A Predictive Model for Transcriptional Control of Physiology in a Free Living Cell

Richard Bonneau,[2,5,7] Marc T. Facciotti,[1] David J. Reiss,[1] Amy K. Schmid,[1] Min Pan,[1] Amardeep Kaur,[1] Vesteinn Thorsson,[1] Paul Shannon,[1] Michael H. Johnson,[1] J. Christopher Bare,[1] William Longabaugh,[1] Madhavi Vuthoori,[1] Kenia Whitehead,[1] Aviv Madar,[2] Lena Suzuki,[4] Tetsuya Mori,[4] Dong-Eun Chang,[4] Jocelyne DiRuggiero,[3] Carl H. Johnson,[4] Leroy Hood,[1] and Nitin S. Baliga[1,6,7,*]

[1]Institute for Systems Biology, Seattle, WA 98103, USA
[2]Center for Genomics & Systems Biology, New York University, New York, NY 10003, USA
[3]University of Maryland, College Park, MD 20742, USA
[4]Vanderbilt University, Nashville, TN 37240, USA
[5]Courant Institute of Mathematical Sciences, Department of Computer Science, New York University, New York, NY 10003, USA
[6]Departments of Microbiology and Molecular and Cellular Biology, University of Washington, Seattle, WA 98195, USA
[7]These authors contributed equally to this work.
*Correspondence: nbaliga@systemsbiology.org

DOI 10.1016/j.cell.2007.10.053

SUMMARY

The environment significantly influences the dynamic expression and assembly of all components encoded in the genome of an organism into functional biological networks. We have constructed a model for this process in *Halobacterium salinarum NRC-1* through the data-driven discovery of regulatory and functional interrelationships among ~80% of its genes and key abiotic factors in its hypersaline environment. Using relative changes in 72 transcription factors and 9 environmental factors (EFs) this model accurately predicts dynamic transcriptional responses of all these genes in 147 newly collected experiments representing completely novel genetic backgrounds and environments—suggesting a remarkable degree of network completeness. Using this model we have constructed and tested hypotheses critical to this organism's interaction with its changing hypersaline environment. This study supports the claim that the high degree of connectivity within biological and EF networks will enable the construction of similar models for any organism from relatively modest numbers of experiments.

INTRODUCTION

Rapid DNA sequencing technology has provided access to a large number of complete genome sequences from diverse and often poorly characterized organisms. The hope is to use this information for engineering new biotechnological solutions to diverse problems spanning bioenergy, bioremediation, and medicine. In principle, it is a reasonable expectation to re-engineer new processes by selectively combining otherwise distinct biochemical capabilities encoded in different genomes. However, in reality this will only be possible when we have a sophisticated understanding of how the proteins encoded in each individual genome dynamically assemble into biological circuits through interactions with the environment. Given that in excess of 500 genomes have already been sequenced and little biological information exists for most of these organisms, a classical gene-by-gene approach is inefficient to accomplish this. Furthermore, since every organism is unique, it is impractical to rely on accumulated sets of known interactions from select model systems to construct really detailed models. A data-driven systems approach, on the other hand, is ideally suited to tackle this problem.

An important goal of applying systems approaches in biology is to understand how a simple genetic change or environmental perturbation influences the behavior of an organism at the molecular level and ultimately its phenotype. High-throughput technologies to interrogate the transcriptome, proteome, protein-protein, protein-DNA interactions, and so forth, present a powerful toolkit to accomplish this goal (DeRisi et al., 1997; Eichenberger et al., 2004; Laub et al., 2000; Liu et al., 2003; Masuda and Church, 2003). However, each of these individual data types captures an incomplete picture of global cellular dynamics. Therefore, these data need to be integrated appropriately to formulate a model that can quantitatively predict how the environment interacts with cellular networks to effect changes in behavior (Facciotti et al., 2004; Faith et al., 2007; Kirschner, 2005; Kitano, 2002). Accurate prediction of quantitative behavior is the ultimate test of our understanding of a given system that will enable re-engineering of cellular circuits. To this end, we have coordinated the integrated development and implementation of experimental and computational approaches to construct a predictive gene regulatory network model covering ~80% of the transcriptome of *Halobacterium salinarum NRC-1*, a free-living cell.

