Construction of regulatory networks using expression time-series data of a genotyped population

Ka Yee Yeung[a], Kenneth M. Dombek[b], Kenneth Lo[a], John E. Mittler[a], Jun Zhu[c], Eric E. Schadt[d], Roger E. Bumgarner[a,1], and Adrian E. Raftery[e,1,2]

[a]Department of Microbiology, [b]Department of Biochemistry, and [e]Department of Statistics, University of Washington, Seattle, WA 98195; [c]Sage Bionetworks, Seattle, WA 98109; and [d]Pacific Biosciences, Menlo Park, CA 94025

Contributed by Adrian Raftery, October 11, 2011 (sent for review March 2, 2011)

The inference of regulatory and biochemical networks from large-scale genomics data is a basic problem in molecular biology. The goal is to generate testable hypotheses of gene-to-gene influences and subsequently to design bench experiments to confirm these network predictions. Coexpression of genes in large-scale gene-expression data implies coregulation and potential gene–gene interactions, but provide little information about the direction of influences. Here, we use both time-series data and genetics data to infer directionality of edges in regulatory networks: time-series data contain information about the chronological order of regulatory events and genetics data allow us to map DNA variations to variations at the RNA level. We generate microarray data measuring time-dependent gene-expression levels in 95 genotyped yeast segregants subjected to a drug perturbation. We develop a Bayesian model averaging regression algorithm that incorporates external information from diverse data types to infer regulatory networks from the time-series and genetics data. Our algorithm is capable of generating feedback loops. We show that our inferred network recovers existing and novel regulatory relationships. Following network construction, we generate independent microarray data on selected deletion mutants to prospectively test network predictions. We demonstrate the potential of our network to discover de novo transcription-factor binding sites. Applying our construction method to previously published data demonstrates that our method is competitive with leading network construction algorithms in the literature.

Large-scale sequencing has provided a wealth of data on the presence, absence, and variation of genes within and between species. However, functional annotation is unavailable for many genes and the majority of genes within most species are not placed within regulatory or biochemical pathways. Classic biochemical methods for placing genes in pathways cannot keep pace with the rapidly increasing amount of genomic information. To address this problem, we and others have been developing methods to infer networks from large-scale functional genomics data (1–5). The overall goals of such methods are to generate predictions of systems behavior and testable hypotheses of gene-to-gene influences. Predictions of systems behavior can be useful even in the absence of detailed mechanistic understanding. For example, the predicted response to the inhibition of a given gene can guide the selection of drug targets (6). The generation of testable hypotheses provides a path to more rapidly gain mechanistic understanding as it focuses bench experiments on subsets of potential gene-to-gene influences. Moreover, network construction and experimental work can be used in an iterative process to converge on underlying mechanisms (7, 8).

At present, the data most used in network construction methods are from large-scale gene-expression studies. Coexpression of genes across a wide variety of experimental conditions implies coregulation (9, 10) and potential gene–gene interactions. However, coexpression cannot predict the outcomes of perturbations (e.g., drug treatment or deletion) as the inferred relationships are undirected. Additional information is needed to assign directionality to edges so as to infer predictive networks. It has been shown that integrating expression data with other data types can lead to the construction of predictive networks (4, 5). Prior knowledge, such as known transcription factor (TF) gene interactions, can be used in some cases to constrain directed edges in networks, but in many systems, such knowledge is incomplete. Hence, additional global data are often needed to construct predictive networks.

One successful approach has been to use DNA variations that are correlated with given gene-expression values (expression QTLs) to infer directionality of edges in networks (4, 5, 11, 12). An alternate approach is to infer networks from time-series data. Because transcriptional regulation is a temporal process in which mRNA is transcribed continuously and new proteins are generated, time-series data can help identify the intermediate events between a given perturbation and expression responses. Time-series data can provide the chronological order of regulatory events, which also provides information about the direction of edges in networks. Time-series data can also help infer feedback loops that are ubiquitous in biology.

In this study, we generated a unique time-series gene-expression dataset from 95 genotyped yeast segregants that were subjected to a perturbation with the macrolide rapamycin. Such a dataset allows one to take advantage of both genetic variations and time dependencies to infer predictive networks. To analyze these data, we developed a Bayesian model averaging (BMA) regression-based algorithm that used a supervised framework to integrate external knowledge. We formulate network construction as a variable selection problem and aim to identify regulators for each gene. Unlike standard Bayesian networks, this method is capable of identifying feedback loops. We showed that the derived networks were enriched for known regulatory relationships that were not used as prior knowledge in network inference.

We also prospectively tested selected network predictions using data generated in our laboratory. Specifically, the child nodes of each of three selected TFs were significantly enriched with genes that responded to the deletion of the corresponding TF. In addition, we compared our method to a leading network-construction algorithm using previously published microarray data (13) and found that our method inferred a network of higher quality by some criteria.

