Reconciling Gene Expression Data With Known Genome-Scale Regulatory Network Structures

Markus J. Herrgård, Markus W. Covert, and Bernhard Ø. Palsson[1]

Department of Bioengineering, Bioinformatics Graduate Program, University of California, San Diego, La Jolla, California 92093–0412, USA

The availability of genome-scale gene expression data sets has initiated the development of methods that use this data to infer transcriptional regulatory networks. Alternatively, such regulatory network structures can be reconstructed based on annotated genome information, well-curated databases, and primary research literature. As a first step toward reconciling the two approaches, we examine the consistency between known genome-wide regulatory network structures and extensive gene expression data collections in *Escherichia coli* and *Saccharomyces cerevisiae*. By decomposing the regulatory network into a set of basic network elements, we can compute the local consistency of each instance of a particular type of network element. We find that the consistency of network elements is influenced by both structural features of the network such as the number of regulators acting on a target gene and by the functional classes of the genes involved in a particular element. Taken together, the approach presented allows us to define regulatory network subcomponents with a high degree of consistency between the network structure and gene expression data. The results suggest that targeted gene expression profiling data can be used to refine and expand particular subcomponents of known regulatory networks that are sufficiently decoupled from the rest of the network.

[Supplemental material is available online at www.genome.org.]

Reconstructing regulatory networks for model organisms has emerged as one of the central tasks in postgenomic biology (Banerjee and Zhang 2002; Stormo and Tan 2002; Wyrick and Young 2002). For well-studied organisms such as *Escherichia coli* or *Saccharomyces cerevisiae*, there is already a wealth of useful information available in databases and the primary literature that can be used to build large-scale regulatory network structures from individual regulatory interactions (RIs; Guelzim et al. 2002; Shen-Orr et al. 2002). In recent years, much excitement has been generated by high-throughput experimental techniques such as genome-wide expression profiling (DeRisi et al. 1997; Eisen et al. 1998; Hughes et al. 2000) and location analysis (Ren et al. 2000; Iyer et al. 2001; Lee et al. 2002) given the promise that these techniques have to allow for rapid reconstruction of regulatory networks.

The utilization of the high-throughput data types on their own as well as in combination with promoter sequence analysis has been shown to provide a powerful platform for regulatory network reconstruction (D'Haeseleer et al. 2000; de Jong 2002; Lee et al. 2002; Wang et al. 2002). However, the network reconstruction task is hampered by the enormous number of potential regulatory network structures that are generated and must be searched in order to identify the structure that is most consistent with the data sets. A variety of different computational frameworks has been proposed for this structural search task including Bayesian networks (Hartemink et al. 2001; Pe'er et al. 2001), combinatorial approaches (Ideker et al. 2000), and methods based on linear models (Yeung et al. 2002; Tegner et al. 2003). Alternatively, methods have been developed to identify coregulated gene modules from large-scale gene expression data (Ihmels et al. 2002; Segal et al. 2003). Despite significant progress in these data-driven reconstruction methods, the combinatorial expansion in the number of potential network structures still presents a major challenge for network reconstruction.

An approach to overcome this shortcoming of purely data-driven reconstruction methods would be to start with what are considered to be established regulatory network structures and use the new data sets to refine and expand these structures. The challenges associated with this approach are that the individual RIs comprising the known network structures have been established using a variety of different experimental methods and not all the interactions have been equally thoroughly studied. Databases representing these interactions are also likely to be incomplete and contain errors as a result of misinterpretation of results provided in experimental papers. In addition, we do not actually know which subcomponents of the known networks have been probed by high-throughput experiments and how much information these experiments provide about these subcomponents. Hence, the natural first step toward reconciling known regulatory network structures with high-throughput data sets is a comprehensive validation of these structures against the heterogeneous data sets that they can be derived from. This validation task will culminate in the identification of consistent and biologically meaningful network subcomponents that can be effectively expanded using high-throughput data sets. We have thus undertaken the present study to accomplish such validation or regulatory networks through systematic reconciliation of multiple data sources.

To our knowledge, this study is the first attempt to comprehensively evaluate the consistency between known regulatory network structures and gene expression data at the genome scale. Previously, the coherence of expression of genes in the same operon in *E. coli* has been studied (Sabatti et al. 2002), but the study was not extended to regulons or other types of network elements. In Kim et al. (2000), an approach for evaluating the agreement between regulatory network structures and gene ex-

[1]Corresponding author.
E-MAIL palsson@ucsd.edu; FAX (858) 822-3120.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1330003. Article published online before print in October 2003.

pression data based on discretized data and the perceptron algorithm is presented and applied to a small number of known transcription factor target genes in a myeloid cell line utilizing a small gene expression data set. General approaches for incorporating prior biological knowledge in the form of known regulatory network structures into reverse-engineering of regulatory networks have also been described (Hartemink et al. 2002; Chrisman et al. 2003), but they have not been applied to the large-scale data sets analyzed in this study. Previously established network structures have also been utilized in order to derive kinetic parameters for regulatory networks based on gene expression data (Ronen et al. 2002), but this approach requires much higher quality and better time resolution data that are currently commonly available.

RESULTS

The known regulatory network structures based on existing regulatory network reconstructions for *E. coli* (Shen-Orr et al. 2002) and yeast (Guelzim et al. 2002) were obtained as described in Methods. The regulatory network structures were represented as directed bipartite graphs with RI edges between a regulator node and a target gene node, with the mode of regulation (activation, repression, or dual/unknown) indicated for each interaction. The *E. coli* network has 123 regulatory genes regulating 721 target genes through 1367 RIs, whereas the yeast network has 107 regulatory genes regulating 413 target genes through 925 RIs. Because of the existence of a comprehensive database on transcriptional regulation for *E. coli*, RegulonDB (Salgado et al. 2001), the *E. coli* regulatory network is more complete and better validated than the yeast network used in this study.

We used gene expression data (1024 separate experiments for yeast, 163 experiments for *E. coli*) preprocessed and normalized as described in Methods organized into one large compendium data set for each organism. The total number of experiments or conditions in each compendium after preprocessing was 904 for yeast and 141 for *E. coli*. These data compendia include the majority of the publicly available gene expression data for both organisms. The data sets represent a large number of different experimental conditions in order to avoid a priori biasing our study toward a particular subcomponent of the regulatory network. Data produced using both oligonucleotide-array based approaches and cDNA microarrays was included in the compendium data sets. While the *E. coli* regulatory network used in this study is more complete than the yeast one, the yeast data compendium represents a more comprehensive view of the possible genomic expression programs than the *E. coli* compendium. Thus studying the consistency between established regulatory networks and gene expression data in both organisms allows us to utilize the complementary strengths of the data sets for each organism.

Computing the consistency between known regulatory network structures and gene expression data requires decomposing the networks into building blocks or elements whose local consistency with the expression data can be evaluated. There were four basic types of regulatory network building blocks analyzed in this study (Fig. 1A): (1) regulator modules (RMs), (2) multiregulator modules (MRMs), (3) RIs, and (4) target modules (TMs). A RM is defined as the set of all target genes for a single transcriptional regulator following Wang et al. (2002), who call these network elements regulatory modules. The RM also corresponds to the traditional notion of a regulon (Wagner 2000). A MRM is defined as the set of target genes that share the same set of regulators corresponding to the notion of a complex regulon. The MRM represents a group of genes that according to the knowledge encoded in the reconstructed regulatory network structure should be coregulated under all conditions. A RI is defined as a single regulator–target pair. A TM is defined as a single target gene together with all of its transcriptional regulators.