H. salinarum NRC-1 represents a class of poorly studied organisms (Archaea) and as such provides an explicit demonstration of how systems approaches can be used to rapidly characterize the already large and growing number of newly sequenced organisms. It also provides a unique window into molecular mechanisms underlying fascinating response physiologies in extreme environments such as above boiling temperatures and in deep sea ocean vents. Specifically, *H. salinarum NRC-1* thrives in an environment of ~4.5 M salinity and can be expected to provide insights into evolutionary adaptation for survival in high-salinity-induced low-water activity, which precludes growth of most organisms (Grant, 2004). Like most organisms it is also subject to daily and seasonal changes in many environmental factors (EFs), and one could expect it to have regulatory circuits that effectively negotiate these complex and often stressful conditions. From a practical standpoint, all these physiological capabilities are encoded in ~2400 nonredundant genes in a very compact and easily manipulable 2.6 Mbp genome (Ng et al., 2000). However, prior to this study only two regulons were characterized in this organism (Baliga et al., 2001; Hofacker et al., 2004). Consequently, we explored the value of a systems approach to rapidly discover and characterize a significant fraction of the gene regulatory network associated with the intercoordination of physiological processes in this organism in differing environmental and genetic backgrounds.

Since the power of a systems approach is in integrating as much information (old and new) into a unified model, in this study we have used data from whole-genome microarray analysis, genome-wide binding location analysis for eight transcription factors (TFs), mass spectrometry-based proteomic analysis, protein structure predictions, computational analysis of genome structure and protein evolution, and also data from public resources such as KEGG (Kanehisa, 2002) and STRING (von Mering et al., 2005). While some of these data are from prior studies (albeit our own recent work), a large fraction of the data, including 234 out of the 413 microarray experiments, were collected exclusively for this study to cover transcriptional responses to a spectrum of genetic and environmental perturbations. More importantly, all of the hypotheses constructed from the network model were verified with new data that were not used for its construction.

RESULTS AND DISCUSSION

EGRIN: A Dynamic Model of Transcriptional Control of Cellular Physiology in *H. salinarum NRC-1*

The basic premise of our approach was to perturb the cells (genetically or environmentally), characterize their growth and/or survival phenotype, quantitatively measure steady-state and dynamic changes in mRNAs, assimilate these changes into a network model that can recapitulate all observations, and, finally, experimentally validate hypotheses formulated from the model. This approach required the integrated development and implementation of computational and experimental technologies (Figure 1) and consisted of the following steps (see Experimental Procedures and Supplemental Data available online for details):

1 Sequence the genome and assign functions to genes using protein sequence and structural similarities (Bonneau et al., 2004; Ng et al., 2000).
2 Perturb cells by changing relative concentrations of EFs and/or gene knockouts (Table S1) (Baliga et al., 2004; Kaur et al., 2006; Kottemann et al., 2005).
3 Measure the resulting dynamic and/or steady-state transcriptional changes in all genes using microarrays (Table S2 and Figure S1) (Baliga et al., 2004; Kaur et al., 2006; Whitehead et al., 2006).
4 Integrate diverse data (mRNA levels, evolutionarily conserved associations among proteins, metabolic pathways, *cis*-regulatory motifs, etc.) with the cMonkey algorithm to reduce data complexity and identify subsets of genes that are coregulated in certain environments (biclusters) (Reiss et al., 2006).
5 Using the machine learning algorithm Inferelator construct a dynamic network model for influence of changes in EFs and TFs on the expression of coregulated genes (Bonneau et al., 2006).
6 Explore the network with Gaggle, a framework for data integration and software interoperability (Shannon et al., 2006), to formulate and then experimentally test hypotheses to drive additional iterations of steps 2–6.

