nature chemical biology

Learning biological networks: from modules to dynamics

Richard Bonneau

Learning regulatory networks from genomics data is an important problem with applications spanning all of biology and biomedicine. Functional genomics projects offer a cost-effective means of greatly expanding the completeness of our regulatory models, and for some prokaryotic organisms they offer a means of learning accurate models that incorporate the majority of the genome. There are, however, several reasons to believe that regulatory network inference is beyond our current reach, such as (i) the combinatorics of the problem, (ii) factors we can't (or don't often) collect genome-wide measurements for and (iii) dynamics that elude cost-effective experimental designs. Recent works have demonstrated the ability to reconstruct large fractions of prokaryotic regulatory networks from compendiums of genomics data; they have also demonstrated that these global regulatory models can be used to predict the dynamics of the transcriptome. We review an overall strategy for the reconstruction of global networks based on these results in microbial systems.

All cellular processes operate as part of a genome-wide system, and an accurate model of the complete regulatory system for a given organism of interest will help us understand how cells coordinate processes as they develop and/or respond to their environment. Large fractions of all genomes are unannotated, and global regulatory models greatly aid the pursuit of the functional roles of these proteins (the function of transcription factors and regulatory RNAs are often defined primarily by their position in the regulatory network, and placing target genes in the regulatory network often gives us information about their likely cellular roles). More complete regulatory models will also facilitate the interpretation of genetic variation (genome-wide studies of genetic variation show that key changes are more likely to occur in regulatory genes). Ever-improving genomics experiments have begun to make possible the reconstruction of large numbers of regulatory relationships from the analysis of large accumulated genomics data collections (protein-DNA interactions, genome-wide mRNA, whole-genome sequencing)[1–4]. The work of large collaborative teams has in several cases resulted in functional genomics projects that integrate computational analysis, experimental designs and data visualization to form highly productive multigroup consortia. This review focuses on just one aspect of these coordinated systems biology efforts: the learning of genome-wide regulatory networks in a manner that enables prediction of unobserved cell states and modeling of dynamical regulatory responses to changes in cell state (**Fig. 1**). As new technologies enable more accurate (and cheaper) measurement of global metabolites, proteins, noncoding RNAs and post-translational modifications, many of the mathematical tools developed to learn and model transcriptional networks will prove powerful enough and general enough to incorporate these important additional informational levels.

Biological regulatory network inference shouldn't, upon first glance, be possible given our current experimental designs and the complexity, scale and number of unobserved system components. A few of the main challenges are outlined below.

First, for most organisms we have large collections of mRNA measurements, but lack coupled protein, post-translational modification and metabolite measurements. These unmeasured systems components must also be inferred (not possible in general) or subsumed into the learned model. As measurements of these quantities become available, we must adapt methods for network reconstruction to optimally use these new sources of information.

Second, in spite of dramatic advances in our ability to measure mRNA and protein levels in cells, nearly all biological systems are underdetermined with respect to the problem of regulatory network inference. Owing to the number of potential interactions between transcription factors and genes, network inference is a very data-intensive problem.

Third, biological networks operate on a wide range of temporal (from nanoseconds to multiple hour/day cycles) and spacial scales (environmental factors, multiple cell consortia and multicellular organisms).

Finally, the datasets required for global network inference are quite large, often requiring the coordinated efforts of several groups, and there are a great number of practical and social challenges associated with executing optimal experimental designs. An operating assumption built into many genomics efforts is that resulting global networks will provide a starting model that will be iteratively refined using experiments directed by early models. The implied experiment-model-repeat design cycle might involve multiple groups and present organizational challenges that are rarely met in any academic or research and development setting.

Biological considerations

Given the challenges above, any statement that suggests that genome-wide reconstruction of regulatory networks is possible might seem wildly optimistic. There are, however, several features of biological regulatory networks that help us to mitigate several of these issues, possibly allowing us to reclassify such affirmative statements as reasonably optimistic.

Biological systems have several components that are conserved in other organisms, enabling comparative analysis. Biological networks are neither random nor designed by a known process, and therefore have yet-to-be-determined design principles. Nature does provide several clues, however, via considerations of evolution. Many

Richard Bonneau is in the Biology and Courant Computer Science Department, New York University, 100 Washington Square East, 1009 Silver Center, New York, New York 10003-6688, USA.
e-mail: bonneau@nyu.edu

Published online 20 October 2008; doi:10.1038/nchembio.122

subcircuits and components of biological systems have evolved from similar systems in common ancestors (and are conserved in operation and/or construction), are under similar selective pressures (that lead to common function motifs or strategies via convergent evolution)[5], or must cooperate with conserved systems. From comparative analysis of gene sequences (combined with curation and other analysis), we can reconstruct large parts of metabolic networks, likely binding site patterns for highly conserved transcription factors, lists of putative transcription factors and so on. Several exciting recent works have demonstrated that we will also be able to gain insights into the composition and operation of co-regulated modules and subnetworks from the comparative analysis of functional genomics data[6].

Biological systems are modular. Clustering and biclustering pervade systems biology analysis and are performed for a very wide variety of reasons (where "biclustering" is condition- or cell-state-specific clustering). The main practical reasons are that biological systems are inherently modular and that grouping genes into modules dramatically reduces the effective complexity of any given dataset. If clustering or biclustering is done incorrectly, however, very little of the downstream analysis is likely to be correct. The problem is complicated by (i) the fact that many genes are active in only subsets of cell states and environmental conditions, (ii) the noise in available data and (iii) the complexity of the underlying regulatory system. When it comes to clustering, one method should not, in principle, be appropriate for all studies, as clusters will have different mechanistic meaning depending on their composition and the overall motivations of the study. Different methods should, ideally, be used to generate clusters meant to represent physical complexes, pathways or co-regulated groups. For example, if the goal is to learn regulatory networks, then a method tailored specifically to learning co-regulated modules should be used[7,8], whereas if the goal is functional complexes, other approaches should be used[9]. Although taking advantage of modularity is key to success in learning biological networks from data, it is still a tough problem and should continue to be an active area of research.