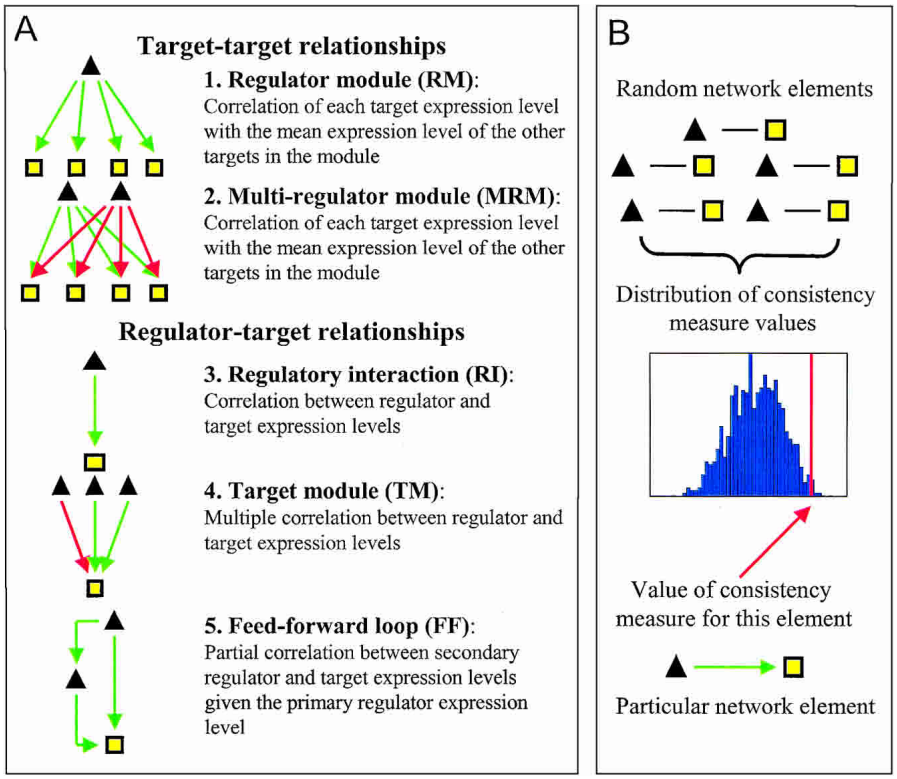

Figure 1 (*A*) Regulatory network elements or building blocks studied. The element types are classified into those involving only target–target relationships and those involving only regulator–target relationships. For each type of element, a diagram illustrating a typical element structure is shown. The black triangles are transcriptional regulators, and yellow boxes are transcription factor target genes. The red and green arrows correspond to repressing and activating regulatory interactions respectively. The consistency measure used to evaluate the agreement between a particular element and gene expression data is also indicated. Feed-forward loops are not considered to be a basic building block of the network, but are included here because ignoring them could lead to overestimating the consistency of regulatory interactions. (*B*) A schematic illustration of the strategy used to calculate the statistical significance of a particular value of a consistency measure (in this case Pearson correlation coefficient for a regulatory interaction). A set of random network elements corresponding to the particular network element studied is created, and the same consistency measure is calculated for each random element. The resulting distribution of values is used to evaluate the statistical significance of the true value of a consistency measure in the form of a *P*-value. The *P*-value is computed as the fraction of random elements with the squared consistency measure value higher than the squared observed value for the actual element. This corresponds to the estimated probability of observing a squared consistency measure value as high as the one for the actual element for a regulatory network element with random regulator and target genes.

The network elements studied in this work can also be related to the three fundamental regulatory network motifs—feed-forward loops, single-input modules, and dense overlapping regulons—that have been identified in the *E. coli* regulatory network (Shen-Orr et al. 2002). Single-input modules are a specific case of the MRM with only one regulator acting on the target genes in the module. Dense overlapping regulons correspond to combinations of the MRMs whose overall consistency could be estimated from the consistencies of the component modules. The dense overlapping regulons in both networks tend to be relatively large and do not represent a sufficiently fine-scale subdivision of the network for our purposes. Feed-forward loops were not included in the set of basic network building blocks studied in this work. However, the effect of specifically accounting for feed-forward loops was studied separately as ignoring these loops could potentially lead to overestimating the number of consistent RIs as described below.

For each instance of the four types of network building blocks present in the network, we computed a consistency measure that indicates the level of support the gene expression data provides for the particular regulatory network element structure. After testing a number of different consistency measures for each type of network element, we decided to utilize a set of measures based on correlation coefficients in order to provide a coherent framework for the study (Fig. 1A and Methods). The consistency measures were adjusted for each instance of a regulatory network building block by weighting the gene expression data sets prior to computing the consistency measure as described in Methods. Experiments or conditions particularly informative about the genes forming the particular building block were preferentially weighted. The weighting process was necessary because for any specific regulatory network element most experiments in the expression data compendia did not significantly induce or repress any of the genes involved in the element. Without any weighting of the expression profiles, the largely random background variability in these noninformative experiments would lead to significant underestimation of the number of consistent network elements. The weighted expression profile approach is similar to the one utilized in (Ihmels et al. 2002) to calculate "condition scores" for a set of potentially coregulated genes in order to identify modules of coregulated genes in the yeast genome using a large expression data compendium.

The statistical significance (reported as a *P*-value) of a particular value of a consistency measure was determined by a nonparametric randomization test based on comparing the observed value to a distribution of consistency measure values calculated for a large number of appropriate randomized network elements as described in Methods (Fig. 1B). The randomization procedure allows us to make conclusions that do not depend strongly on the particular consistency measures chosen. The fraction of regulatory network elements of a particular type with *P*-value lower than a given threshold value is used below as the actual measure of consistency for this type of element (Figs. 2, 3). While the results did not depend strongly on the *P*-value cutoff for reason-

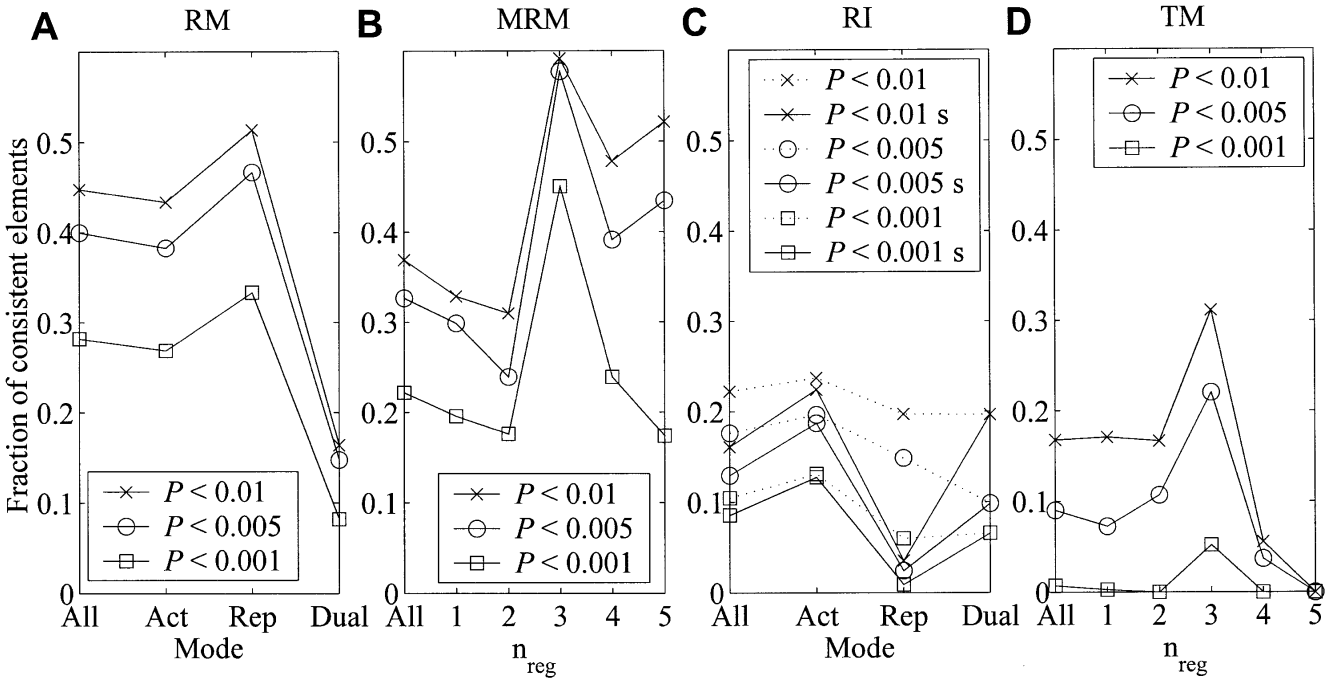

Figure 2 Summary of the results of the consistency calculations for *Escherichia coli*. The results are presented in the form of fraction of consistent network elements for three different significance levels ($P < 0.01$, $P < 0.005$, $P < 0.001$). (*A*) Results for regulator modules classified by the mode of regulation of the regulator acting on the module. All, all modules with more than one target gene (number of regulator–target pairs n = 1355); Act, activator-regulated modules (844); Rep, repressor-regulated modules (455); Dual, dual function regulator–regulated modules (61). (*B*) Results for multiregulator modules shown for all modules as well as modules with different numbers of regulators (n_{reg}). The results are only shown for cases with >10 instances of the module with the indicated number of regulators. The total number of target genes in multiregulator modules with more than one target is 658, and the numbers of target genes in multiregulator modules with different numbers of regulators are (in format: *number of regulators* [number of targets with this number of regulators]): *1* (368), *2* (142), *3* (71), *4* (46), and *5* (23). (*C*) Results for regulatory interactions classified by the mode of the regulator. The numbers of regulatory interactions in the network are: All (1367), Act (855), Rep (451), and Dual (61). In addition to the *P*-value threshold, results are also shown for consistency criteria utilizing both *P*-value threshold and sign criteria (s) as described in the text for pairwise correlations. (*D*) Results for target modules classified by the number of regulators acting on a target gene. The results are only shown for cases with >10 instances of the module with the indicated number of regulators. The numbers of target modules with different numbers of regulators are (in format: *number of regulators* [number of targets]): *1* (386), *2* (168), *3* (77), *4* (54), and *5* (25).