Using this approach we collectively analyzed transcriptional responses to individual and combinatorial perturbations in 10 EFs including light, oxygen, UV radiation, gamma radiation, manganese (Mn), iron (Fe), cobalt (Co), nickel (Ni), copper (Cu), and zinc (Zn) and 32 genes including TFs, signal transducers, and metabolic enzymes (Tables S1 and S2). This classified 1929 of the total 2400 predicted genes into 300 biclusters that were often highly enriched in genes with known metabolic processes (Table S3). Each of these biclusters represents a subset of genes that are potentially coregulated in a defined set of environmental conditions. We then constructed subcircuits that model expression changes in each of these biclusters as a function of corresponding changes in 72 TFs (Table S4) and 9 EFs (although Co was included as a potential predictor it did not make it into the final network). The resulting model is a set of equations that can take as input measured changes in a few TFs and/or EFs to predict kinetic and steady-state transcriptional changes in ~80% of genes in this organism with an average (Pearson) correlation of ~0.8 to their actual measured changes. Importantly, this

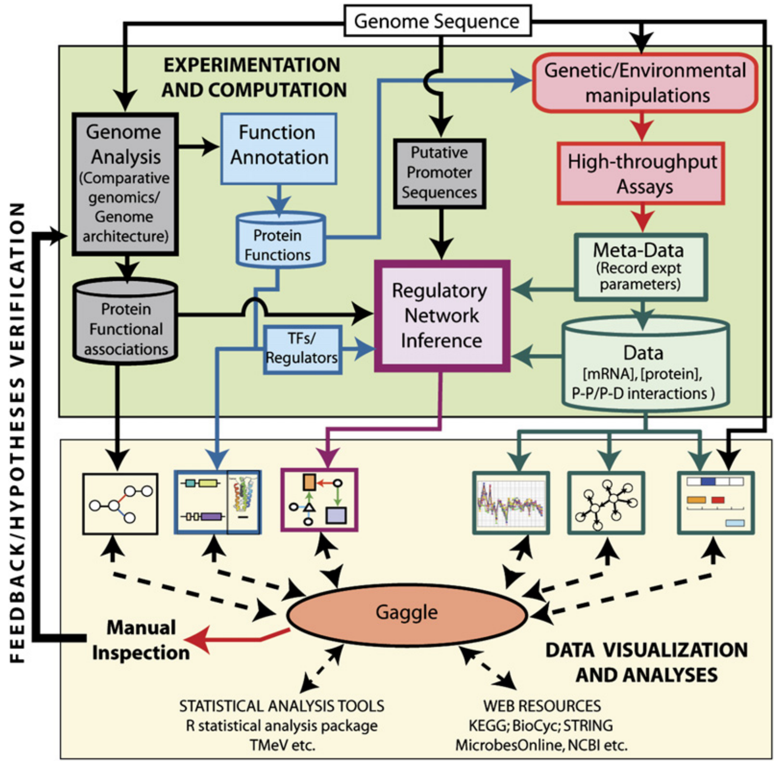
Genome Sequence
EXPERIMENTATION AND COMPUTATION
Genetic/Environmental manipulations
Genome Analysis (Comparative genomics/ Genome architecture)
Function Annotation
Putative Promoter Sequences
High-throughput Assays
Protein Functions
Meta-Data (Record expt parameters)
Regulatory Network Inference
Protein Functional associations
TFs/ Regulators
Data [mRNA], [protein], P-P/P-D interactions)
FEEDBACK/HYPOTHESES VERIFICATION
Gaggle
Manual Inspection
DATA VISUALIZATION AND ANALYSES
STATISTICAL ANALYSIS TOOLS
R statistical analysis package
TMeV etc.
WEB RESOURCES
KEGG; BioCyc; STRING
MicrobesOnline, NCBI etc.

