica* serovar Typhimurium. *PLoS Pathog.* **5**, e1000306 (2009).
12. Bonneau, R. *et al.* A predictive model for transcriptional control of physiology in a free living cell. *Cell* **131**, 1354–1365 (2007).
An example of the use of an integrated computational–experimental approach to chart the regulatory network of a largely uncharacterized archaeon, including experimental validation of the predicted network.
13. Bansal, M., Belcastro, V., Ambesi-Impiombato, A. & di Bernardo, D. How to infer gene networks from expression profiles. *Mol. Syst. Biol.* **3**, 78 (2007).
14. Bonneau, R. Learning biological networks: from modules to dynamics. *Nature Chem. Biol.* **4**, 658–664 (2008).
15. Karlebach, G. & Shamir, R. Modelling and analysis of gene regulatory networks. *Nature Rev. Mol. Cell Biol.* **9**, 770–780 (2008).
16. Babu, M. M. & Teichmann, S. A. Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res.* **31**, 1234–1244 (2003).
17. Draghici, S., Khatri, P., Eklund, A. C. & Szallasi, Z. Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet.* **22**, 101–109 (2006).
18. Marshall, E. Getting the noise out of gene arrays. *Science* **306**, 630–631 (2004).
19. Johnson, D. S. *et al.* Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Res.* **18**, 393–403 (2008).
20. Ma, H. W., Buer, J. & Zeng, A. P. Hierarchical structure and modules in the *Escherichia coli* transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinformatics* **5**, 199 (2004).
21. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).
22. Ihmels, J., Bergmann, S. & Barkai, N. Defining transcription modules using large-scale gene expression data. *Bioinformatics* **20**, 1993–2003 (2004).
23. Qi, Y. & Ge, H. Modularity and dynamics of cellular networks. *PLoS Comput. Biol.* **2**, e174 (2006).
24. Madeira, S. C. & Oliveira, A. L. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **1**, 24–45 (2004).
25. Bonneau, R. *et al.* The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo*. *Genome Biol.* **7**, R36 (2006).
26. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genet.* **34**, 166–176 (2003).
Pioneering work introducing module-based network inference.
27. Margolin, A. A. *et al.* ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7**, S7 (2006).
28. Basso, K. *et al.* Reverse engineering of regulatory networks in human B cells. *Nature Genet.* **37**, 382–390 (2005).
29. Michoel, T., De Smet, R., Joshi, A., Van de Peer, Y. & Marchal, K. Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. *BMC Syst. Biol.* **3**, 49 (2009).

30. Ernst, J. *et al.* A semi-supervised method for predicting transcription factor–gene interactions in *Escherichia coli*. *PLoS Comput. Biol.* **4**, e1000044 (2008).
The first integrative reconstruction of the *E. coli* TRN using a supervised method, combining motif information and the expression compendium from reference 31.
31. Mordelet, F. & Vert, J. P. SIRENE: supervised inference of regulatory networks. *Bioinformatics* **24**, i76–i82 (2008).
32. Faith, J. J. *et al.* Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* **5**, e8 (2007).
The first global reconstruction of the *E. coli* TRN from an Affymetrix gene expression compendium, along with experimental validation of the predicted network.
33. Foster, J. W. *Escherichia coli* acid resistance: tales of an amateur acidophile. *Nature Rev. Microbiol.* **2**, 898–907 (2004).
34. Joshi, A., De Smet, R., Marchal, K., Van de Peer, Y. & Michoel, T. Module networks revisited: computational assessment and prioritization of model predictions. *Bioinformatics* **25**, 490–496 (2009).
35. Anastassiou, D. Computational analysis of the synergy among multiple interacting genes. *Mol. Syst. Biol.* **3**, 83 (2007).
36. Watkinson, J., Liang, K. C., Wang, X., Zheng, T. & Anastassiou, D. Inference of regulatory gene interactions from expression data using three-way mutual information. *Ann. NY Acad. Sci.* **1158**, 302–313 (2009).