Biological systems are robust and often have reproducible responses to their environment that enable replicate measurement. A very important aspect of biological complexity is that aspects of it are reproducible. Bacteria will often mount similar global transcription changes in response to environmental changes. For example, there is some random chance that any given member of a population of *Bacillus subtilis* will sporulate under starvation conditions, but once that decision is made for any given cell, a highly reproducible cascade of transcriptional changes is executed, with reproducibility in aspects of relative concentrations of transcription factors and targets, metabolites and temporal patterning of changes. Thus, repeated measurements of these cells following genetic and environmental perturbations allow for reconstruction of much of this sporulation circuit[10]. The reproducible behavior of many subcircuits allows us to combine results from replicate measurements and derive meaningful relationships from the genome-wide responses of cell populations (especially populations synchronized by a perturbation, or entrained by passage through a common protocol).

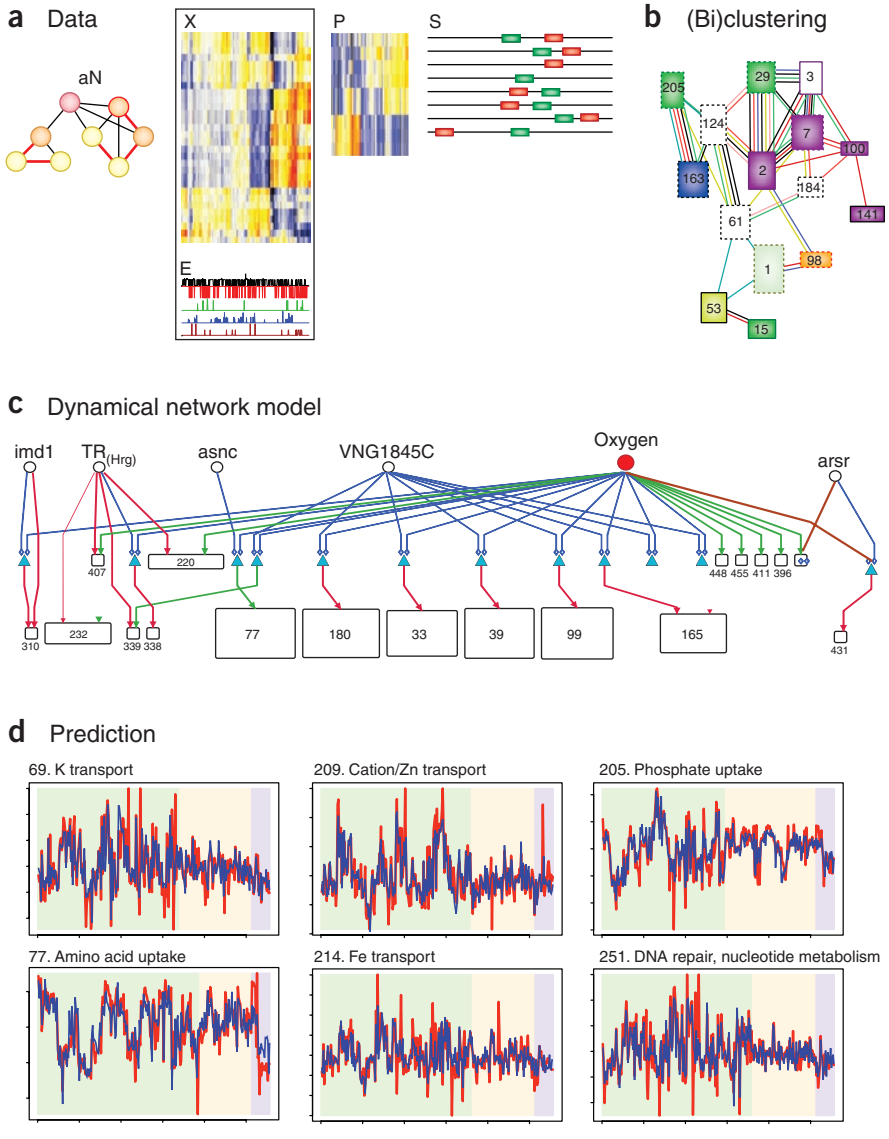

Figure 1 Network inference workflow. We show an overall schematic of the network inference strategy described. (**a**) Data available before network inference includes not just expression data (X), but often environmental parameters and experimental metadata (E), sequence elements surrounding genes (S), protein measurements (P) and association networks (aN). (**b**) A critical first step is to reduce genes to co-regulated groups; here biclusters containing multiple genes are shown color-coded by function[63] with colored lines representing relationships between biclusters. (**c**) A subnetwork resulting from network inference, with transcription factors and elongation factors at the top and bicluster targets represented as white squares below. Activation (red), repression (green) and transcription factor interactions (triangles) are shown. (**d**) Lastly, models are used to make predictions of the dynamical expression of biclusters (representing transporters, DNA repair and so on) under new conditions[63]. Predictions are made for biclusters, with numbers indicating the bicluster number (from **b**) and text indicating the functional role of the bicluster.