Figure 1. Systems Approach for Predictive Modeling of Cellular Responses

Subsequent to genome sequencing there were two major interconnected and iterative components: experimentation and computation followed by data visualization and analyses. Within the first component the major efforts included computational genomic analyses for discovering functional associations among proteins (black boxes); putative functional assignment to proteins using sequence- and structure-based methods (blue boxes); and high-throughput microarray, proteomic, and ChIP-chip assays on genetically and/or environmentally perturbed strains (red boxes). All data (with the exception of proteomic and ChIP-chip data) from these approaches along with associated records of experiment design (green boxes) were analyzed with regulatory network inference algorithms (purple box). The resulting EGRIN was explored along with underlying raw data using software visualization tools within Gaggle (yellow box), which enables seamless software interoperability and database integration. Gaggle also provides a cost-effective interface to third party tools and databases. This manual exploration and analysis enabled hypothesis formulation and provided feedback for additional iterations of systems analyses.

predictive capability reduces significantly when the time component is removed from the model, strongly suggesting that a significant fraction of the influences have causal properties (Bonneau et al., 2006). Although we provide evidence that some of the regulatory influences are mediated directly via TF-DNA interactions, we expect that a large fraction, especially EF influences, act indirectly, for example, via interactions with signal-transducing environmental sensors. We, therefore, refer to this network as environment and *g*ene *r*egulatory *i*nfluence *n*etwork (EGRIN).

The dominant influence of the environment on the assembly of EGRIN was evident in two observations. First, we find that many of our strongest predicted interaction terms represent interactions between EFs and TFs—implying that the activity of the relevant TFs is dependent on presence of certain environmental conditions. Second, we observe that the transcription of 423 genes in 70 biclusters is predicted to be influenced by changes in one or more of the 9 EFs. With a few selected examples we discuss below both how EGRIN recapitulates and extends our understanding of biological processes we have previously studied—providing a mechanistic understanding of relevant interrelationships—and as well how it has generated insights into fascinating new biology of *H. salinarum NRC-1*.

Coregulated Modules within EGRIN Recapitulate and Extend Known Biology

A central aspect of this integrated effort is the data-driven grouping of genes into biologically meaningful biclusters. A good example of new biological insight discovered through integration of diverse data is provided through the analysis of energy production in *H. salinarum NRC-1*. Prior knowledge (gathered from literature and our own studies) has shown that three of the four known halobacterial energy production processes (arginine fermentation, phototrophy [using bacteriorhodopsin], and dimethyl sulfoxide [DMSO] respiration) operate in anoxic conditions, and the fourth (oxidative phosphorylation) requires oxic conditions (Baliga et al., 2002; Muller and DasSarma, 2005; Ruepp and Soppa, 1996). Specifically, we observed that all five genes of the phototrophy regulon (*bop*, *blp*, *brp*, *bat*, and *crtB1*) and the 6 genes responsible for DMSO respiration (*dmsR*/*E*/*A*/*B*/*C*/*D*) (Muller and DasSarma, 2005) cocluster within the 29 gene bicluster #208 (*bc*208) (Figure S2 and Table S5A), suggesting that these processes are coregulated under certain environmental conditions (Reiss et al., 2006). The composition of *bc*208 also suggested that the phototrophy regulon includes 9 additional genes. The evidence for this hypothesized expansion was that the promoters of these genes contain the putative binding site for the phototrophy regulator Bat (AtaCcCcAtgtgtTTGggTgTT-, $p < 10^{-10}$; Baliga and DasSarma, 1999; Table S5B); they are connected to the characterized phototrophy genes by one or more of three types of functional associations (operons, conserved chromosomal linkages across diverse organisms, and similar phylogeny); and they are coexpressed with the phototrophy genes under the conditions included in *bc*208 (Figure S2). Also, these 9 additional genes are not present in any other biclusters, suggesting that they may be exclusively associated with the phototrophy process.

This is an important contribution because it stresses the power of global systems approaches to reveal new aspects of biology by suggesting that a process we previously considered well understood still has a potentially large number of uncharacterized genes associated with it.