Author contributions: K.Y.Y., J.Z., E.E.S., R.E.B., and A.E.R. designed research; K.Y.Y. and K.M.D. performed research; K.Y.Y. and A.E.R. contributed new reagents/analytic tools; K.Y.Y., K.M.D., K.L., and J.E.M. analyzed data; and K.Y.Y., R.E.B., and A.E.R. wrote the paper.

The authors declare no conflict of interest.

Data deposition: The microarray data have been deposited in ArrayExpress [accession nos. E-MTAB-412 (time series) and E-MTAB-446 (deletion validation)].

[1]R.E.B. and A.E.R. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: raftery@u.washington.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1116442108/-/DCSupplemental.

Results

Overview of Algorithm. We developed a BMA regression-based framework for inference about regulatory networks integrating external data sources. Our approach consists of two stages: (*i*) We used a supervised framework to compute prior probabilities of regulatory relationships using diverse data sources that include genetics data, genome-wide binding data, additional expression data, interaction data, and literature curation. (*ii*) For each gene, we incorporated these prior probabilities into a Bayesian variable selection method to select regulators using time-series expression data. Fig. 1 summarizes our method.

Time-Series Microarray Data for Network Construction. We measured time-dependent gene-expression levels of 95 genotyped haploid yeast segregants treated with the macrolide drug rapamycin. These segregants were constructed and genotyped by Brem et al. (13), and were derived from two genetically diverse parental yeast strains BY4716 (BY) and RM11-1a (RM). Time-series data were generated by sampling each yeast culture at 10-min intervals for up to 50 min after rapamycin addition and profiling the purified RNA with Affymetrix Yeast 2.0 microarrays. Rapamycin treatment was chosen because it is relatively easy to implement and induces widespread changes in global transcription, based on a screen of the public microarray data repositories (14–16). This perturbation allowed us to capture a large subset of all regulatory interactions encoded by the yeast genome. Before applying network construction algorithms to our time-series data, genes that did not vary much over time were filtered out. The resulting data contained expression levels for 3,556 genes at six time points for each of the 95 genotyped yeast segregants and parental strains.

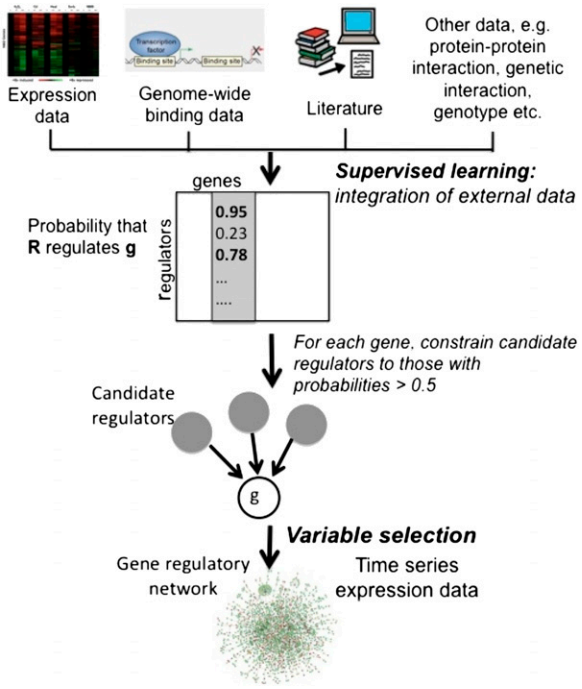

Fig. 1. Overview of our algorithm. We used a supervised framework to compute the probabilities of transcriptional regulation by integrating external data sources. Our goal is to infer the parent nodes (regulators) for each gene *g* in a regression framework. The probabilities computed in the supervised step were used to constrain potential regulators for gene *g*. Subsequently, a variable selection algorithm (iBMA) was applied to these candidate regulators using the time-series expression data. Specifically, the expression levels of putative regulators at the previous time point were used to predict the expression level of the gene of interest at the current time point.

Network Construction as a Variable Selection Problem. To build a causal network with directed edges, regulators that predict the expression level of each gene need to be identified from a large set of potential regulators. To facilitate this process, we formulated network construction as a variable selection problem. Variable selection identifies a subset of relevant features for building statistical models, and has been applied in a regression framework to the construction of regulatory networks (5, 17, 18). Here, we extended a variable selection method (BMA) to network construction. BMA takes model uncertainty into account by averaging over the posterior distributions of a quantity of interest based on multiple models, weighted by their posterior model probabilities. BMA has been shown to be effective in many different applications (19–22).