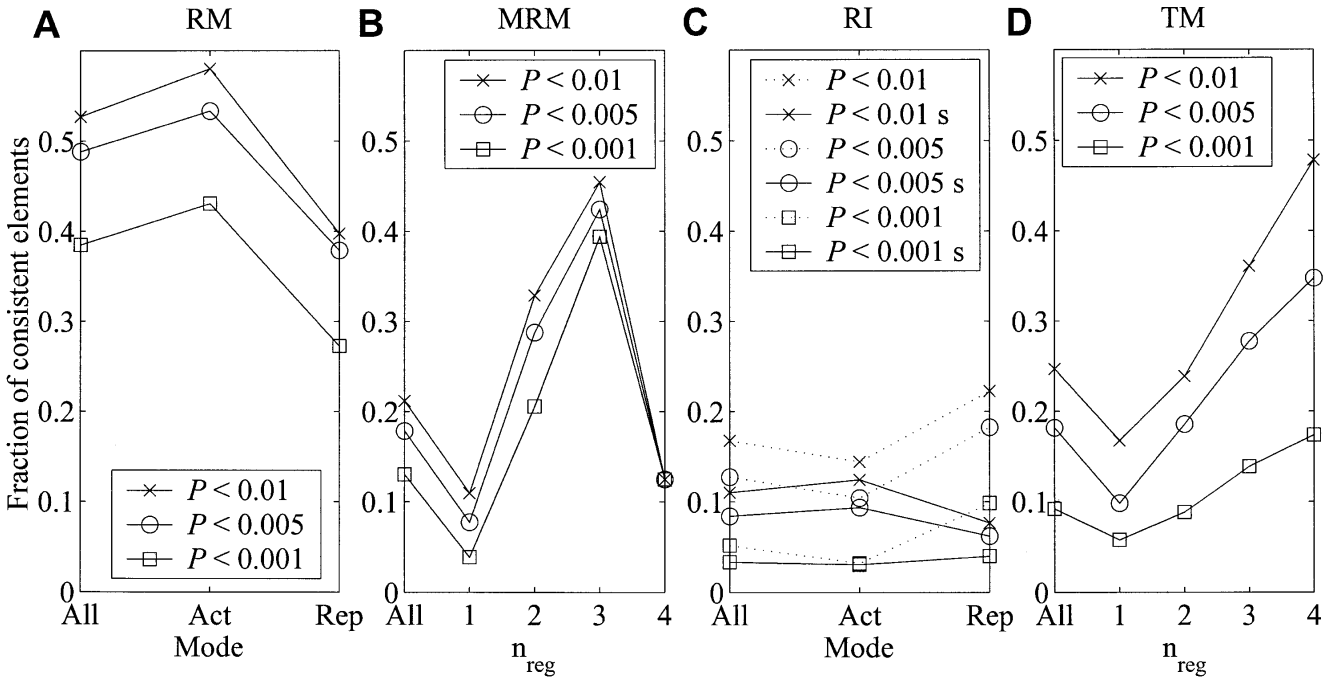

Figure 3 Summary of the results of the consistency calculations for yeast. See the caption of Figure 2 for explanations of the legends and abbreviations. (*A*) Results for regulator modules. The numbers of regulator–target pairs are: All (907), Act (643), and Rep (264). (*B*) Results for multiregulator modules. The numbers of target genes in multiregulator modules with different numbers of regulators are: All (269), *1* (155), *2* (73), *3* (33), and *4* (8). (*C*) Results for regulatory interactions. The numbers of interactions are: All (925), Act (651), and Rep (274). (*D*) Results for target modules. The numbers of target genes with different numbers of regulators are: All (413), *1* (173), *2* (113), *3* (72), and *4* (23).

able choices of cutoff values (0.001–0.05), the Supplemental Material (available online at www.genome.org) includes figures that show the dependency of the fraction of consistent network elements as a function of the *P*-value cutoff for a larger range of values than those shown in Figures 2 and 3.

The four types of basic building blocks described above can be divided into two classes—ones involving only putatively coregulated target genes (RM and MRM), and ones involving both targets and their regulators (RI and TM). The significance of this division is to separate the calculations involving target–target correlations in gene expression (RM and MRM) from calculations involving regulator–target correlations (RI and TM). Below we will discuss the results for each of the four types of regulatory network building blocks separately beginning with the first class of building blocks involving only target–target relationships. We conclude with drawing together all the results in order to identify consistent regulatory network subcomponents.

Regulator Modules

For RMs, we used the correlation R_{RM} of the weighted expression profile of each target gene in a RM with the weighted average expression profile of the other genes in the same module as a consistency measure. Because each target gene participates in as many distinct RMs as there are RIs involving the particular gene, there is a separate R_{RM} value for each interaction. It should be emphasized, however, that the regulator module consistency measure does not depend on the expression profile of the regulator that defines the module. Figures 2A (*E. coli*) and 3A (yeast) show the fractions of regulator modules consistent with gene expression data at a given *P*-value and for which the correlation R_{RM} is positive. Results for modules controlled by activating, repressing, or dual activity regulators are shown separately. The most interesting feature of the results is the different patterns observed for *E. coli* and yeast for repressor-controlled RMs—in yeast these RMs have lower degree of consistency whereas in *E. coli*, these RMs have higher degree of consistency than RMs controlled by activators. Overall, the fraction of consistent RMs is relatively high in both organisms (45% for *E. coli* and 53% for yeast at $P < 0.01$).

Multiregulator Modules

For MRMs, the consistency measure used was the same as for RMs, that is, the correlation between the weighted expression profile of each gene in the module with the mean weighted expression profile of all the other genes in the module R_{MRM}. This consistency measure indicates how coherent the expression of each target gene is with the expression of other target genes regulated by the same set of regulators. As every target gene in the network participates in exactly one MRM, the R_{MRM} value is computed separately for each target gene. The fractions of consistent target genes in MRMs at a given *P*-value and with positive correlation R_{MRM} are shown in Figures 2B (*E. coli*) and 3B (yeast) for all target genes as well as MRMs with different numbers of regulators. In both organisms, the MRMs with three regulators are the most consistent. Overall, the MRMs appear to be somewhat less consistent with gene expression data than RMs especially in yeast (37% for *E. coli* and 21% for yeast at $P < 0.01$). For RMs and MRMs, cases with only one target gene in the module were not included in the calculations presented in Figures 2 and 3.

Pairwise Regulatory Interactions

The simplest elements in the regulatory network that involve both regulators and targets are pairwise regulator–target interactions. Pearson correlation coefficients (R_{RI}) between weighted target and regulator expression profiles were used as a consistency measure for RIs. Figures 2C (*E. coli*) and 3C (yeast) show the fraction of consistent RIs classified by the mode of interaction. In

addition to the *P*-value criterion based on randomization tests (dotted lines), it was necessary to further require that for activator–target interactions, the correlation coefficient is positive and that for repressor–target interactions, the correlation coefficient is negative (solid lines). Including the sign criterion lowers the fraction of consistent repressing interactions significantly as many repressing interactions actually had a positive correlation between the regulator and target gene expression profiles. This somewhat surprising property of repressors results in overall a low fraction of consistent interactions (16% for *E. coli* and 11% for yeast for $P < 0.01$).

Target Modules

TMs allow accounting for combinatorial interactions between transcription factors acting on a target gene in the consistency calculations. The consistency metric we utilized for target modules is multiple coefficient of determination R^2_{TM} based on fitting a linear multiple regression model with the target gene expression profile as the dependent variable and its regulator expression profiles as independent variables. Figures 2D (*E. coli*) and 3D (yeast) show the fractions of consistent TMs as a function of the number of regulators in the TM. In yeast, the fraction of consistent TMs increases strongly with increasing number of regulators in the module. In *E. coli*, this tendency is less clear although the fraction of consistent TMs peaks at three regulators. In general, a larger fraction of TMs compared with RIs are consistent (17% in *E. coli* and 25% in yeast at $P < 0.01$).