Although phototrophy and DMSO respiration were both known to be associated with anoxic metabolism, little was known about their operational relationships. The analysis of *bc*208 shows that while phototrophy and DMSO respiration genes appear to be coregulated under some environmental conditions, their expression diverges under others. While we are fairly certain that the regulation of phototrophy genes is mediated by Bat (Baliga et al., 2001), the absence of the putative Bat-binding site in DMSO metabolism-related genes in *bc*208 suggests that coregulation of these two processes under certain conditions may be either an indirect influence of Bat function or due to factors other than Bat that are common to both sets of genes. The motif search discovered at least two other promoter motifs that appear to be shared by all DMSO and phototrophy genes and represent putative binding sites for regulatory proteins common to both pathways. We predict that the combinatorial logic with which these various regulators operate is responsible for the conditional coregulation of the two processes.

EGRIN Predicts Novel Regulatory Control for Known Biological Processes

The Inferelator algorithm uncovers TF and EF influences on the expression of genes in biclusters. Transcriptional influences may be of three types: (1) direct influences, in which a TF acts through physical interactions with promoters of genes in the bicluster, as was the case in the previous example, wherein presence of a putative Bat-binding site in conjunction with genetic analyses (Baliga and DasSarma, 1999; Baliga et al., 2001) suggests that Bat directly influences the expression of phototrophy genes; (2) indirect influences, where a TF might act through a secondary TF; or (3) via an unknown mechanism that leads to the coexpression of the TF and genes in a bicluster. EF influences, on the other hand, mostly act via environment-sensitive TFs, or through signaling processes that direct changes in transcription and thus are always indirect (Figure S3A). Understanding the possible nature of these influences allows one to formulate testable hypotheses to characterize the regulatory mechanisms in the appropriate environmental context.

A good example of how TF and EF influences are integrated to describe the transcriptional changes in biclusters is provided by analyzing *bc*66 (Table S6 and Figures S3B and S3C). The transcriptional behavior of the 34 genes in *bc*66, including cytochrome oxidase, ribosomal proteins, and RNA polymerase, is nearly perfectly modeled by corresponding changes of four factors—two EFs (oxygen and light) and two TFs (Cspd1 and TFBf) (Figures 2A and 2B). We were able to further characterize the influence of oxygen on the expression of these genes by analyzing data from a controlled experiment in which only oxygen was perturbed (Figure 2C). Meanwhile independent ChIP-Chip experiments showed that TFBf interacted physically with a significant number of promoters in this biclusters (i.e., promoters of 24 out of 34 genes, $p < 10^{-10}$) (Figure 2D) (Facciotti et al., 2007). This fact strongly suggests that TFBf acts to influence the expression of these genes directly. In fact, of the 181 genes whose expression is modeled in EGRIN as a function of TFBf, promoters of at least 62 genes have binding sites for TFBf, implying a significant relationship between the statistically learned influence and actual promoter association ($p < 10^{-4}$). For a subset of genes under the direct influence of TFBf (24 genes in *bc*66), EGRIN has now provided an environmental context in which to further investigate the regulatory mechanisms. This predicted central role of *tfbF* in control of these critical functions was further substantiated by our inability to construct a viable knockout strain for this gene (Facciotti et al., 2007). In a similar manner, we were also able to assign specialized regulatory functions to two additional members of the seven gene TFB family (TFBb and TFBg) (Table 1; Facciotti et al., 2007; see Supplemental Data for details). While the regulatory influences of oxygen and TFBf on the expression of genes in *bc*66 seem relatively clear, the roles for light and Cspd1 are still to be determined. It is particularly interesting that the influence of TFBf acts through an AND logic gate with light, implying that the influence of TFBf on these genes is somehow dependent on the presence of light. This type of information provides valuable environmental context for further investigating the function of this general transcription factor (GTF).