To construct regulatory networks from our time-series data, we used expression levels for selected genes at the previous time point to predict the expression levels of genes at the current time point. Our key approximations are that the regulatory relationships remain constant over time and that the expression level of gene *g* at time *t* is a linear function of the expression levels of potential regulators from the same segregant at time $(t - 1)$. To apply BMA to data with thousands of potential regulators, we used the iterative BMA (iBMA) algorithm that we previously developed for classification and survival analysis. When applied to microarray data, iBMA selected fewer genes and produced higher prediction accuracies than other methods (23, 24). Here, we extend iBMA to linear regression by rank ordering putative regulators using the coefficient of determination (R^2) from single-variable models and then iteratively applying BMA to the top ranked genes, removing variables with low posterior probabilities (see *Materials and Methods*). Other methods for implementing BMA for linear regression with high-dimensional data have also been proposed more recently (25–29).

Integration of External Data Sources. Zhu et al. (4) showed that incorporating genotypic, TF binding sites and protein–protein interaction data improved the quality of Bayesian networks. This finding prompted us to develop a supervised framework leveraging on external data sources to guide our search for the regulators in the regression framework. Our supervised framework aims to learn the relative importance (weights) of different data sources. We modeled the prior probability that a given gene is regulated by a given TF as a function of variables representing evidence of transcriptional regulation. We trained our supervised framework using known regulatory relationships from various yeast databases (30–32) as positive examples. We generated negative training examples by randomly sampling TF-gene pairs that were not documented in any of the yeast databases.

We compiled extensive data sources to create variables representing evidence of regulation, including genetics data, ChIP-chip data (33), physical interactions (34, 35), genetic interactions (36), additional expression data, and literature curation (34) (see *Materials and Methods* and Table S1). For example, we defined variables measuring coexpression of regulator-gene pairs across diverse experimental conditions, binding strength in ChIP-chip data, existence of physical or genetic interactions, numbers of common gene ontology (GO) terms, and known regulatory roles. We made use of two types of genetics data in this supervised step. We used the genotype data of the segregants (13) to determine whether sequence variations of candidate regulators correlated with expression levels of nearby genes. We also used the sequence polymorphisms of the two parental strains to determine nonsynonymous SNPs in the coding regions and the number of SNPs in the promoter regions. See *SI Materials and Methods* and Table S2 for an example of how genetics data contributed to network construction.

Assessment: Recovery of Known Regulatory Interactions. We evaluated our networks by quantifying how well the constructed network recovers regulatory relationships inferred from data

sources that were not used in the network construction process. The YEASTRACT database is a curated repository of regulatory associations between TFs and target genes in *Saccharomyces cerevisiae*, based on more than 1,200 literature references (37). Although a small subset of the interactions documented in YEASTRACT is derived from the same data sources, it represents a much larger set of regulatory interactions than those used in our supervised framework. We adopted the regulatory interactions documented in YEASTRACT as our independent assessment criterion.

We defined "direct evidence" as the number of edges in the inferred network that were supported by the independent assessment criteria (i.e., the number of recovered regulatory relationships from YEASTRACT). Sometimes it might not be possible to recover regulatory relationships in our networks because of the lack of signal in the data. If two genes have very similar expression patterns across all of the segregants, for example, it would be difficult to decide which regulates a third gene without additional information. Therefore, we defined "indirect" and "same path" as evidence capturing network inference that was proximal to the independent assessment criteria. The indirect evidence accounted for inferred regulators that were highly correlated with the known regulator (see *Materials and Methods* and Fig. S1). The "same path evidence" accounted for network inference involving an additional intermediate node than the known regulatory relationship.

We quantified associations between our network inference and independent assessment criteria using contingency tables, summarized by the precision and *P* values from the χ^2 test (Fig. S1). "Precision" is defined as the proportion of inferred edges that are supported by YEASTRACT. It measures how the inferred network edges matched the regulatory relationships documented in YEASTRACT. Pearson's χ^2 goodness-of-fit statistic is an adjusted sum of the squared differences between observed and expected frequencies, and is a classic test of association in categorical data analysis.

Networks Constructed Using our Algorithm. We applied our iBMA algorithm to the time-dependent expression profiles of yeast segregants treated with rapamycin. The inferred network (network A) consists of 3,556 nodes and 65,122 edges. The number of edges for which there is direct evidence is 662 (i.e., 662 edges in network A represent regulatory interactions documented in the YEASTRACT database). This number is 2.3 times more than would be expected if the association between YEASTRACT and our results were random. Fig. S1*C* shows the corresponding contingency table. Our assessment criteria consist of regulatory relationships with TFs and our algorithm does not constrain regulators to be known TFs (although the supervised step favors regulators with a known regulatory role). The total possible number of edges that span the subset of TFs and genes covered by YEASTRACT is 6,636, and hence: precision = 662/6,638 = 10.0%.