Feed-Forward Loops

Analysis of pairwise interactions could overestimate correlations between transcription factor and target gene expression levels in the presence of transcriptional feed-forward loops. In such cases, two or more transcription factors act on the same gene, but some of them (primary regulators) also regulate another (secondary) regulator directly. Feed-forward loops can lead to an indirect effect by which the secondary regulator–target correlation is solely because of the influence of the primary regulators. There are 206 (*E. coli*) and 240 (yeast) secondary regulator–target gene interactions that participate in feed-forward loops so that ignoring feed-forward loops could potentially bias the results for pairwise interactions significantly. In the framework used here, the effect described above can be accounted for by replacing standard correlation coefficients with partial correlation coefficients (R_{FF}) for secondary regulator–target interactions. This approach allows for the elimination of the effect solely because of primary regulator–secondary regulator correlation from the calculations involving secondary regulator–target gene correlations. While for individual RIs, R_{FF} values were in many cases quite different from R_{RI} values, the overall effect of accounting for feed-forward loops on the fractions of consistent RIs was quite small. The percentage of consistent RIs dropped to 14% at $P < 0.01$ for *E. coli* with similar changes for other *P*-value thresholds, and for yeast, this percentage was unchanged.

Consistent Subnetworks

All the results described above for both *E. coli* and yeast can be displayed on maps of the regulatory network (Figs. 4, 5) in order to help identifying subcomponents of the regulatory networks that are consistent with the gene expression data compendiums used in this study. The results for target–target relationships (RMs and MRMs) are presented in Figures 4B (*E. coli*) and 5B (yeast), whereas the results for regulator–target relationships (RIs and TMs) are shown in Figures 4C (*E. coli*) and 5C (yeast). These figures show both the computed values of consistency metrics (widths of links and sizes of nodes) and the elements deemed to be consistent by the randomization test at $P < 0.01$ (darker color links and nodes). The network maps are also included in the Supplemental Material with the gene names indicated to allow closer inspection of the network structures.

For both organisms, there are clearly identifiable consistent subcomponents of the regulatory network involving multiple transcriptional regulators. Examples of these include regulation of amino-acid utilization and biosynthesis in yeast (A in Fig. 5A) and flagellar biosynthesis in *E. coli* (F in Fig. 4A). These subnetworks are also consistent when both target–regulator (Figs. 4C, 5C) and target–target (Figs. 4B, 5B) relationships are considered unlike most of the subcomponents identified that only have consistent target–target relationships. On the other hand, there are major subcomponents of the network where there are few consistent elements such as most components of the carbon utilization machinery in both organisms (C in Figs. 4A, 5A).

The high variability in consistency depending on the functional class of the genes involved in the particular network element evident in Figures 4 and 5 is also clearly observable when MRMs and TMs are classified by the functional class of the target gene (Tables 1, 2). For *E. coli*, the targets involved in flagellar biosynthesis (classes "motility", "flagellum", and "cellular component biosynthesis") and nucleotide biosynthesis are the most consistent considering both target–target relationships (MRMs) and regulator–target relationships (TMs). The flagellar biosynthesis genes also typically have three regulators explaining partially the observed peak in the fraction of consistent TMs and MRMs in Figure 2. However, the number of regulators does not seem to be the only determinant of consistency as the targets involved in nucleotide biosynthesis usually have only one regulator. For yeast, the target genes involved in pheromone response are the most consistent when target–target relationships are considered. However, the amino acid, nitrogen, and sulfur-metabolism related targets have most consistent regulator–target relationships. Interestingly, targets in these functional classes do not have particularly high correlation between their expression levels and the expression levels of other genes in the same MRM.

DISCUSSION

We have analyzed the consistency between known regulatory network structures and gene expression data in *E. coli* and yeast using previously reconstructed regulatory networks taken from published literature and extensive compendia of gene expression data. The subdivision of the networks into different classes of network elements allowed us to establish local consistency measures between each element structure and gene expression data. We investigated consistency both at the level of relationships between target genes (RMs and MRMs) and at the level of regulator–target relationships (RIs and TMs). Taken together, the results described above can be used to evaluate both the value of literature-derived reconstructed networks and gene expression data in fully reconstructing transcriptional regulatory networks in microbial organisms. Below, we will discuss specific determinants of consistency between regulatory network structures and gene expression data.

Difference Between Repressors and Activators

For both pairwise RIs (*E. coli* and yeast) and RMs (yeast only), we found that elements involving repressors tend to be less consistent with gene expression data than elements involving activators. This effect is most likely because for a repressor–target gene pair, either the repressor expression level is low or the target expression level is low, but it is not likely that both have high levels of expression at the same time. Because of the limited-