EGRIN Connects Biological Processes in Previously Uncharacterized Combinatorial Relationships

The assembly of the regulatory influence subcircuits for all biclusters into the complete EGRIN has reconstructed known relationships among cellular processes that are connected in metabolic networks and play complementary roles (Figure S4). More importantly, based on the confidence gained from recapitulating these known relationships, we can investigate the architecture of EGRIN to discover new experimentally testable relationships. We illustrate this point by selecting genes distributed across 9 biclusters (*bc*20, *bc*28, *bc*45, *bc*48, *bc*61, *bc*75, *bc*76, *bc*163, and *bc*174) that bring together components of pyruvate metabolism, glutamate-glutamine metabolism, and ATP synthesis as well as some accessory functions required for enzyme cofactor biosynthesis and raw material transport to support these metabolic processes (Figure S5 and Table S7). The predicted subnetwork controlling these biclusters is presented in Figure 3A.

It is useful to first look directly at genes that cocluster into biclusters to understand what relevant information this first level of analysis can provide. We note that genes for cobalamin biosynthesis cocluster with components of pyruvate dehydrogenase (PDH: *pdhB* and *pdhA2*) and ribonucleotide reductase (NrdB2) into three distinct biclusters (*bc*45, *bc*61, and *bc*174). The coclustering of cobalamin biosynthesis genes with these two enzymes