To assess the merits of the supervised step and the importance of the external data sources, we applied iBMA to the time-series microarray data without using any external data sources (network B). Table 1 shows the precisions and *P* values from the χ^2 test for each type of evidence. We showed that network A generally outperformed network B in terms of each type of evidence. In particular, 6,986 or 10.7% (= 6,986/65,122) of the edges in network A are supported by at least one type of evidence (i.e., union of edges supported by direct, same path and indirect evidence), compared with 4.6% (= 2,913/63,026) in network B.

As an alternative assessment strategy, we downloaded all of the binding sites documented in JASPAR (38), and computed enrichment between the gene targets containing the known binding sites upstream and the inferred child nodes of the corresponding TFs in networks A and B. Because these binding sites were not used in the supervised framework, this represents another independent assessment criterion. There are a total of 129 TFs with documented binding sites in JASPAR. Network A contained 38 TFs with enriched gene targets and network B contained only 20 such TFs (Tables S3 and S4). Consistent with YEASTRACT, the supervised framework and the external data sources provided important contributions to the accuracy of inferred networks.

Table 1. Merits of the external data sources in network construction

	Network A		Network B	
	Precision	*P* value	Precision	*P* value
Direct	10.0%	1.7×10^{-111}	8.9%	1.8×10^{-23}
Same path	6.0%	1.5×10^{-177}	5.6%	6.0×10^{-25}
Indirect	4.1%	2.5×10^{-4}	3.8%	1.1×10^{-9}

Network A (65,122 edges) was constructed using iBMA with external data sources, and network B (63,026 edges) was constructed using the time-series expression data only (without any external data sources). Precision is defined as the fraction of edges in the inferred network that are supported by regulatory interactions documented in YEASTRACT. The *P* value is derived from the χ^2 test measuring the significance of the association between our inferred network and YEASTRACT.

Prospective Validation: Independent Deletion Experiments. We generated additional independent data to prospectively validate selected network inference about the impact of one gene on other gene-expression values. We selected TFs with child nodes with high posterior probabilities that were stable with respect to bootstrapping of the data. We also selected TFs with different characteristics (e.g., numbers of citations, known binding sites, response to rapamycin over time). We selected three TFs (ARO80, DAT1, and RTG3), each of which has ~50 edges with high posterior probabilities in network A. RTG3 (YBL103C) has 83 curated references in the *Saccharomyces* Genome Database (SGD) (34), and ARO80 (YDR421W) and DAT1 (YML113W) each have under 20 curated references. ARO80 and RTG3 have known TF binding sites and increase over time in response to rapamycin. On the other hand, DAT1 has no known binding site and decreases over time in response to rapamycin. See Table S5 for a summary of these TFs.

We profiled the expression levels of the wild-type strain BY4742, which is closely related to BY4716, and each single-deletion mutant, each with three biological replicates, at 50 min after rapamycin perturbation using microarrays. We compared our network predictions to the genes that respond to the deletion in the presence of rapamycin. Specifically, we compared the child nodes of the three selected TFs in network A to the differentially expressed genes comparing each deletion mutant to the WT (Fig. 2). We observed significant overlap (adjusted *P* values <0.05) between our network predictions and the independent deletion experiments (Table 2).

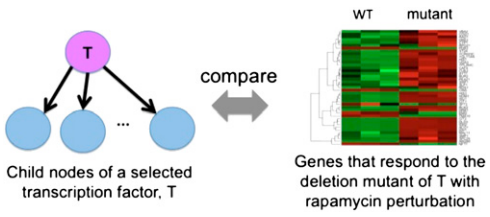

Fig. 2. Design of prospective validation experiments. We generated independent deletion data to confirm selected network predictions. Specifically, we compared the child nodes of selected transcription factors to the genes that respond to the deletion of the same transcription factor after rapamycin perturbation.

Table 2. Comparison of network inference to the independent validation experiments

TF	No. child nodes	No. genes responded to deletion	No. child nodes responded to deletion	*P* value
ARO80	51	10	4	9.3×10^{-6}
DAT1	57	784	20	0.04
RTG3	47	2,288	39	0.03

We compared the child nodes of selected transcription factors in network A to the genes that responded to the deletion of the same transcription factors in the presence of rapamycin.