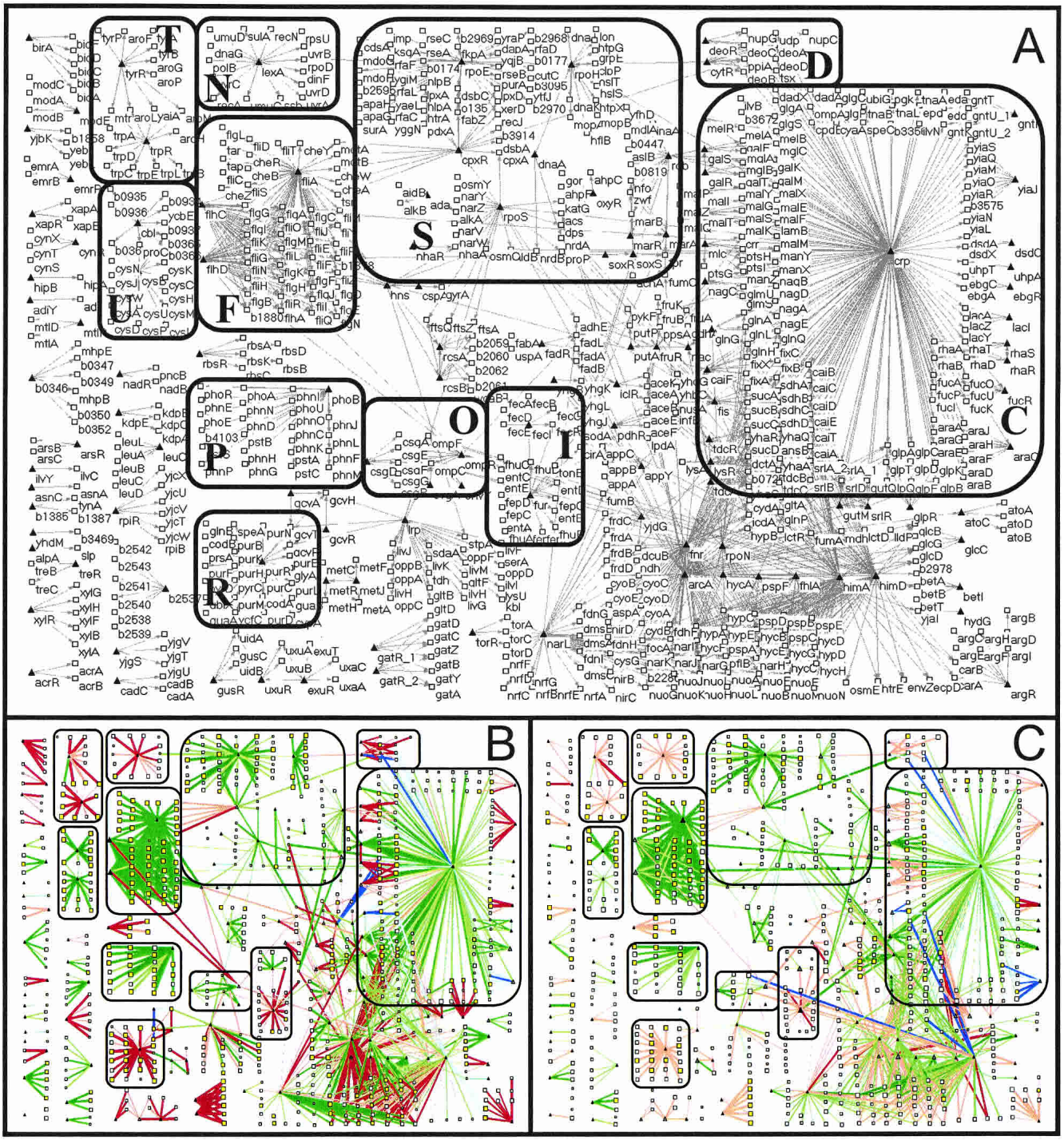

Figure 4 Maps of the *Escherichia coli* regulatory network indicating the magnitudes of the consistency metrics and consistent network elements at a *P*-value threshold of 0.01. (*A*) Network maps with each gene labeled by its name. Triangles correspond to regulators and boxes to targets. Particular subcomponents of the network are identified by labeled boxes (C, carbon utilization; D, DNA metabolism; F, flagellar biosynthesis; I, iron utilization; N, DNA damage; O, osmotic stress; P, phosphate utilization; R, purine utilization; S, stress response; T, Trp/Tyr utilization; U, sulfur utilization). (*B*) Consistent target–target relationships in the regulatory network. The widths of the links are proportional to R_{RM}^2 indicating how coherent the target gene expression is with the expression of other genes in the regulator module defined by the regulator involved in the interaction. For target genes (and regulators that are also targets of other regulators) the sizes of the nodes are proportional to R_{MRM}^2 indicating how coherent the target gene expression is with the expression of other genes in the same multiregulator module. The colors of the links indicate the mode of regulation (green, activation; red, repression; blue, dual). The darker shades of colors indicate that this component of the regulator module (for links) or multiregulator module (for nodes) is consistent with gene expression data (at $P < 0.01$). (*C*) Consistent regulator–target relationships in the regulatory network. The widths of the links are proportional to R_{RI}^2 indicating the correlation between regulator and target expression levels. The sizes of the nodes are proportional to R_{TM}^2 and describe how well the target expression level is explained by the regulator expression levels. The color scheme is the same as in *B*. The sign criterion was used to evaluate the consistency of regulatory interactions as described in the text.

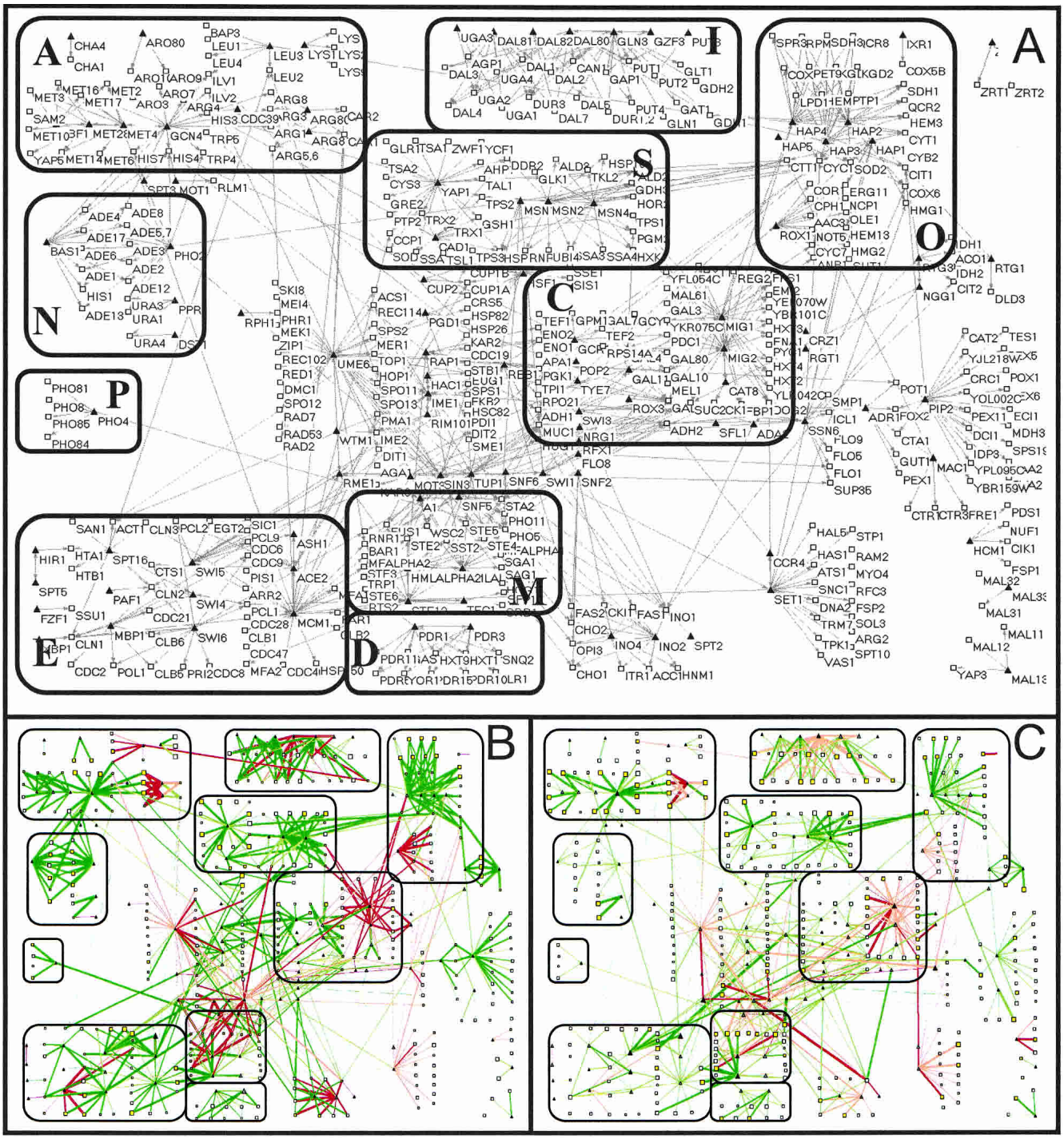

Figure 5 Maps of the yeast regulatory network indicating the magnitudes of the consistency metrics and consistent network elements at a *P*-value threshold of 0.01. See caption for Figure 4 for details. (*A*) Network maps with each gene labeled by its name. The subcomponents identified for the yeast network are (A, amino acid utilization; C, carbon utilization; D, drug response; E, cell cycle control; I, nitrogen utilization; N, nucleotide utilization; O, oxygen response; P, phosphate utilization; S, stress response). (*B*) Consistent target–target relationships. (*C*) Consistent regulator–target relationships.

dynamic range of high-throughput gene expression profiling experiments, this would result in low absolute correlation between repressor and target gene expression levels. However, in *E. coli*, the RMs with repressing interactions are more consistent than other types of regulatory modules (especially those involving transcription factors with dual/unknown activity). As repressors are more commonly utilized in prokaryotes than in eukaryotes (Struhl 1999), the higher degree of consistency for repressor-controlled regulator modules in *E. coli* may result from the preference for repression-based transcriptional regulation. In general, the development of more sensitive high- or moderate-throughput gene expression assays (Ronen et al. 2002) will potentially allow solving the problems involved in reconstructing repressor-regulated network elements.

Table 1. Fractions of Consistent Transcription Factor Target Genes in the *E. Coli* Regulatory Network Classified by the Functional Class of the Target

	MRM				TM				Number of regulators				
Functional class	**n**	***P* < 0.01**	***P* < 0.005**	***P* < 0.001**	**n**	***P* < 0.01**	***P* < 0.005**	***P* < 0.001**	**1**	**2**	**3**	**4**	**>4**
Motility	44	0.95	0.95	0.89	46	0.78	0.48	0.09	12	6	28	0	0
Flagellum	38	0.95	0.95	0.87	40	0.68	0.45	0.1	6	2	32	0	0
Nucleotide biosynthesis	18	0.83	0.83	0.72	18	0.5	0.11	0	16	0	2	0	0
Metabolism of other compounds	46	0.8	0.67	0.37	47	0.23	0.21	0	39	6	1	1	0
Cellular component biosynthesis	64	0.58	0.56	0.52	68	0.37	0.22	0.06	23	13	30	2	0
Chaperoning	28	0.57	0.57	0.39	30	0.3	0.13	0.03	17	7	3	3	0
Posttranscriptional regulation	17	0.53	0.53	0.35	19	0.37	0.21	0	12	5	1	0	1
Primary active transporters	94	0.51	0.41	0.28	95	0.14	0.07	0	62	16	0	4	13
Group translocators	16	0.38	0.38	0.38	17	0.35	0.18	0	5	5	7	0	0
Macromolecule degradation	14	0.36	0.36	0.29	15	0.07	0	0	10	4	1	0	0
Membrane components	172	0.34	0.32	0.23	186	0.14	0.08	0.01	86	52	18	14	16
Central intermediary metabolism	91	0.34	0.26	0.14	99	0.07	0.03	0	42	26	10	14	7
Energy metabolism	93	0.33	0.3	0.22	102	0.06	0.03	0	27	22	7	23	23
Carbon compound utilization	115	0.33	0.29	0.22	126	0.18	0.1	0	47	52	20	6	1
Amino acid biosynthesis	66	0.33	0.27	0.17	74	0.09	0.03	0	56	8	5	4	1
Energy production/transport	55	0.31	0.27	0.15	56	0.04	0.02	0	10	18	4	11	13
Adaptation to stress	37	0.27	0.22	0.14	46	0.11	0.09	0	31	7	6	2	0
Transcriptional regulation	42	0.26	0.24	0.12	63	0.1	0.08	0	36	15	5	6	1
Amino acid utilization	22	0.23	0.23	0.14	27	0.11	0.07	0	14	4	0	3	6
Transcription	40	0.23	0.2	0.1	60	0.08	0.07	0	36	13	4	6	1
Cofactor biosynthesis	28	0.18	0.18	0.07	30	0.03	0.03	0	24	4	2	0	0
EC potential driven transporters	36	0.17	0.11	0.08	42	0.12	0.07	0.02	17	15	4	4	2
Cell protection	25	0.08	0.08	0	31	0.19	0.03	0	24	5	1	0	1