37. Shaw, O. J., Harwood, C., Steggles, L. J. & Wipat, A. SARGE: a tool for creation of putative genetic networks. *Bioinformatics* **20**, 3638–3640 (2004).
38. Schmitt, W. A. Jr, Raab, R. M. & Stephanopoulos, G. Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data. *Genome Res.* **14**, 1654–1663 (2004).
39. Gutierrez-Rios, R. M. *et al.* Regulatory network of *Escherichia coli*: consistency between literature knowledge and microarray profiles. *Genome Res.* **13**, 2435–2443 (2003).
40. Herrgard, M. J., Covert, M. W. & Palsson, B. O. Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Res.* **13**, 2423–2434 (2003).
An informative study illustrating the limitations of expression-based network inference for *E. coli* and *S. cerevisiae*.
41. Bar-Joseph, Z. *et al.* Computational discovery of gene modules and regulatory networks. *Nature Biotech.* **21**, 1337–1342 (2003).
The first large-scale integration of ChIP-chip and expression data, applied to yeast (including experimental validation).
42. Lemmens, K. *et al.* Inferring transcriptional modules from ChIP-chip, motif and microarray data. *Genome Biol.* **7**, R37 (2006).
43. Sabatti, C. & James, G. M. Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics* **22**, 739–746 (2006).
44. Tanay, A., Sharan, R., Kupiec, M. & Shamir, R. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc. Natl Acad. Sci. USA* **101**, 2981–2986 (2004).
45. Myers, C. L. & Troyanskaya, O. G. Context-sensitive data integration and prediction of biological networks. *Bioinformatics* **23**, 2322–2330 (2007).
46. Keseler, I. M. *et al.* EcoCyc: a comprehensive view of *Escherichia coli* biology. *Nucleic Acids Res.* **37**, D464–D470 (2009).
47. Reiss, D. J., Baliga, N. S. & Bonneau, R. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics.* **7**, 280 (2006).
48. Huttenhower, C. *et al.* Detailing regulatory networks through large scale data integration. *Bioinformatics* **25**, 3267–3274 (2009).
49. Freckleton, G., Lippman, S. I., Broach, J. R. & Tavazoie, S. Microarray profiling of phage-display selections for rapid mapping of transcription factor–DNA interactions. *PLoS Genet.* **5**, e1000449 (2009).
50. Butala, M., Busby, S. J. & Lee, D. J. DNA sampling: a method for probing protein binding at specific loci on bacterial chromosomes. *Nucleic Acids Res.* **37**, e37 (2009).
51. Lu, L. J., Xia, Y., Paccanaro, A., Yu, H. & Gerstein, M. Assessing the limits of genomic data integration for predicting protein networks. *Genome Res.* **15**, 945–953 (2005).
52. Sheng, Q., Moreau, Y. & De Moor, B. Biclustering microarray data by Gibbs sampling. *Bioinformatics* **19**, ii196–ii205 (2003).
53. Getz, G., Levine, E. & Domany, E. Coupled two-way clustering analysis of gene microarray data. *Proc. Natl Acad. Sci. USA* **97**, 12079–12084 (2000).
54. Tanay, A., Sharan, R. & Shamir, R. Discovering statistically significant biclusters in gene expression data. *Bioinformatics* **18**, S136–S144 (2002).
55. Lazzeroni, L. & Owen, A. Plaid models for gene expression data. *Stat. Sin.* **2**, 61–86 (2002).
56. Murali, T. M. & Kasif, S. Extracting conserved gene expression motifs from gene expression data. *Pac. Symp. Biocomput.* **2003**, 77–88 (2003).
57. Cheng, Y. & Church, G. M. Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8**, 93–103 (2000).
58. Ben-Dor, A., Chor, B., Karp, R. & Yakhini, Z. Discovering local structure in gene expression data: the order-preserving submatrix problem. *J. Comput. Biol.* **10**, 373–384 (2003).
59. Kluger, Y., Basri, R., Chang, J. T. & Gerstein, M. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res.* **13**, 703–716 (2003).