Furthermore, we retrieved the known binding sites for ARO80 and RTG3 from JASPAR (38), determined the gene targets for which the promoter regions contain the known binding sites, and compared these gene targets to the child nodes of ARO80 and RTG3 in network A. Interestingly, all four genes (*ARO9*, *ARO10*, *NAF1*, and *ESBP6*) that were among the children of ARO80 in network A and responded to the deletion of ARO80 contained the known binding site of ARO80 in their promoter regions (Fig. S2). Both ARO9 and ARO10 were shown to be regulated by Aro80p (39). The regulatory roles of Aro80p on NAF1 and ESBP6 are not documented in SGD or PubGene (40), but are supported by independent ChIP-chip data (41) not used in the construction of our networks.

We repeated our analysis with RTG3, and to our surprise the overlapping genes between the child nodes of RTG3 in network A and the genes that responded to the deletion of RTG3 were not enriched with targets of RTG3's known binding site from JASPAR (*P* value = 0.31). Further investigation showed that the binding sites of both ARO80 and RTG3 in JASPAR were derived from protein binding microarray data (42). However, SGD documented an additional binding site for RTG3 (GGTCAC), determined from traditional bio-chemical methods (43). We showed that the overlapping genes between the child nodes of RTG3 in network A and the genes that responded to the deletion of RTG3 were significantly enriched for the binding site GGTCAC (*P* value = 0.01) (Fig. S3). See Table S6 for additional binding site analyses for ARO80 and RTG3.

Encouraged by the concordance between our network inference and previously determined binding sites, we identified binding sites for DAT1 that have no known binding sites in JASPAR, using computational methods. We applied MEME (Multiple Em for Motif Elicitation) (44) to the 500-bp upstream regions of the 20 overlapping genes between the child nodes of DAT1 and the genes that responded to the deletion of DAT1 in the presence of rapamycin, and obtained a highly significant motif (*e*-value = 4.5×10^{-30}) shown in Fig. 3. Furthermore, DAT1 is known to bind to poly-A sequences (45) and the poly-A sequence is the second ranked overrepresented motif from our MEME analysis.

Comparison with Leading Methods in the Literature. We compared the performance of our network-construction algorithm to a leading network construction method called Lirnet (5). Because Lirnet is designed for steady-state microarray data without any time points, we modified iBMA to be applied to a published microarray data by Brem et al. (13) (*SI Materials and Methods*). The Brem data (13) measured the steady-state expression levels of 112 yeast segregants, 95 of which were profiled in our time series microarray data. Because Lirnet constrained the inference of regulators to known TFs, we constructed network C using the same constraint. We constructed network L by applying Lirnet to the same preprocessed microarray data, the same subset of 3,152 genes, and the same external data sources used in network C. We then evaluated networks C and L using the same independent assessment criteria. Our iBMA algorithm (network C) consistently outperformed Lirnet in terms of every type of evidence (Table S7). Most strikingly, our algorithm recovered almost twice the fraction of edges supported by the independent assessment criteria (precision of direct evidence = 12.0% in network C compared with 6.8% in network L).

Discussion

We generated a microarray dataset measuring the time-dependent expression levels of 95 genotyped yeast segregants subject to an extensive perturbation. This dataset is a valuable resource for the network-construction community, as it contains both genotype data and time dependencies on a genome-wide scale. The genotype data can be used to map DNA variation to RNA variation, and the time-series data shed light on the chronological order of regulatory events. Hence, both can be used to infer the directionality of edges in regulatory networks.

We showed the usefulness of these time-series data by developing a BMA regression-based framework for the inference of regulatory networks integrating external data sources. We evaluated our inferred networks in two ways: (*i*) recovery of known regulatory relationships and (*ii*) prospective validation to confirm selected network predictions. We showed that our networks recovered many of the regulatory relationships documented in a yeast database that was not used in the construction of the networks. We showed that the supervised step (and hence, the external data resources) improved the quality of inferred networks. Because known regulatory relationships documented in databases typically span diverse experimental conditions and may not represent interactions under rapamycin perturbation, we generated additional independent data to test selected network predictions. In particular, we generated deletion data for three selected TFs (ARO80, DAT1, and RTG3) after the rapamycin perturbation. We showed that our network predictions were consistent with the independent deletion data in all three cases, and that the child nodes of ARO80 and RTG3 in our inferred network were enriched with targets of known binding sites. In addition, we found an overrepresented TF-binding site motif among the child nodes of DAT1, for which no known binding site exists in JASPAR. Applying our algorithm to a published microarray data (13), we found that our method performed better than a leading network construction algorithm in the literature.