Biological constraints and new experiments reduce the space of possible models. There is a lot known about the likely layout of biological networks. Several network motifs are found to be over-represented in the best characterized regulatory networks, and these patterns are expected to be relevant across all organisms[5]. We also know that regulatory networks are likely to be sparse (for example, most transcription factors don't regulate most genes). Although using such prior information about regulatory motifs and topology is nontrivial, it is possible that significant performance boosts can be achieved by using this information. Experiments that directly measure protein-DNA binding are another important source of regulatory-model constraints. These experiments include genome-wide experiments such as chromatin immunoprecipitation (ChIP)-chip[11,12], ChIP-sequation[13,14] and yeast one-hybrid[15]. These experiments have significant systematic and random error, and we cannot use these constraints in a way that does not allow for nonfunctional binding, missed bindings, missed or convoluted combinatorial bindings, and spurious binding in significant quantities.

Dynamics and experimental design. A key aspect of biological networks is that they respond to changes in their environment and cell state, and they execute these responses on timescales that can, in many cases, be observed via genomic technologies. As the cost of methods for measuring mRNA, protein and other indicators continues to fall, it becomes reasonable to design experiments that capture the dynamic response of the system to a perturbation, entrainment, release from quiescence or other synchronization. Measuring the dynamics of biological systems at the global level is key to any effort to reconstruct biological networks (transcriptional, signaling, metabolic and others).

Methods for network inference

A significant body of work has been devoted to the modeling and learning of regulatory networks, and this review covers only a very narrowly focused swath of that literature (for additional reviews of this topic, see refs. 16–20). In these studies, regulatory interactions and dynamics are modeled with varying degrees of detail and model flexibility, and accordingly such models can be separated into general classes based on the level of detail with which they model individual regulatory interactions. Differential equations and stochastic models, which provide detailed descriptions of regulatory systems that include physically relevant units, can be used to simulate systems dynamics, are computationally demanding, and require accurate measurement of a large number of parameters. Several groups have also used dynamical models that scale to genome-wide analysis, but often with some limiting assumptions that restrict model complexity and form. At the other end of the model complexity spectrum lie Boolean networks, which assume that genes are simply on or off, and include standard logic interactions (and, or, xor and so on). Despite this simplification of regulatory dynamics and interactions, these approaches have the advantages of robustness and ease of interpretation[21]. Several probabilistic approaches to modeling regulatory networks on the genome-wide scale use Bayesian networks to model regulatory structure, *de novo*, at the Boolean level[22–29]. Additive linear or generalized linear models take an intermediate approach, in terms of model complexity and robustness[30–34]. Such models describe each gene's expression level as a weighted sum of the levels of its putative predictors. Given the diversity of methods for network inference, it is quite important that a regular forum be organized to evaluate the relative strengths and weaknesses of these methods; the DREAM (Dialogue for Reverse Engineering Assessments and Methods, http://wiki.c2b2.columbia.edu/dream/index.php/Main_Page) initiative provides such a venue. Several challenges are presented annually, and predictions are collected from several participating groups and evaluated, revealing the relative strengths of the often quite different approaches.

The remainder of this work will describe a possible overall strategy that includes methods for optimal design of experiments, integration of multiple datatypes to learn co-regulated modules, learning likely regulatory associations, and parameterization of a dynamical model that includes physically meaningful units, such as time.

Experimental design

Experimental design can be dramatically improved for nearly all prokaryotic functional genomic projects. The major flaws in functional genomics experimental designs are often unavoidable owing to practical reasons (optimality of experimental designs must be balanced with biological and other practical constraints). There are, however, several key principles that should be considered if global network inference is the goal.

Optimal multifactorial design. Multifactorial designs (where several perturbations are simultaneously applied to the system, often using a randomized layout) provide more information about larger fractions of the underlying system, per experiment, than single-factor designs[35,36]. Most prokaryotic functional genomics projects are, however, single-factor experiments. This is often the case for two practical reasons: (i) in many cases the perturbations are also stresses, and multifactor perturbations might be lethal and thus quite uninformative and (ii) it is often more straightforward to publish results from single-factor experiments (by focusing just on the strongest effects following perturbation). Mathematical methods for designing optimal multifactor experiments have been described, although these mathematical methods have model assumptions that might not hold for biological systems, that might be difficult to compute or that produce designs that would be too costly[37,38]. Although there are several issues that must be investigated, it is likely that (i) as the cost of experiments continues to fall, the relevance of optimal, robust and randomized designs in network inference should increase and (ii) we can do a lot better by at least considering randomized or optimal design principles when we coordinate large projects.

Sampling rate. The recipe for network inference that we espouse involves collecting mostly time series data (observing the dynamics of biological systems greatly aids our ability to reconstruct the underlying regulatory networks). The collection of time series data increases the cost of experiments by many fold, and thus one must (i) carefully balance the number of time points collected against the cost and (ii) carefully consider where along a time series to concentrate observations. The design of cost-effective time series presents several challenges: for example, a clear phenotype might emerge several hours after a perturbation (or several days into a developmental program), but the initial regulatory response governing that change might occur on a faster timescale of minutes or hours. A simple guiding principle is that one should sample on timescales comparable to the process one aims to learn about, and those samples should be more concentrated following relevant perturbations.

Sampling the sweet spot, not measuring the death response. Many perturbations (genetic or environmental) lead to either cell death or quiescent states that do not have much measurable information about the global regulatory network. Thus, it is desirable to perform cheaper experiments (such as kill curves or similar experiments to titrate a phenotype's response curve). For example, in exploring the response to toxic metals and UV stresses[39], care must be taken that samples are taken in regions of the kill curve where the majority of cells survive the stress but a clear effect is observable. Tegner and Collins note that finding this sweet spot is a

nontrivial task and that experiments to select perturbation size might be required if the system's properties are not known a priori (measurements of response curves before more expensive genomics experiments).