60. Dhollander, T. *et al.* Query-driven module discovery in microarray data. *Bioinformatics* **23**, 2573–2580 (2007).
61. Ihmels, J. *et al.* Revealing modular organization in the yeast transcriptional network. *Nature Genet.* **31**, 370–377 (2002).
62. Zwir, I., Huang, H. & Groisman, E. A. Analysis of differentially-regulated genes within a regulatory network by GPS genome navigation. *Bioinformatics* **21**, 4073–4083 (2005).
63. Pena, J. M., Bjorkegren, J. & Tegner, J. Growing Bayesian network models of gene networks from seed genes. *Bioinformatics* **21**, ii224–ii229 (2005).
64. Gat-Viks, I. & Shamir, R. Refinement and expansion of signaling pathways: the osmotic response network in yeast. *Genome Res.* **17**, 358–367 (2007).
65. Tanay, A. & Shamir, R. Computational expansion of genetic networks. *Bioinformatics* **17**, S270–S278 (2001).
66. Honkela, A. *et al.* Model-based method for transcription factor target identification with limited data. *Proc. Natl Acad. Sci. USA* **107**, 7793–7798 (2010).
67. Zwir, I. *et al.* Dissecting the PhoP regulatory network of *Escherichia coli* and *Salmonella enterica*. *Proc. Natl Acad. Sci. USA* **102**, 2862–2867 (2005).
68. de Hoon, M. J. *et al.* Predicting gene regulation by sigma factors in *Bacillus subtilis* from genome-wide data. *Bioinformatics.* **20**, i101–i108 (2004).
69. Gama-Castro, S. *et al.* RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.* **36**, D120–D124 (2008).
70. 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.* **36**, D93–D96 (2008).
71. McDermott, J. E., Taylor, R. C., Yoon, H. & Heffron, F. Bottlenecks and hubs in inferred networks are important for virulence in *Salmonella typhimurium*. *J. Comput. Biol.* **16**, 169–180 (2009).
72. Taylor, R. C. *et al.* A network inference workflow applied to virulence-related processes in *Salmonella typhimurium*. *Ann. NY Acad. Sci.* **1158**, 143–158 (2009).
73. Fredrickson, J. K. *et al.* Towards environmental systems biology of *Shewanella*. *Nature Rev. Microbiol.* **6**, 592–603 (2008).
74. Toepel, J., McDermott, J. E., Summerfield, T. C. & Sherman, L. A. Transcriptional analysis of the unicellular, diazotrophic cyanobacterium *Cyanothece* sp. ATCC 51142 grown under short day/night cycles. *J. Phycol.* **45**, 610–620 (2009).
75. Mendes, P., Sha, W. & Ye, K. Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics* **19**, ii122–ii129 (2003).
76. Van den Bulcke, T. *et al.* SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics.* **7**, 43 (2006).
77. Van den Bulcke, T., Lemmens, K., Van de Peer, Y. & Marchal, K. Inferring transcriptional networks by mining 'omics' data. *Curr. Bioinform.* **1**, 301–331 (2006).
78. Stolovitzky, G., Monroe, D. & Califano, A. Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. *Ann. NY Acad. Sci.* **1115**, 1–22 (2007).
79. Cantone, I. *et al.* A yeast synthetic network for *in vivo* assessment of reverse-engineering and modeling approaches. *Cell* **137**, 172–181 (2009).
80. Marbach, D. *et al.* Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl Acad. Sci. USA* **107**, 6286–6291 (2010).
A discussion about the current limitations of network inference methods based on submissions to the DREAM3 *in silico* challenge.
81. Hibbs, M. A. *et al.* Directing experimental biology: a case study in mitochondrial biogenesis. *PLoS Comput. Biol.* **5**, e1000322 (2009).
82. Stolovitzky, G., Prill, R. J. & Califano, A. Lessons from the DREAM2 Challenges. *Ann. NY Acad. Sci.* **1158**, 159–195 (2009).