Materials and Methods

Time-Series Microarray Data. We profiled the time-dependent expression levels of a set of 95 genomically characterized haploid yeast segregants constructed by Brem et al. (13), which were derived from two genetically diverse parental yeast strains, BY4716 and RM11-1a. Both parental strains have been sequenced. The genotype data of these yeast segregants of over ~3,000 markers are publicly available (13). A 70-mL Yeast Proteome Database (YPD) culture of each parental strain and segregant was grown to log-phase in shaken flasks at 30 °C. An aliquot of cells from each culture was taken as time point 0 and saved for RNA analysis. Then rapamycin was added to the culture

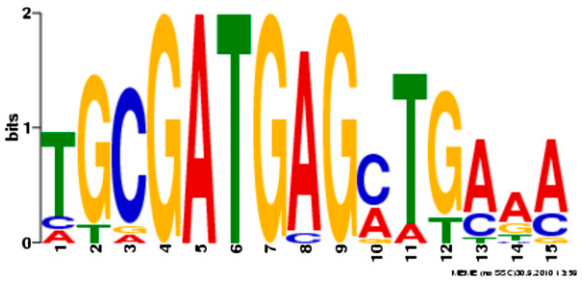

Fig. 3. Overrepresented motif of DAT1 (E-value = 4.5×10^{-30}) among the overlapping genes between the child nodes of DAT1 in network A and the genes that respond to the deletion of DAT1 in the presence of rapamycin.

at a concentration of 100 nM to induce perturbations in gene expression. Each culture was sampled at 10-min intervals after the addition for up to 50 min. Total RNA was prepared from these cell samples using RNeasy kits from Qiagen and then profiled using Yeast Genome 2.0 Arrays from Affymetrix.

The CEL files were summarized using the Robust Multiple-Array average method (46) after removing intensity data for probes whose sequence overlapped one or more sequence polymorphisms present in the RM11-1a SNP data (47) or in the RM11-1a deletion data (48). We filtered our time-series data to remove genes that do not vary much over time or across segregants, resulting in a filtered dataset consisting of 3,556 genes over six time points across 95 genotyped yeast segregants and two parental strains.

Bayesian Model Averaging. BMA takes model uncertainty into account by averaging over the posterior distributions of a quantity of interest based on multiple models, weighted by their posterior model probabilities (49, 50). Let Δ be the quantity of interest. In BMA, the posterior distribution of Δ given the data D is $\Pr(\Delta|D) = \sum_{k=1}^{K} \Pr(\Delta|D, M_k)*\Pr(M_k|D)$, where $M_1,\ldots,M_K$ are the models considered. In our context, a model is defined by a set of regulators. The reduced set of "good" models M_k for the weighted average calculations is efficiently identified using the leaps-and-bounds algorithm (51), which rapidly returns the best *nbest* models of each size up to *w* genes (19) (*w* = 30, *nbest* = 10 in our experiments). A set of parsimonious and data-supported models is then selected using the Occam's window method (52). This method consists of discarding models that are much less likely than the best model supported by the data (the default is 20-times less likely in terms of posterior model probability). Therefore, the set of "good" models used in the weighted average calculations is chosen by first applying the leaps-and-bounds algorithm, and then the Occam's window method. We used the Bayesian Information Criterion to approximate the posterior probability of a model M_k (49).

Iterative BMA for Network Construction. We formulated network construction as a variable selection problem: we modeled the expression level of gene *g* at time *t* as a linear regression of the expression levels of potential regulators at time $(t-1)$ from the same segregant. Let $X(g, t, s)$ be the expression level of gene *g* under time *t* in segregant *s*, where t = 0, 10, 20, 30, 40, 50 min and s = 1, 2, ..., 95. Mathematically, $X(g,t,s) = \beta_0 + \sum_{\text{h is a putative regulator}} \beta_h * X(h, t-1, s) + \varepsilon$, where the β_h's are regression coefficients. We applied the iBMA for network construction to select significant regulators from potentially thousands of variables and to compute regression coefficients.

In the iBMA algorithm for network construction, we ranked the variables (putative regulators for the current gene of interest) in descending order of the coefficient of determination (R^2) from fitting single-variable models. We then applied the original BMA algorithm to the *w* top-ranked genes (*w* = 30 in our experiments). Variables to which BMA assigned low posterior probabilities (<5% in our experiments) were removed. Suppose *m* variables were removed. The next *m* variables from the rank ordered R^2 were added to maintain a window of *w* variables and the original BMA was again applied. These steps of gene swaps and iterative applications of BMA were continued until we had considered all top *v* variables in our univariate ranked gene list (*v* = 100 in our experiments).

Supervised Framework for the Integration of Public Data Sources. We extracted ~550 regulatory relationships derived from non-high-throughput sources from SCPD (30), YPD (31), and TRANSFAC (32), and used these regulatory relationships as positive training examples. We generated negative training examples by randomly sampling TF-gene pairs that were not documented in any yeast databases. These positive (Y = 1) and negative (Y = 0) training examples served as response variables in our supervised learning step. We computed variables representing evidence of regulation from various yeast data resources (Table S1). Let R and G be the regulator and gene of interest. As an example, we computed variables representing the correlation coefficients between regulator R and gene G in each of three large-scale yeast gene-expression datasets: the environmental stress data (53), consisting of 225 experiments; the compendium data (54), consisting of 300 experiments; and the Stanford Microarray Database data, consisting of 671 experiments (14). As another example, another variable was equal to log(*P* value) from ChIP-chip data (33) measuring the strength of binding between R and the upstream region of gene G. We also used the functionally relevant polymorphisms collated by Lee et al. (5).