The functional classes were obtained from the GenProtEC database (Serres and Riley 2000). Only functional classes with at least 15 genes in the network are included. Results considering both multiregulator modules (MRM) and target modules (TM) are shown. n, number of targets in this class (for MRMs this is the number of targets in this class participating in MRMs with more than one target). Fractions of consistent targets in a particular class at particular *P*-value cutoffs are shown. The number of regulator columns show the number of target genes in a functional class with the indicated number of regulators.

Correlation Between Transcription Factor and Target Gene Expression

The consistency measures used in this work for pairwise interactions, feed-forward loops, and TMs assume that some level of correlation would exist between the transcription factor and its target gene expression at least under some specific conditions—an assumption that is not necessarily, in general, true. A major reason for this lack of correlation would be that most transcription factors themselves are not significantly transcriptionally regulated and their expression remains at a low constitutive level. Instead, many transcription factors such as Mig1 glucose repressor in yeast are regulated by phosphorylation and localization as well as other posttranscriptional regulatory mechanisms (Carlson 1999). However, there is evidence from previous studies (Birnbaum et al. 2001; Zhu et al. 2002) that in many cases the correlation assumption is at least partially true, that is while the major mode of regulation of transcription factor activity is probably posttranscriptional, there is also a transcriptional component that can be utilized in analyzing the relationship between regulator and target expression. Based on these considerations, the relatively low fraction of consistent interactions can be thought to be more indicative of the fraction of transcription factors that are significantly transcriptionally regulated and are subject to only minor posttranscriptional regulation. In any case, the low fraction of consistent pairwise interactions exposes the inherent limitations in regulatory network reconstruction strategies that fundamentally rely on the correlation between transcription factor and target gene expression. These considerations do not imply that this correlation should not be utilized where it exists, as it can provide valuable additional information for regulatory network reconstruction.

How Complete Are Known Regulatory Networks?

In both organisms, we observed that in addition to depending on the mode of regulation and the functional class of the genes involved in the network element, the consistency of a particular element can also significantly depend on the structural features of the element. The most significant of these features affecting the consistency of TMs and MRMs is the number of regulators a particular target gene has. In particular, modules with only one regulator appear to be less consistent than those with more regulators. This lesser degree of consistency indicates that the single regulator modules might in fact be regulated by more than one transcription factor, but the regulatory mechanisms have not been completely characterized. In particular, yeast, like other eukaryotes, typically utilizes general transcription factors and chromatin-modifying enzymes in addition to specific transcription factors (Featherstone 2002). Because the targets of these general factors are not as well characterized as those of specific factors, the network utilized in this work is expected to contain only a small fraction of the RIs in the complete regulatory network.

Origin of Consistent Subnetworks

Some of the variability in the consistency between regulatory network structures and gene expression data appears to be a result of the types of data gene expression sets utilized in this work. For example, the response to amino-acid depletion was specifically studied in one of the yeast data sets (Gasch et al. 2000) potentially giving rise to the high degree of consistency in the amino-acid utilization subnetwork. A key determinant of the degree of consistency in a subcomponent is also the nature of the transcriptional regulators in the component. For example, in *E. coli*, the flagellar biosynthesis process is controlled by a transcrip-

Table 2. Fractions of Consistent Transcription Factor Target Genes in the Yeast Regulatory Network Classified by the Functional Class of the Target

	MRM				TM				Number of regulators				
Functional class	**n**	**$P < 0.01$**	**$P < 0.005$**	**$P < 0.001$**	**n**	**$P < 0.01$**	**$P < 0.005$**	**$P < 0.001$**	**1**	**2**	**3**	**4**	**>4**
Pheromone response	10	0.4	0.4	0.3	30	0.37	0.27	0.13	6	8	5	3	8
Detoxification	16	0.38	0.38	0.31	27	0.33	0.22	0.07	11	10	4	0	2
Nucleotide metabolism	21	0.33	0.24	0.19	25	0.28	0.2	0.2	3	16	3	3	0
DNA synthesis and replication	11	0.27	0.27	0.18	18	0.17	0.06	0	8	6	2	1	1
Amino acid metabolism	35	0.26	0.2	0.11	61	0.39	0.33	0.18	23	14	16	5	3
Carbohydrate metabolism	50	0.26	0.18	0.12	83	0.2	0.13	0.06	26	22	20	3	12
Mitotic cell cycle and cell cycle control	27	0.22	0.22	0.19	43	0.14	0.14	0.02	18	14	5	1	5
Nitrogen and sulfur metabolism	14	0.14	0.14	0.07	28	0.39	0.39	0.32	5	9	6	6	2
Cellular import	18	0.11	0.11	0.06	21	0.29	0.29	0.14	5	9	5	2	0
Stress response	18	0.11	0.11	0.11	29	0.1	0.07	0.07	12	8	5	1	3
Meiosis	20	0.1	0.1	0.1	26	0.04	0.04	0	12	7	2	1	4
mRNA synthesis	34	0.06	0.06	0.03	64	0.16	0.11	0.06	30	18	6	1	9
Lipid, fatty-acid and isoprenoid metabolism	30	0.03	0.03	0.03	39	0.13	0.13	0.05	20	8	5	4	2
DNA recombination and DNA repair	11	0	0	0	19	0.05	0.05	0	10	4	3	1	1
Sporulation and germination	8	0	0	0	17	0.12	0.06	0	5	6	2	2	2

See caption for Table 1 for details. The functional classes were obtained from the MIPS database (Mewes et al. 2002).

tional regulatory cascade involving regulators *fliDC* and *fliA* (Kalir et al. 2001) in which both regulator–target and target–target correlations in gene expression are expected. However, there are also general network structural features that appear to influence consistency. The most prominent feature is the tendency of relatively isolated subcomponents of the network such as flagellar biosynthesis and phosphate utilization in *E. coli* or amino-acid utilization in yeast to be consistent with gene expression data, whereas highly interconnected components such as carbon utilization regulation typically have a lower degree of consistency. However, not every isolated subnetwork has a large number of consistent network elements indicating that the network reconstruction may be incomplete, and these subnetworks may in fact be more strongly connected to other parts of the network than is currently known. Alternatively, the data compendiums used may not contain experiments where regulators in these subnetworks are activated strongly enough to cause changes in target gene expression detectable using current gene expression profiling technologies.

Expanding Known Regulatory Networks

Based on the results of the current study focusing on evaluating and validating the consistency between known network structures and high-throughput data sets, the next step is to utilize the information gained to expand known regulatory networks. In principle, the problem of expanding known regulatory networks by utilizing gene expression and location analysis data could be cast as a supervised data mining problem, where known RIs would be used to train a prediction algorithm that would predict RIs from the available data. However, the limited sizes of the training sets available for each TM/RM (only a few interactions or genes) would make training extremely challenging. The most appropriate approach would then seem to be a semisupervised approach, where the known RIs would be used to seed an iterative process to add new members to TMs/RMs in an otherwise unsupervised manner. The method described in Ihmels et al. (2002) implements this type of approach but does not specifically initialize the process with known network structures. Furthermore, this method would potentially benefit from utilizing regulator–target relationships in cases where these relationships would provide additional useful constraints for network reconstruction. The type of approach described above could also readily include other types of information such as location analysis (Lee et al. 2002) and promoter sequence data (Wang et al. 2002) to form a principled basis for network expansion.