83. Nachman, I. & Regev, A. BRNI: modular analysis of transcriptional regulatory programs. *BMC Bioinformatics* **10**, 155 (2009).
84. Sorek, R. & Cossart, P. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nature Rev. Genet.* **11**, 9–16 (2010).
85. MacLean, D., Jones, J. D. & Studholme, D. J. Application of 'next-generation' sequencing technologies to microbial genetics. *Nature Rev. Microbiol.* **7**, 287–296 (2009).
86. Sharma, C. M. & Vogel, J. Experimental approaches for the discovery and characterization of regulatory small RNA. *Curr. Opin. Microbiol.* **12**, 536–546 (2009).
87. Coppins, R. L., Hall, K. B. & Groisman, E. A. The intricate world of riboswitches. *Curr. Opin. Microbiol.* **10**, 176–181 (2007).
88. Vora, T., Hottes, A. K. & Tavazoie, S. Protein occupancy landscape of a bacterial genome. *Mol. Cell* **35**, 247–253 (2009).
89. Madar, A., Greenfield, A., Ostrer, H., Vanden Eijnden, E. & Bonneau, R. The Inferelator 2.0: a scalable framework for reconstruction of dynamic regulatory network models. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2009**, 5448–5451 (2009).
90. Rockman, M. V. & Kruglyak, L. Genetics of global gene expression. *Nature Rev. Genet.* **7**, 862–872 (2006).
91. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nature Rev. Genet.* **10**, 184–194 (2009).
92. Cooper, T. F., Remold, S. K., Lenski, R. E. & Schneider, D. Expression profiles reveal parallel evolution of epistatic interactions involving the CRP regulon in *Escherichia coli*. *PLoS Genet.* **4**, e35 (2008).
93. Fong, S. S., Joyce, A. R. & Palsson, B. O. Parallel adaptive evolution cultures of *Escherichia coli* lead to convergent growth phenotypes with different gene expression states. *Genome Res.* **15**, 1365–1372 (2005).
94. Mitchell, A. *et al.* Adaptive prediction of environmental changes by microorganisms. *Nature* **460**, 220–224 (2009).
95. Tagkopoulos, I., Liu, Y. C. & Tavazoie, S. Predictive behavior within microbial genetic networks. *Science* **320**, 1313–1317 (2008).
96. Litvin, O., Causton, H. C., Chen, B. J. & Pe'er, D. Modularity and interactions in the genetics of gene expression. *Proc. Natl Acad. Sci. USA* **106**, 6441–6446 (2009).
97. Lee, S. I. *et al.* Learning a prior on regulatory potential from eQTL data. *PLoS Genet.* **5**, e1000358 (2009).
98. Lee, S. I., Pe'er, D., Dudley, A. M., Church, G. M. & Koller, D. Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proc. Natl Acad. Sci. USA* **103**, 14062–14067 (2006).
99. Gat-Viks, I., Meller, R., Kupiec, M. & Shamir, R. Understanding gene sequence variation in the context of transcription regulation in yeast. *PLoS Genet.* **6**, e1000800 (2010).
100. Herring, C. D. *et al.* Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. *Nature Genet.* **38**, 1406–1412 (2006).
101. Barrick, J. E. *et al.* Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* **461**, 1243–1247 (2009).

102. Conrad, T. M. *et al.* Whole-genome resequencing of *Escherichia coli* K-12 MG1655 undergoing short-term laboratory evolution in lactate minimal media reveals flexible selection of adaptive mutations. *Genome Biol.* **10**, R118 (2009).
103. Brem, R. B. & Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl Acad. Sci. USA* **102**, 1572–1577 (2005).
104. Isalan, M. *et al.* Evolvability and hierarchy in rewired bacterial gene networks. *Nature* **452**, 840–845 (2008).
105. Barrett, C. L., Kim, T. Y., Kim, H. U., Palsson, B. O. & Lee, S. Y. Systems biology as a foundation for genome-scale synthetic biology. *Curr. Opin. Biotechnol.* **17**, 488–492 (2006).