Taking advantage of modularity: clustering and biclustering

A natural first step in the analysis of functional genomics data is the learning of co-regulated clusters. Early methods for clustering genes assumed that genes cluster across all observed cell states (or genetic backgrounds) and that genes participate in only one cluster. Newer methods allow for genes to participate in multiple clusters and for those clusters to be condition-specific[7,40–47]. Identifying which conditions a bicluster is relevant over (in addition to the gene membership of a bicluster) is especially important in cases where genes are not expressed over significant numbers of conditions and in cases where clusters are split into multiple clusters by additional regulatory factors only active under subsets of conditions. Most biclustering algorithms can place genes into more than one cluster (genes can play more than one role). Several biclustering methods are guaranteed to converge in a reasonable time, but only after mapping the data or problem onto a new data structure or problem with significant loss of information (such as discretizing the expression data or representing expression data as a graph of co-expression), or by using methods that have ill-defined convergence criteria[7,47]. These works empirically show that the biclusters discovered are of quality sufficient to distill biologically sound conclusions, the methods converge in reasonable amounts of computer time and the biclustering results are reproducible from run to run.

Integrative biclustering. Methods for learning co-regulated conditions can also take advantage of the fact that many co-regulated groups are also cofunctional and often share detectable binding sites for transcription factors and/or regulators. For example, genes whose products form a protein complex are likely to be co-regulated. These associations can be derived either experimentally or computationally, and it is common practice to use one or more of these associations as a post facto measure of the biological quality of a gene cluster. cMonkey[7] groups genes and conditions into biclusters on the basis of (i) coherence in expression data across subsets of experimental conditions, (ii) co-occurrence of putative *cis*-acting regulatory motifs in the regulatory regions of bicluster members and (iii) the presence of highly connected subgraphs in metabolic, signaling, protein-protein and comparative genomics networks[48–50]. cMonkey identifies relevant conditions in which the genes within a given bicluster are expected to be co-regulated (importantly, in later stages of analysis we use only these conditions to learn transcription factors that influence each bicluster). The method separates the calculation of the score components associated with each data type into individual calculations but still effectively samples biclusters that optimally satisfy multiple model components (each representing a separate data type). The method was designed as a preprocessing step for network inference and performed well in comparison to all other methods tested when the trade-offs between sensitivity, specificity and coverage (fraction of conditions and genes included in one or more biclusters) were considered, particularly in the context of the other bulk characteristics (cluster size, residual). The main test was the ability to build predictive models based on biclusterings resulting from different methods; cMonkey biclusters resulted in more predictive network models when used as a first step for our full network inference pipeline.

Learning regulatory interactions

Methods for learning regulatory interactions can be separated into two groups: (i) methods that aim to learn a network of unitless regulatory influences and (ii) methods that attempt to learn regulatory networks and the dynamical parameters needed to predict transcript levels measured in new or future experiments. In general this dichotomy is not absolute, but it is a convenient construct for laying out our overall strategy. We present methods for learning associations as a first step. In many cases these algorithms begin by computing relatedness metrics such as the Pearson correlation, a rank-based relatedness metric such as Spearman's correlation, or the mutual information between a transcription factor and a possible target. The methods then search for significant interactions between transcription factors and targets based on the relatedness of the levels of a given transcription factor or elongation factor to all possible gene targets. Two such algorithms, the context likelihood relatedness (CLR)[51] algorithm and ARACNe[52], are based on a mutual information.

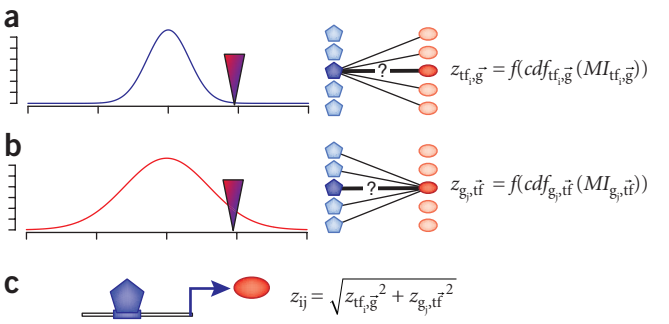

Figure 2 Context likelihood relatedness significance calculation. At the heart of the CLR algorithm is a correction that allows for the fact that different genes will have different distributions of mutual information (MI) when compared to all other genes. (**a**,**b**) Thus for a given transcription factor–gene pair, we compute the *Z*-score of the MI value for that pair compared to the distribution of all MI values for that transcription factor (**a**) and a *Z*-score for the MI value for the pair compared to the distribution for all MI values for that gene (**b**). (**c**) These two *Z*-scores are then combined to compute a *Z*-score for the transcription factor–gene interaction, thus correcting for the large differences in transcription factor and gene MI distributions.

The CLR algorithm first computes the mutual information between all transcription factors and potential gene targets. The mutual information is a commonly used information theoretic similarity measure. A key aspect of the method is that it does not assume the functional form of the interaction between any transcription factor and its target. Once a matrix of the mutual information values between transcription factors and targets is calculated, the CLR algorithm uses the full set of mutual information values to estimate a significance value for each transcription factor–gene pair (**Fig. 2**). The algorithm takes advantage of the fact that biological networks are, on average, quite sparse, and treats the majority of the mutual information values for each gene and transcription factor as insignificant or background. This background correction allows the CLR algorithm to filter out those genes that have spurious similarities with large numbers of other genes. To validate the results from their CLR algorithm, Salgado *et al.* used the RegulonDB database[53] (a set of known interactions for *Escherichia coli*). Using these known interactions, they found that at a 60% precision rate, CLR identified 1,079 interactions, of which 338 were known and 741 were putative or new. They performed *in vivo* validation using ChIP-qPCR and real-time quantitative PCR for several of the putative regulatory interactions.