We used logistic regression to model the probability of a regulatory relationship as a function of a linear combination of these independent variables: that is, $\Pr(Y = 1) = f(\Sigma\, \alpha_i\, x_i)$, where *f* is the inverse logit function, the x_i's are independent variables and the α_is are regression coefficients. BMA for logistic regression was applied to determine the weights α_i's and the posterior probabilities of the independent variables. The estimated weights were used to compute the probabilities of regulatory relationships for all regulator-gene pairs. We used these predicted probabilities to constrain potential regulators before iBMA was applied.

Hard-Coding Known Regulatory Relationships. We hard-coded the ~550 regulatory relationships from the supervised framework in the construction of network A. We used residuals, the differences between the responses, and the fitted values to account for the effects of the known regulators. Suppose T1 and T2 are known regulators for gene *g*, and h_is are putative regulators for gene *g*. We computed the residuals of *g* on T1 and T2, resid(*g*) = residuals $[X(g,t,s) \sim X(\mathrm{T1}, t-1, s) + X(\mathrm{T2}, t-1, s)]$, and the residuals of h_j on T1 and T2, resid(h_i) = residuals $(X(h_j, t-1, s) \sim X(\mathrm{T1}, t-1, s) + X(\mathrm{T2}, t-1, s))$. We then applied iBMA using the residuals of *g* as the response and the residuals of each of the putative regulator h_i as the independent variables.

Assessment: Recovery of Existing Knowledge. The YEASTRACT database (37) was used to assess the inferred networks. To avoid bias, we removed the ~550 hard-coded regulatory relationships that were also used in the supervised training step from the assessment criteria from YEASTRACT. This process resulted in a total of 17,173 regulatory pairs, spanning 127 TFs.

Suppose our network inferred an edge R → G. Direct evidence refers to the edges for which R → G was also a regulatory relationship in the assessment criteria. Indirect evidence accounted for highly correlated genes and regulators. We called (R,G) indirect evidence if T − >G was a documented relationship from the assessment criteria, and T,R were highly correlated. In same path evidence (R → R′ → G), R′ was an intermediate node between R and G. We used a two-way contingency table to quantify the association between the inference drawn from our networks and the independent assessment criteria. See Fig. S1 for details.

ACKNOWLEDGMENTS. We thank Dr. Rachel Brem for providing us with the yeast segregants; Dr. Su-In Lee for sharing her software and external data sources; Dr. Larry Ruzzo for discussions on strongly connected components; and Kurt Hardesty, Emily Mitchell, and James Wacker for their assistance in data generation. This work was supported by Grants 5R01GM084163, 3R01GM084163-02S2, and R01 HD54511 from the National Institutes of Health, and grants from Merck.

1. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM (1999) Systematic determination of genetic network architecture. *Nat Genet* 22:281–285.
2. Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data. *J Comput Biol* 7:601–620.
3. Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 4:Article17.
4. Zhu J, et al. (2008) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat Genet* 40:854–861.
5. Lee SI, et al. (2009) Learning a prior on regulatory potential from eQTL data. *PLoS Genet* 5:e1000358.
6. Schadt EE (2009) Molecular networks as sensors and drivers of common human diseases. *Nature* 461:218–223.
7. Ideker T, et al. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292:929–934.
8. Ideker T, Galitski T, Hood L (2001) A new approach to decoding life: Systems biology. *Annu Rev Genomics Hum Genet* 2:343–372.
9. Spellman PT, et al. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* 9: 3273–3297.
10. Yeung KY, Medvedovic M, Bumgarner RE (2004) From co-expression to co-regulation: How many microarray experiments do we need? *Genome Biol* 5:R48.
11. Lee SI, Pe'er D, Dudley AM, Church GM, Koller D (2006) Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proc Natl Acad Sci USA* 103:14062–14067.
12. Schadt EE, et al. (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* 37:710–717.
13. Brem RB, Kruglyak L (2005) The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci USA* 102:1572–1577.
14. Ball CA, et al. (2005) The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Res* 33(Database issue):D580–D582.
15. Barrett T, et al. (2007) NCBI GEO: Mining tens of millions of expression profiles—Database and tools update. *Nucleic Acids Res* 35(Database issue):D760–D765.