106. Joshi, A., Van, P. T., Van de Peer, Y. & Michoel, T. Characterizing regulatory path motifs in integrated networks using perturbational data. *Genome Biol.* **11**, R32 (2010).
107. Ye, C., Galbraith, S. J., Liao, J. C. & Eskin, E. Using network component analysis to dissect regulatory networks mediated by transcription factors in yeast. *PLoS Comput. Biol.* **5**, e1000311 (2009).
One of the pioneering methods that tries to explain mechanistically how genomic variations result in observed expression changes.
108. Zhu, J. *et al.* Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genet.* **40**, 854–861 (2008).
109. Hwang, D. *et al.* A data integration methodology for systems biology: experimental verification. *Proc. Natl Acad. Sci. USA* **102**, 17302–17307 (2005).
110. Lee, I., Date, S. V., Adai, A. T. & Marcotte, E. M. A probabilistic functional network of yeast genes. *Science* **306**, 1555–1558 (2004).
111. Suthram, S., Beyer, A., Karp, R. M., Eldar, Y. & Ideker, T. eQED: an efficient method for interpreting eQTL associations using protein networks. *Mol. Syst. Biol.* **4**, 162 (2008).
112. Liao, J. C. *et al.* Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl Acad. Sci. USA* **100**, 15522–15527 (2003).
113. Gardner, T. S., di Bernardo, D., Lorenz, D. & Collins, J. J. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301**, 102–105 (2003).
114. Grainger, D. C., Hurd, D., Harrison, M., Holdstock, J. & Busby, S. J. Studies of the distribution of *Escherichia coli* cAMP-receptor protein and RNA polymerase along the *E. coli* chromosome. *Proc. Natl Acad. Sci. USA* **102**, 17693–17698 (2005).
115. Grainger, D. C., Hurd, D., Goldberg, M. D. & Busby, S. J. Association of nucleoid proteins with coding and non-coding segments of the *Escherichia coli* genome. *Nucleic Acids Res.* **34**, 4642–4652 (2006).
116. Grainger, D. C., Aiba, H., Hurd, D., Browning, D. F. & Busby, S. J. Transcription factor distribution in *Escherichia coli*: studies with FNR protein. *Nucleic Acids Res.* **35**, 269–278 (2007).

Acknowledgements
We thank the anonymous reviewers as well as Y. Van de Peer and J. Vanderleyden for their useful comments on the manuscript. R.D.S. is a research assistant of the agency for Innovation by Science and Technology (IWT, Belgium). This work is further supported by the Katholieke Universiteit Leuven (GOA AMBioRICS, GOA/08/011, CoE EF/05/007, SymBioSys and CREA/08/023), by the IWT through the SBO-BioFrame project, by the Interuniversity Attraction Poles (IUAP, Belgium) (BioMaGNet grant P6/25), by the National Fund for Scientific Research (FWO, Belgium) (grant IOK-B9725-G.0329.09) and by the Human Frontier Science Program (grant HFSP-RGY0079/2007C).

Competing interests statement
The authors declare no competing financial interests.

DATABASES
Entrez Genome Project: http://www.ncbi.nlm.nih.gov/genomeprj
Bacillus subtilis | *Escherichia coli* | *Halobacterium salinarum* | *Saccharomyces cerevisiae* | *Salmonella enterica* subsp. *enterica* serovar Typhimurium | *Shewanella oneidensis*
UniProtKB: http://www.uniprot.org
AraC | BetI | Crp | Fis | Fnr | GadE | GadW | GadX | GutR | H-NS | IscR | Lrp | PhoP | PurR | RpoS

FURTHER INFORMATION
Kathleen Marchal's homepage:
http://homes.esat.kuleuven.be/~kmarchal/
EcoCyc: http://ecocyc.org
DBTBS: http://dbtbs.hgc.jp
RegulonDB: http://regulondb.ccg.unam.mx

SUPPLEMENTARY INFORMATION
See online article: S1 **(table)**

ALL LINKS ARE ACTIVE IN THE ONLINE PDF