Dynamical models of gene regulation

Time-lagged relatedness. The idea that explicit consideration of the dynamics of a system improves our ability to learn causation (when com-

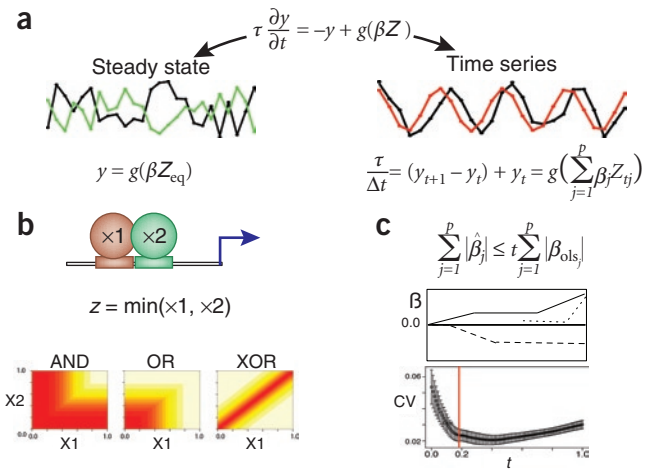

Figure 3 The Inferelator. (**a**) The core model is shown for the Inferelator algorithm; the forms of the core model for steady state data (left) and time series data (right) are shown. The algorithm can use both steady state and time series data simultaneously. (**b**) Our method for modeling interactions between transcription factors and transcription factors (as well as elongation factors and transcription factors) is shown. The minimum of an interacting pair is included in the matrix Z (the design matrix); depending on the weights learned by the Inferelator for individual (x_i) and min (x_i, x_j) terms, arbitrary interactions can be learned (including xor). (**c**) Cross-validation is used to select the shrinkage parameter, t, which controls the strength of the L1 constraint that controls model size, and prevents over-fitting[63].

pared with collections of just steady state data) is most certainly not new, and it has a rich history that precedes its application to reconstruction of biological networks[54,55]. One such early method (the correlation metric construction method, CMC) was used to reconstruct chemical reaction networks from dynamical measurements of metabolite levels[55]. The authors found that they could learn a surprising number of direct reactions and that clustering using their time-lagged correlation as a distance metric proved useful in co-clustering reactions that were adjacent in the test networks. Later works applied similar time-lagged correlation metrics to discovering regulatory relationships from microarray data[56–58].

Scalable differential equation models. Several works have described methods for learning dynamical models of regulatory networks from data that model regulatory networks as ordinary differential equations (ODEs)[59–64]. For example, Tegner *et al.* described one such method where the change in each gene's expression is modeled as a linear process[60]:

$$\frac{dx_i(t)}{dt} = P_i + \gamma_i(x_i(t) - a_i) + \sum_{j=1}^{p} w_{ij}(x_j(t) - a_j),\ i = 1,...,N \qquad (1)$$

Where x_i is the gene of interest (a separate model of this type is fit for each of the N genes), P_i is the perturbation for that gene (a given, part of the experimental design), tau is an overall time constant, and the constant W represents the couplings between gene x_i and the other regulators (X). The result is a coupled set of linear ODEs. They use their model to investigate the number of experiments one would need to collect to reconstruct networks to different levels of completeness using this type of method. Although their results were based on analysis of synthetic data, the predictions from their model largely agree with empirical and experimental results from later studies[51,63]. Gustafsson *et al.* expanded on this work by formulating a similar model (representing the regulatory dynamics of a whole cell as a set of coupled ODEs)[65]. They tested their method on the yeast *Saccharomyces cerevisiae* and showed that the use of linear representations of cellular dynamics could represent and reconstruct a surprising number of regulatory relationships.

The Inferelator[62] expanded on these earlier works and uses an ODE model for regulatory dynamics and L1 shrinkage as a means of selecting parsimonious models (**Fig. 3**). Key developments included modeling of environment and transcription factor interactions, the use of a model-fitting procedure that allows for missing data and irregular sampling intervals without the need to impute missing data, and the testing of the model on new time series collected after the publication of the initial model[2,12,39,63,66,67]. The Inferelator learns a separate sparse model for each gene or gene (bi)cluster, y, as a function of $x(t)$ by assuming that the evolution of y is governed by an equation of the form:

$$\frac{dy_i(t)}{dt} = \sum_{j=1}^{p} \beta_{ij} f_j(x_j(t)),\ i = 1,...,N \qquad (2)$$

where

$$\beta = \begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,p} \\ \beta_{2,1} & \beta_{2,2} & \cdots & \beta_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{N,1} & \beta_{N,2} & \cdots & \beta_{N,p} \end{pmatrix} \qquad (3)$$

is a set of parameters to be estimated. Interactions were incorporated into the model by allowing the function $f_j(x_j(t))$ to be either the identity function of a single variable or the minimum of two variables. These interaction terms represent transcription factor–transcription factor interactions as well as interactions between transcription factors and environmental factors. Several types of biologically relevant interactions involving two or more components can be fit by this encoding (and, xor, or). With this scheme for encoding interactions in the predictor matrix, we were able to capture many of the interactions between predictors necessary for modeling realistic regulatory networks, in a readily interpretable form. The Inferelator learns β by minimizing the following objective function, which amounts to a least square estimate based on the difference between observed data and the output of a finite difference approximation of the ODE:

$$\varepsilon_i(\beta) = \sum_{k=1}^{K} \left| \frac{y_i(t_{k+1}) - y_i(t_k)}{t_{k+1} - t_k} - \sum_{j=1}^{p} \beta_{ij} f_j(x_j(t_k)) \right|^2 \qquad (4)$$

Note that the matrix β is typically assumed to be sparse, and that the overall problem setup is compatible with L1 shrinkage (the method for enforcing model parsimony)[32,33,68]. We impose an L1 constraint on the size of β as follows:

$$\sum_{j=1}^{p} \left| \beta_{i,j} \right| = s \sum_{j=1}^{p} \left| \beta_{i,j}^{ols} \right| \qquad (5)$$

Where $\beta_{i,j}^{obs}$ the over-fit ordinary least squares estimate (that is, the minimizer of equation (4) without constraints) and s is a number between 0 and 1 referred to as the shrinkage parameter ($s = 0$ is the null model). Cross-validation is used to select the value of s that results in models with good estimated predictive performance on new data. Each resulting model is then an ODE describing the time evolution of a single bicluster or gene. The objective function (4) can be slightly modified for steady state data, thereby allowing for the simultaneous learning of networks from steady state and time series data. In order to dramatically improve the computational cost of this cross-validation procedure (to find s and β), we use least angle regression (LARS) in

the Lasso limit (which gives a result identical to that from the original Lasso code but requires much less compute time)[68].