Conclusion

Taken together, the results shown here indicate that combining information on known regulatory network structures with gene expression data is a productive way to refine and expand regulatory networks structures. The results show that different features of the network structure influence consistency. In particular, we observe that network elements involving repressors are typically less consistent than those involving activators indicating that reconstruction of these types of network components would pose a challenge. We also find that gene expression data provide much better support overall for target–target relationships (RMs and MRMs) than for regulator–target relationships (RIs and TMs). This result shows that a clustering-like approach to analyzing gene expression data by grouping target genes with similar expression patterns should indeed be successful in at least partially reconstructing individual RMs and MRMs. The observed increase in the fraction of consistent elements as a function of the number of regulators involved in the element can be interpreted as potential incompleteness of the established regulatory networks in cases where target genes are controlled by only one known regulator. The discovery of highly consistent network subcomponents indicates that a gene expression data-based reconstruction of regulatory networks can be a powerful strategy for particular subcomponents that are sufficiently isolated and for which sufficient quantities of relevant data are available. The increasing availability of other high-throughput data types such as genome-wide location-analysis data (Lee et al. 2002) will further improve the prospects of such reconstruction, as additional data types can be used to resolve inconsistencies (Wyrick and Young 2002).

The full utilization of all high-throughput data types, however, will require the combination of prior biological knowledge

extracted from databases and literature with the statistical analysis of the large-scale data sets. Thus, full reconstruction of regulatory networks will rely on a combination of "bottom-up" (based on descriptions of individual interactions and regulons in the literature) and "top-down" (based on large systemic data sets) approaches with targeted prospective experimentation to successively resolve inconsistencies between the two. Ultimately, all such data types are expected to be reconciled in the context of genome-scale, in silico models of regulatory networks that can be used to analyze, interpret, and ultimately predict their function (Covert and Palsson 2002; Palsson 2002).

METHODS

Regulatory Networks

We utilized the recently published reconstructions of regulatory network structures for *E. coli* (Shen-Orr et al. 2002) and yeast (Guelzim et al. 2002). These reconstructions are primarily based on literature-derived information stored in databases, RegulonDB for *E. coli* (Salgado et al. 2001) and YPD for yeast (Costanzo et al. 2001), with some additional manual curation based on recent research literature. The procedures used for these reconstructions are similar to well-established reconstruction procedures for metabolic networks (Covert et al. 2001). All autoregulatory interactions were removed from the networks, as these cannot be studied using steady-state gene expression data. Each operon in the *E. coli* network was split into individual genes that comprise the operon, and the regulators of the operon were assigned as regulators of each individual gene. In both networks, all genes that did not have gene expression data in the final compendium data sets were removed resulting in somewhat smaller networks than those reported in Guelzim et al. (2002) and Shen-Orr et al. (2002).

Gene Expression Data

For yeast, the gene expression data were downloaded directly from the source described in each individual publication (see Supplementary Material for the references). For *E. coli*, the data were loaded from either the source described in the publication (see Supplementary Material), Stanford Microarray Database (Sherlock et al. 2001), or the ASAP database at the University of Wisconsin (Glasner et al. 2003). The data produced using cDNA microarrays were converted into $\log_2$ ratios if they were not in that form already. For data sets produced using oligonucleotide arrays, the data were first converted into $\log_2$ expression ratios between each experimental condition and a chosen reference condition. The individual preprocessed data sets were combined into a compendium data set for both yeast (904 experiments) and *E. coli* (141 experiments). Genes with over two thirds of missing values in the compendiums were removed from the final data set. The final compendiums were standardized by both experiment and gene and organized into a data matrix X of size N genes times M experiments.

Consistency Measures

For every instance of each of the four basic types of regulatory network elements studied in this work, we derived a weight for every experiment in the gene expression data compendium. For regulator module k, the weight (normalized over experiments) for experiment j was calculated as

$$w_{kj} = \frac{|x_{rj} \cdot \bar{x}_{tj}|}{\sum_{s=1}^{M} w_{ks}} \quad (1)$$

where x_{rj} is the expression level of the regulator gene of the module in experiment j and $\bar{x}_{tj}$ is the mean expression level of all target genes in the module. For MRMs, the weight was calculated as in equation 1, but x_{rj} was replaced by the mean expression level of all regulators of the module $\bar{x}_{rj}$. For RIs, the weight was calculated as in equation 1, but $\bar{x}_{tj}$ was replaced by the expression level of the single target gene x_{tj}. The same weight was used for feed-forward loops. Finally, for target modules, x_{rj} was replaced by the mean expression level of all the regulators of the target $\bar{x}_{rj}$and $\bar{x}_{tj}$ was replaced by the expression level of the single target gene x_{tj}. Below we will assume that the expression profiles have already been weighted with the appropriate weight vector for each element so that the expression level of gene i in experiment j is redefined as $x_{ij} \rightarrow W_{k(i)j}x_{ij}$, where $k(i)$ is the index to the element that gene i belongs to when the particular consistency measure is calculated (e.g., for RMs, $k(i)$ is the index to the module in which gene i is a target gene).

The consistency measures for RMs, MRMs, and RIs are all based on the standard Pearson product–moment correlation coefficient between two variables X and Y with M observations:

$$R(X,Y) = \frac{1/M \sum_{j=1}^{M} x_j y_j - \overline{xy}}{s_X s_Y}, \quad (2)$$

where $\bar{x}$ and s_X denote the mean and standard deviation of x respectively. The consistency measure we utilized for RMs is the Pearson correlation coefficient $R_{RM} = R(x_t, \bar{x}_{T\setminus t})$ between the expression profile each target gene in the module (x_t) with the mean of the expression profiles of the other targets in the module ($\bar{x}_{T\setminus t}$). The same measure was also used for MRMs except that the set of target genes in the same module with a particular target gene was different. For MRMs with only one regulator, both regulator and MRM measures are exactly the same. The consistency measure for RIs was simply the correlation coefficient between regulator and target expression profiles $R_{RI} = R(x_r,x_t)$.

For evaluating the consistency of the TMs, we utilized multiple coefficients of determination (Johnson and Wichern 2002) obtained by computing a multiple regression fit with the target expression level as a dependent variable and the regulator expression levels as independent variables:

$$x_{tj} = \beta_0 + \sum_{i \in R} \beta_i x_{ij} + \varepsilon_j. \quad (3)$$

Here, x_{tj} denotes the expression level of the target gene in experiment j, R is the set of regulator gene indices for this target module, β_i is the regression coefficient for the ith regulator, and ε_j is a residual error term. The multiple coefficient of determination for a target module is then defined as

$$R_{TM}^2 = \frac{\sum_{j=1}^{M} (\hat{x}_{tj} - \bar{x}_t)^2}{\sum_{j=1}^{M} (x_{tj} - \bar{x}_t)^2}, \quad (4)$$

where $\bar{x}_t$ is the average expression level of the target across all M experiments, and $\hat{x}_{tj}$ is the estimate for the expression level of the target gene in the jth experiment obtained using the regression model (equation 3). The multiple coefficient of determination for target modules with only one regulator is the same as the squared Pearson correlation coefficient R_{RI}^2 for the corresponding RI.

Partial correlation coefficients R_{FF} for feed-forward loops were computed by first calculating multiple regressions with both target gene expression and secondary regulator expression separately as the dependent variable and the set of primary regulator expression levels as the independent variables, and then calculating the Pearson correlation for the residuals from these regressions (Johnson and Wichern 2002). Except for R_{TM}^2 all the measures described above range from -1 to +1 with +1 indicating perfect correlation, -1 perfect anticorrelation, and 0 lack of correlation (R_{TM}^2 ranges between 0 and +1).

The major drawback with the consistency measures described above is that they underestimate the consistency if the dependencies between log-expression levels are nonlinear. However, we also investigated nonlinear measures of association such as mutual information computed based on discretized gene expression data and observed that the overall conclusions of this study were not dependent on the particular measure chosen.

Randomized Network Elements

To evaluate the significance of a particular value for each of the consistency measures described above in a network structure-dependent manner, we devise a separate nonparametric randomization strategy for each basic type of network element. In all the randomization tests described below, the weight vector used to evaluate the consistency measure for a particular network element was also used to evaluate the corresponding measure for the randomly generated network elements. The randomization process is illustrated in Figure 1B.