16. Brazma A, et al. (2003) ArrayExpress—A public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* 31:68–71.
17. Jensen ST, Chen G, Stoeckert C (2007) Bayesian variable selection and data integration for biological regulatory networks. *Annals of Applied Statistics* 1:612–633.
18. James GM, Sabatti C, Zhou N, Zhu J (2010) Sparse regulatory networks. *Ann Appl Stat* 4:663–686.
19. Raftery AE (1995) Bayesian model selection in social research (with discussion). *Sociol Methodol* 25:111–193.
20. Volinsky CT, Raftery AE (2000) Bayesian information criterion for censored survival models. *Biometrics* 56:256–262.
21. Viallefont V, Raftery AE, Richardson S (2001) Variable selection and Bayesian model averaging in case-control studies. *Stat Med* 20:3215–3230.
22. Raftery AE, Zheng Y (2003) Discussion: Performance of Bayesian model averaging. *J Am Stat Assoc* 98:931–938.
23. Yeung KY, Bumgarner RE, Raftery AE (2005) Bayesian model averaging: Development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics* 21:2394–2402.
24. Annest A, Bumgarner RE, Raftery AE, Yeung KY (2009) Iterative Bayesian Model Averaging: A method for the application of survival analysis to high-dimensional microarray data. *BMC Bioinformatics* 10:72.
25. Dobra A (2009) Variable selection and dependency networks for genomewide data. *Biostatistics* 10:621–639.
26. Hans C, Dobra A, West M (2007) Shotgun stochastic search for "large p" regression. *J Am Stat Assoc* 102:507–516.
27. Bottolo L, Richardson S (2010) Evolutionary stochastic search for Bayesian model exploration. *Bayesian Anal* 5:583–618.
28. Hans C (2010) Model uncertainty and variable selection in Bayesian lasso regression. *Stat Comput* 20:221–229.
29. Tsai MY, Hsiao CK, Chen WJ (2011) Extended Bayesian model averaging in generalized linear mixed models applied to schizophrenia family data. *Ann Hum Genet* 75: 62–77.
30. Zhu J, Zhang MQ (1999) SCPD: A promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics* 15:607–611.
31. Costanzo MC, et al. (2000) The yeast proteome database (YPD) and *Caenorhabditis elegans* proteome database (WormPD): Comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res* 28: 73–76.
32. Matys V, et al. (2003) TRANSFAC: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* 31:374–378.
33. Harbison CT, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* 431:99–104.
34. Saccharomyces Genome Database. Available at http://www.yeastgenome.org/. Accessed September, 2010.
35. Stark C, et al. (2006) BioGRID: A general repository for interaction datasets. *Nucleic Acids Res* 34(Database issue):D535–D539.
36. Costanzo M, et al. (2010) The genetic landscape of a cell. *Science* 327:425–431.
37. Teixeira MC, et al. (2006) The YEASTRACT database: A tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*. *Nucleic Acids Res* 34 (Database issue):D446–D451.
38. Wasserman WW, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* 5:276–287.
39. Iraqui I, Vissers S, André B, Urrestarazu A (1999) Transcriptional induction by aromatic amino acids in *Saccharomyces cerevisiae*. *Mol Cell Biol* 19:3360–3371.
40. Jenssen TK, Laegreid A, Komorowski J, Hovig E (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet* 28:21–28.
41. Workman CT, et al. (2006) A systems approach to mapping DNA damage response pathways. *Science* 312:1054–1059.
42. Zhu C, et al. (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res* 19:556–566.
43. Jia Y, Rothermel B, Thornton J, Butow RA (1997) A basic helix-loop-helix-leucine zipper transcription complex in yeast functions in a signaling pathway from mitochondria to the nucleus. *Mol Cell Biol* 17:1110–1117.
44. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 28–36.
45. Winter E, Varshavsky A (1989) A DNA binding protein that recognizes oligo(dA).oligo (dT) tracts. *EMBO J* 8:1867–1877.
46. Irizarry RA, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4:249–264.
47. Liti G, et al. (2009) Population genomics of domestic and wild yeasts. *Nature* 458: 337–341.
48. Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L (2009) Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature* 458:342–345.
49. Kass RE, Raftery AE (1995) Bayes factors. *J Am Stat Assoc* 90:773–795.
50. Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: A tutorial. *Stat Sci* 14:382–401.
51. Furnival GM, Wilson RW (1974) Regression by leaps and bounds. *Technometrics* 16: 499–511.
52. Madigan D, Raftery AE (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *J Am Stat Assoc* 89:1335–1346.
53. Gasch AP, et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* 11:4241–4257.
54. Hughes TR, et al. (2000) Functional discovery via a compendium of expression profiles. *Cell* 102:109–126.