The simplified kinetic description at the core of the original Inferelator encompasses several essential elements required to describe gene transcription, such as control by specific transcriptional activators (or repressors), activation kinetics and transcript decay, while at the same time facilitating access to computationally efficient methods for searching among an astronomically large number of possible regulators and regulator combinations. One of the major drawbacks of the infelerator is that the objective function (4) relies on an approximation of the time derivative, which may be very crude if the sampling interval (Δt) is large. The result of the procedure is a large system of ODEs that can be used to simultaneously model equilibrium and time course expression levels, such that both kinetic and equilibrium expression levels may be predicted by the resulting models.

An application of the method to *Halobacterium salinarum* resulted in a regulatory network model that could predict mRNA levels of 2,000 out of a total 2,400 genes found in the genome (a network with 1,431 predicted regulatory influences controlling 459 biclusters that represent 85% of the genes in the genome). The network model was tested (it was first trained on the 268 conditions available at the time) using 130 additional new measurements that were collected after model fitting and, in fact, after the first publication of the *H. salinarum* NRC-1 ODE network model. It was found that the prediction error over the training set was essentially the same as that over the new dataset (predicting the genome-wide transcript levels based on the levels measured in the previous time point; that is, predicting the time evolution of the transcriptome). This is encouraging, as the new data included environmental perturbations, new combinations of environmental and genetic perturbations, and time series measurements after new entrainments of the cell. Parts of the network were validated using ChIP-chip[12,63]. A similar method has also been applied to the learning of human regulatory networks mediating TLR-5–mediated stimulation of macrophages[64], and to the learning of several other microbial networks.

Future directions

Global dynamics. We can expect more explicit treatments of global regulatory dynamics in future works. Many of the works described above consider one gene or gene cluster at a time, and they involve several approximations that can be improved on (such as the finite difference approximation, $\Delta x/\Delta t$, of the true but not directly measurable derivative of genes, $\partial x/\partial t$). The Inferelator, for example, has several drawbacks associated with the assumptions it uses to simplify regulatory dynamics: (i) the dynamics are not stable over longer timescales owing to the finite difference approximation, (ii) the balance of degradation and synthesis rates is often skewed toward degradation by our method for preventing over-fitting and (iii) the approximations used in Inferelator version 1.0, although extremely efficient, preclude several attractive methods for modeling complex dynamical systems. Thus, future work needs to be done to improve the stability of dynamics over longer timescales (hour-long transitions between cell states, cell cycle, and so on).

New technologies. Technologies such as metabolomics and proteomics are continuing to change the face of systems biology. As new methods for measuring protein, metabolite and protein-modification levels become cheaper, more accurate and more ubiquitous, we will see datasets where measurements of metabolite, protein and transcript are coupled (performed on the same, or similarly treated, biological samples)[69–75]. This will enable methods, mathematically much like the ones described above, to infer the regulatory effect of protein modifications and the interactions of regulators with metabolites, noncoding RNAs and small molecules. Although most of the mathematical framework described above is in principle compatible with metabolite measurements, a few important considerations are likely to require new work. One major concern is that modeling and learning the effects of metabolites on regulation will require integration of global regulatory models (or methods for learning these global regulatory models) with current methods for modeling metabolite flux[76,77], as changes in metabolite levels will be due to changes in the balance of metabolic flux through pathways as well as regulation[78]. It is also possible that, in addition to incorporating models of metabolic fluxes and dynamics, we will have to use methods that explicitly deal with the differences in the timescales of metabolic, signaling and regulatory dynamics. Proven methods for modeling across multiple timescales (multiscale modeling) from other fields are likely to contribute to an integration of metabolic, signaling and regulatory models.

Comparative functional genomics. In most cases we have access to not just data for our organism of interest, but also functional genomics data for several related species. Some of the best described comparative methods with bearing on regulatory network inference are methods for detecting co-expressed groups conserved across multiple species[6]. These comparative functional genomics methods will improve our estimation of co-regulated modules, the analysis of promoter-region transcription factor binding sites and the performance of downstream inference procedures. We will be able to use the possibility of conserved transcription factor–target pairings and regulatory motifs as priors in our learning procedures, although work needs to be done to determine automatic data-driven methods for the use of this information. An intuitive understanding of the extreme importance of comparative functional analysis comes when considering the amazing diversity of just the microbial systems confined to well-defined environments and hosts[79]. In addition to improving network inference, new methods for learning regulatory networks from multispecies datasets are key to applying insights gained from the analysis of large model system compendiums to strains and species found in the environment or clinic.

Several experimental and computational frontiers remain, and the next decade of research is certainly going to contain a lot of surprises for those trying to solve the regulatory network reconstruction problem. As we learn prokaryotic networks based on combinations of different global measurements (that is, protein, post-translational modification, RNA, small molecule), we will then face additional levels of complexity as we reconcile our models with the variation (and new networks) found in the environment[80,81], in multispecies consortia and in host-pathogen interactions[82]. Clearly it will be some time before our dynamical models of these combined processes reach the level of completeness and accuracy we have achieved (so far) for only a few of the most well-studied prokaryotes. That said, the future is certainly bright, and we now have global regulatory models that can predict regulatory dynamics accurately enough to model the dynamical responses of several prokaryotic systems following genetic and environmental perturbations.