For pairwise correlation coefficients (RIs) and partial correlation coefficients (feed-forward loops), we created a null distribution by generating 1000 random gene pairs from the lists of all genes included in the compendium gene expression data set and evaluated R_{RI}* for each pair. A *P*-value for the *k*th interaction in the real regulatory network is then determined by computing the probability of observing a specific value $R_{RI,k}^2$ given the distribution of the randomized values $(R_{RI}^*)^2$ (Fig. 1B). Similarly for multiple coefficients of determination (TMs) we simulate a null distribution for TMs with different numbers of regulators by randomly choosing 1000 random target gene–regulator set pairs form the list of all genes so that the regulator set is of the same size as the true regulator set for a particular module. We then computed a multiple coefficient of determination R_{TM}^{2*} for this random TM. The probability of observing a specific value of $R_{TM,k}^2$ is evaluated given the distribution of R_{TM}^{2*} values for the same regulator set size. For RMs and MRMs, we generate a null distribution by choosing 1000 random sets of target genes of the same size as the individual modules and evaluating the probability of observing the particular value of $R_{RM,k}$ (or $R_{MRM,k}$) for the *k*th module given the null distribution for modules of this size. All the calculations and data processing were done using Matlab 6.1 (Mathworks) and Perl v.5.6.1.

ACKNOWLEDGMENTS

The authors acknowledge support by the National Institutes of Health (GM57089) and the ASLA-Fulbright Scholarship Program (Graduate Scholarship to M.J.H.). We thank Frederick Blattner for advance access to data in the ASAP database and Shankar Subramaniam, George Church, Julio Collado-Vides, and Leroy Hood for stimulating discussions.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Banerjee, N. and Zhang, M.Q. 2002. Functional genomics as applied to mapping transcription regulatory networks. *Curr. Opin. Microbiol.* **5:** 313–317.

Birnbaum, K., Benfey, P.N., and Shasha, D.E. 2001. *cis* element/transcription factor analysis (*cis*/TF): A method for discovering transcription factor/*cis* element relationships. *Genome Res.* **11:** 1567–1573.

Carlson, M. 1999. Glucose repression in yeast. *Curr. Opin. Microbiol.* **2:** 202–207.

Chrisman, L., Langley, P., Bay, S., and Pohorille, A. 2003. Incorporating biological knowledge into evaluation of causal regulatory hypotheses. *Pac. Symp. Biocomput.* **8:** 128–139.

Costanzo, M.C., Crawford, M.E., Hirschman, J.E., Kranz, J.E., Olsen, P., Robertson, L.S., Skrzypek, M.S., Braun, B.R., Hopkins, K.L., Kondu, P., et al. 2001. YPD, PombePD and WormPD: Model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res.* **29:** 75–79.

Covert, M.W. and Palsson, B.O. 2002. Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. *J. Biol. Chem.* **277:** 28058–28064.

Covert, M.W., Schilling, C.H., Famili, I., Edwards, J.S., Goryanin, I.I., Selkov, E., and Palsson, B.O. 2001. Metabolic modeling of microbial strains in silico. *Trends Biochem. Sci.* **26:** 179–186.

de Jong, H. 2002. Modeling and simulation of genetic regulatory systems: A literature review. *J. Comput. Biol.* **9:** 67–103.

DeRisi, J.L., Iyer, V.R., and Brown, P.O. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278:** 680–686.

D'Haeseleer, P., Liang, S., and Somogyi, R. 2000. Genetic network inference: From co-expression clustering to reverse engineering. *Bioinformatics* **16:** 707–726.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **95:** 14863–14868.

Featherstone, M. 2002. Coactivators in transcription initiation: Here are your orders. *Curr. Opin. Genet. Dev.* **12:** 149–155.

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., and Brown, P.O. 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11:** 4241–4257.

Glasner, J.D., Liss, P., Plunkett, I.G., Darling, A., Prasad, T., Rusch, M., Byrnes, A., Gilson, M., Biehl, B., Blattner, F.R., et al. 2003. ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Res.* **31:** 147–151.

Guelzim, N., Bottani, S., Bourgine, P., and Kepes, F. 2002. Topological and causal structure of the yeast transcriptional regulatory network. *Nat. Genet.* **31:** 60–63.

Hartemink, A.J., Gifford, D.K., Jaakkola, T.S., and Young, R.A. 2001. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomput.* **7:** 422–433.

Hartemink, A.J., Gifford, D.K., Jaakkola, T.S., and Young, R.A. 2002. Combining location and expression data for principled discovery of genetic regulatory network models. *Pac. Symp. Biocomput.* **7:** 437–449.

Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., et al. 2000. Functional discovery via a compendium of expression profiles. *Cell* **102:** 109–126.

Ideker, T.E., Thorsson, V., and Karp, R.M. 2000. Discovery of regulatory interactions through perturbation: Inference and experimental design. *Pac. Symp. Biocomput.* **292:** 305–316.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., and Barkai, N. 2002. Revealing modular organization in the yeast transcriptional network. *Nat. Genet.* **31:** 370–377.

Iyer, V.R., Horak, C.E., Scafe, C.S., Botstein, D., Snyder, M., and Brown, P.O. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409:** 533–538.

Johnson, R.A. and Wichern, D.W. 2002. *Applied multivariate statistical analysis*. Prentice Hall, Upper Saddle River, NJ.

Kalir, S., McClure, J., Pabbaraju, K., Southward, C., Ronen, M., Leibler, S., Surette, M.G., and Alon, U. 2001. Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. *Science* **292:** 2080–2083.

Kim, S., Dougherty, E.R., Chen, Y., Sivakumar, K., Meltzer, P., Trent, J.M., and Bittner, M. 2000. Multivariate measurement of gene expression relationships. *Genomics* **67:** 201–209.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., et al. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298:** 799–804.

Mewes, H.W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S., and Weil, B. 2002. MIPS: A database for genomes and protein sequences. *Nucleic Acids Res.* **30:** 31–34.

Palsson, B.O. 2002. In silico biology through "omics". *Nat. Biotechnol.* **20:** 649–650.

Pe'er, D., Regev, A., Elidan, G., and Friedman, N. 2001. Inferring subnetworks from perturbed expression profiles. *Bioinformatics* **17:** S215–S224.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., et al. 2000. Genome-wide location and function of DNA binding proteins. *Science* **290:** 2306–2309.

Ronen, M., Rosenberg, R., Shraiman, B.I., and Alon, U. 2002. Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl. Acad. Sci.* **99:** 10555–10560.

Sabatti, C., Rohlin, L., Oh, M.K., and Liao, J.C. 2002. Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res.* **30:** 2886–2893.

Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Millán-Zárate, D., Díaz-Peredo, E., Sánchez-Solano, F., Pérez-Rueda, E., Bonavides-Martínez, C., and Collado-Vides, J. 2001. RegulonDB (version 3.2): Transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.* **29:** 72–74.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. 2003. Module networks: Identifying regulatory

modules and their condition- specific regulators from gene expression data. *Nat. Genet.* **34:** 166–176.
Serres, M.H. and Riley, M. 2000. MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. *Microb. Comp. Genomics* **5:** 205–222.
Shen-Orr, S.S., Milo, R., Mangan, S., and Alon, U. 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* **31:** 64–68.
Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., Binkley, G., Matese, J.C., Dwight, S.S., Kaloper, M., Weng, S., Jin, H., Ball, C.A., et al. 2001. The Stanford Microarray Database. *Nucleic Acids Res.* **29:** 152–155.
Stormo, G.D. and Tan, K. 2002. Mining genome databases to identify and understand new gene regulatory systems. *Curr. Opin. Microbiol.* **5:** 149–153.
Struhl, K. 1999. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell* **98:** 1–4.
Tegner, J., Yeung, M.K., Hasty, J., and Collins, J.J. 2003. Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling. *Proc. Natl. Acad. Sci.* **100:** 5944–5949.
Wagner, R. 2000. *Transcription regulation in prokaryotes*. Oxford University Press, Oxford, UK.
Wang, W., Cherry, J.M., Botstein, D., and Li, H. 2002. A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci.* **99:** 16893–16898.
Wyrick, J.J. and Young, R.A. 2002. Deciphering gene expression regulatory networks. *Curr. Opin. Genet. Dev.* **12:** 130–136.
Yeung, M.K., Tegner, J., and Collins, J.J. 2002. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci.* **99:** 6163–6168.
Zhu, Z., Pilpel, Y., and Church, G.M. 2002. Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm. *J. Mol. Biol.* **318:** 71–81.

Received March 12, 2003; accepted in revised form June 18, 2003.