ACKNOWLEDGMENTS

We thank E. Vanden-Eijnden, D. Reiss, A. Madar, N. Baliga, B. Church and P. Waltman. We thank D. Shasha and the anonymous reviewers for detailed and insightful comments. R.B. is supported by the US National Science Foundation (DBI-0820757), the US Department of Energy GTL program and the US Department of Defense Computing and Society program.

Published online at http://www.nature.com/naturechemicalbiology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Ideker, T., Galitski, T. & Hood, L. A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.* **2**, 343–372 (2001).
2. Baliga, N.S. *et al.* Genomic and genetic dissection of an archaeal regulon. *Proc. Natl.*

Acad. Sci. USA **98**, 2521–2525 (2001).
3. Barrett, C.L. *et al.* Systems biology as a foundation for genome-scale synthetic biology. *Curr. Opin. Biotechnol.* **17**, 488–492 (2006).
4. Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662–1664 (2002).
5. Milo, R. *et al.* Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827 (2002).
6. Tirosh, I., Bilu, Y. & Barkai, N. Comparative biology: beyond sequence analysis. *Curr. Opin. Biotechnol.* **18**, 371–377 (2007).
7. 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).
8. Ihmels, J. *et al.* Revealing modular organization in the yeast transcriptional network. *Nat. Genet.* **31**, 370–377 (2002).
9. Gunsalus, K.C. *et al.* Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. *Nature* **436**, 861–865 (2005).
10. Eichenberger, P. *et al.* The program of gene transcription for a single differentiating cell type during sporulation in *Bacillus subtilis*. *PLoS Biol.* **2**, e328 (2004).
11. Lee, T.I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804 (2002).
12. Facciotti, M.T. *et al.* General transcription factor specified global gene regulation in archaea. *Proc. Natl. Acad. Sci. USA* **104**, 4630–4635 (2007).
13. Schmid, C.D. & Bucher, P. ChIP-Seq data reveal nucleosome architecture of human promoters. *Cell* **131**, 831–832 (2007).
14. Mardis, E.R. ChIP-seq: welcome to the new frontier. *Nat. Methods* **4**, 613–614 (2007).
15. Deplancke, B. *et al.* A gateway-compatible yeast one-hybrid system. *Genome Res.* **14**, 2093–2101 (2004).
16. De Jong, H. Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* **9**, 67–103 (2002).
17. Alm, E. & Arkin, A.P. Biological networks. *Curr. Opin. Struct. Biol.* **13**, 193–202 (2003).
18. Herrgard, M.J., Covert, M.W. & Palsson, B.O. Reconstruction of microbial transcriptional regulatory networks. *Curr. Opin. Biotechnol.* **15**, 70–77 (2004).
19. Bansal, M. *et al.* How to infer gene networks from expression profiles. *Mol. Syst. Biol.* **3**, 78 (2007).
20. Hayete, B., Gardner, T.S. & Collins, J.J. Size matters: network inference tackles the genome scale. *Mol. Syst. Biol.* **3**, 77 (2007).
21. Shmulevich, I. & Kauffman, S.A. Activities and sensitivities in boolean network models. *Phys. Rev. Lett.* **93**, 048701 (2004).
22. Friedman, N. *et al.* Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **7**, 601–620 (2000).
23. Bar-Joseph, Z. *et al.* Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.* **21**, 1337–1342 (2003).
24. Segal, E. *et al.* Rich probabilistic models for gene expression. *Bioinformatics* **17** (suppl. 1), S243–S252 (2001).
25. Segal, E., Yelensky, R. & Koller, D. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics* **19** (suppl. 1), i273–i282 (2003).
26. Stuart, J.M. *et al.* A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
27. Pe'er, D. *et al.* Inferring subnetworks from perturbed expression profiles. *Bioinformatics* **17** (suppl. 1), S215–S224 (2001).
28. Box, G.E.P. & Tiao, G.C. *Bayesian Inference in Statistical Analysis* (Wiley-Interscience, New York, 1992).
29. Pearl, J. *Causality: Models, Reasoning, and Inference* 8th ed. (Cambridge University Press, Cambridge, UK, 2001).
30. D'Haeseleer, P. *et al.* Linear modeling of mRNA expression levels during CNS development and injury. *Pac. Symp. Biocomput.* **1999**, 41–52 (1999).
31. Weaver, D.C., Workman, C.T. & Stormo, G.D. Modeling regulatory networks with weight matrices. *Pac. Symp. Biocomput.* **1999**, 112–123 (1999).
32. van Someren, E.P. *et al.* Genetic network modeling. *Pharmacogenomics* **3**, 507–525 (2002).
33. van Someren, E.P., Wessels, L.F. & Reinders, M.J. Linear modeling of genetic networks from experimental data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8**, 355–366 (2000).
34. Hastie, T., Tibshirani, R. & Friedman, J.H. *The Elements of Statistical Learning* (Springer-Verlag, New York, 2001).
35. Flaherty, P., Jordan, M.I. & Arkin, A. Robust design of biological experiments. *Proc. Neural Inf. Process. Symp.* **18**, 363–370 (2005).
36. Fisher, R.A. *Statistical Methods, Experimental Design and Scientific Inference* (Oxford University Press, Oxford, 1935).
37. Atkinson, A.C. & Donev, A.N. *Optimum Experimental Designs* (Oxford University Press, Oxford, 1992).
38. Box, G.E.P., Hunter, W.G. & Hunter, J.S. *Statistics for Experimenters* (John Wiley & Sons, New York, 1978).
39. Baliga, N.S. et al. Systems level insights into the stress response to UV radiation in the halophilic archaeon Halobacterium NRC-1. *Genome Res.* **14**, 1025–1035 (2004).
40. Tanay, A., Sharan, R. & Shamir, R. Discovering statistically significant biclusters in gene expression data. *Bioinformatics* **18** (suppl. 1), S136–S144 (2002).
41. Sheng, Q., Moreau, Y. & De Moor, B. Biclustering microarray data by Gibbs sampling. *Bioinformatics* **19** (suppl. 2), II196–II205 (2003).
42. Shamir, R. *et al.* EXPANDER–an integrative program suite for microarray data analysis. *BMC Bioinformatics* **6**, 232 (2005).
43. Prelic, A. *et al.* A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* **22**, 1122–1129 (2006).
44. Lee, H., Kong, S.W. & Park, P.J. Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes. *Bioinformatics* **24**, 889–896 (2008).
45. Kluger, Y. *et al.* Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res.* **13**, 703–716 (2003).
46. Grothaus, G.A., Mufti, A. & Murali, T.M. Automatic layout and visualization of biclusters. *Algorithms Mol. Biol.* **1**, 15 (2006).
47. Cheng, Y. & Church, G.M. Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8**, 93–103 (2000).
48. Mellor, J.C. *et al.* Predictome: a database of putative functional links between proteins. *Nucleic Acids Res.* **30**, 306–309 (2002).
49. Bowers, P.M. *et al.* Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.* **5**, R35 (2004).
50. Price, M.N. *et al.* A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res.* **33**, 880–892 (2005).
51. 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).
52. Margolin, A.A. *et al.* ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7** (suppl. 1), S7 (2006).
53. Salgado, H. *et al.* RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res.* **34**, D394–D397 (2006).
54. Vance, W., Arkin, A. & Ross, J. Determination of causal connectivities of species in reaction networks. *Proc. Natl. Acad. Sci. USA* **99**, 5816–5821 (2002).
55. Arkin, A. & Ross, J. Statistical construction of chemical reaction mechanism from measured time series. *J. Phys. Chem.* **99**, 970–979 (1995).
56. Dewey, T.G. & Galas, D.J. Dynamic models of gene expression and classification. *Funct. Integr. Genomics* **1**, 269–278 (2001).
57. Ramsey, S.A. *et al.* Uncovering a macrophage transcriptional program by integrating evidence from motif scanning and expression dynamics. *PLoS Comput. Biol.* **4**, e1000021 (2008).
58. Shi, Y., Mitchell, T. & Bar-Joseph, Z. Inferring pairwise regulatory relationships from multiple time series datasets. *Bioinformatics* **23**, 755–763 (2007).
59. Gardner, T.S. *et al.* Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301**, 102–105 (2003).
60. Tegner, J. *et al.* Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proc. Natl. Acad. Sci. USA* **100**, 5944–5949 (2003).
61. Yeung, M.K., Tegner, J. & Collins, J.J. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. USA* **99**, 6163–6168 (2002).
62. 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).
63. Bonneau, R. *et al.* A predictive model for transcriptional control of physiology in a free living cell. *Cell* **131**, 1354–1365 (2007).
64. Gilchrist, M. *et al.* Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4. *Nature* **441**, 173–178 (2006).
65. Gustafsson, M., Hornquist, M. & Lombardi, A. Constructing and analyzing a large-scale gene-to-gene regulatory network–lasso-constrained inference and biological validation. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2**, 254–261 (2005).
66. Kaur, A. *et al.* A systems view of haloarchaeal strategies to withstand stress from transition metals. *Genome Res.* **16**, 841–854 (2006).
67. Whitehead, K. *et al.* An integrated systems approach for understanding cellular responses to gamma radiation. *Mol. Syst. Biol.* **2**, 47 (2006).
68. Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. Least angle regression. *Ann. Stat.* **32**, 407–499 (2004).
69. Goo, Y.A. *et al.* Proteomic analysis of an extreme halophilic Archaeon, *Halobacterium* sp. NRC-1. *Mol. Cell. Proteomics* **2**, 506–524 (2003).
70. Gygi, S.P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**, 994–999 (1999).
71. Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–934 (2001).
72. Zhang, H. *et al.* UniPep, a database for human N-linked glycosites: a resource for biomarker discovery. *Genome Biol.* **7**, R73 (2006).
73. Hoefgen, R. & Nikiforova, V.J. Metabolomics integrated with transcriptomics: assessing systems response to sulfur-deficiency stress. *Physiol. Plant.* **132**, 190–198 (2008).
74. Weckwerth, W. Integration of metabolomics and proteomics in molecular plant physiology–coping with the complexity by data-dimensionality reduction. *Physiol. Plant.* **132**, 176–189 (2008).
75. Gomase, V.S. *et al.* Metabolomics. *Curr. Drug Metab.* **9**, 89–98 (2008).
76. Price, N.D., Reed, J.L. & Palsson, B.O. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* **2**, 886–897 (2004).
77. Reed, J.L. *et al.* An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol.* **4**, R54 (2003).
78. Covert, M.W. & Palsson, B.O. Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. *J. Biol. Chem.* **277**, 28058–28064 (2002).
79. Hunt, D.E. *et al.* Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* **320**, 1081–1085 (2008).
80. Pignatelli, M. *et al.* Metagenomics reveals our incomplete knowledge of global diversity. *Bioinformatics* **24**, 2124–2125 (2008).
81. Blow, N. Metagenomics: exploring unseen communities. *Nature* **453**, 687–690 (2008).
82. Schnappinger, D. Genomics of host-pathogen interactions. *Prog. Drug Res.* **64**, 313–343 (2007).