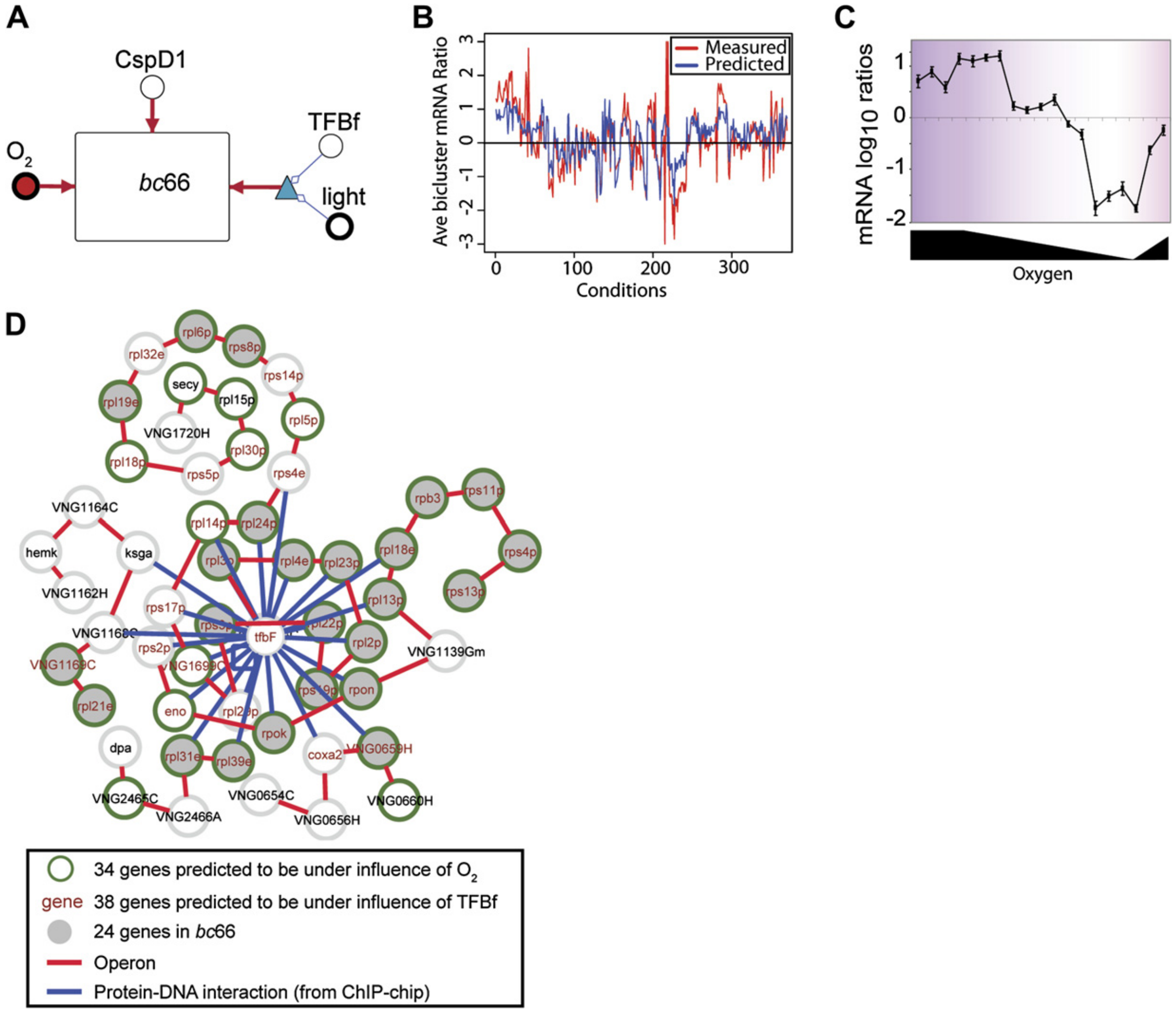

Figure 2. EGRIN Predicts Novel Regulatory Influences for Known Biological Processes

(A) Transcription of 34 genes encoding ribosomal proteins, RNA polymerase subunits, and cytochrome oxidase in *bc*66 are predicted to be positively influenced (red arrows) by two EFs (oxygen and light) and two TFs (CspD1 and TFBf). The influences from TFBf and light act through an AND logic gate (triangle).

(B) The mRNA profile of *bc*66 recreated by the combined TFs and environmental influences is nearly identical to the actual (averaged) mRNA levels over 398 experiments.

(C) The transcript levels of genes in *bc*66 changes proportionally with changes in oxygen tension in controlled experiments. The profile represents average transcription level changes of genes in bc66. The error bars indicate the standard deviation among mRNA level changes of genes in bc66.

(D) Crosscorrelation of predicted influences in EGRIN with physically mapped binding sites (Facciotti et al., 2007) suggests that the TFBf influence may be directly effected via binding of this GTF to promoters of ~70% of the genes (and operons) in *bc*66 ($p < 10^{-10}$).

is consistent with their requirement of vitamin B12 as a cofactor (Peel, 1962; Sintchak et al., 2002). However, *bc*174 also contains phosphate (PO_4) and Mn transport genes that share a conserved *cis*-regulatory motif (-AttTaGcttTAcAtA-; $p < 10^{-6}$) with a cobalamin biosynthesis operon. To our knowledge neither PO_4 nor Mn transport are directly related with cobalamin biosynthesis, PDH, or NrdB2. Since genes can be present in multiple biclusters (Figure S5) or share regulatory influences with genes in other biclusters (Figure 3A), a simultaneous analysis of all genes in the 9 biclusters helps to connect the dots and draw connections among these seemingly unrelated processes. A metabolic reconstruction (Figure 3B) from this integrated analysis illustrates that expression of components of PDH, electron transport flavoproteins, and ATP synthase genes cocluster in *bc*45, *bc*61, and *bc*75 and are coordinated by common influences from Snp, TBPe, KaiC, and VNG0320H (Figures 3A and 3B). Further, phosphate-transport genes cocluster with glutamine synthetase, peptide transport, and peptidase genes in *bc*76. These functional overlaps and shared regulatory influences among the biclusters just noted (*bc*45, *bc*61, *bc*75, *bc*76, and *bc*174) link energy production (ATP biosynthesis) with energy-requiring processes such as

Table 1. Novel Biological Insights Gleaned through Experimental Tests on EGRIN Predictions

Prediction	Verification
TFBg influences transcription of 149 genes	TFBg binds the promoters of 85 of these genes ($p < 10^{-15}$)
TFBg regulates the sodium extrusion pump NhaC3	TFBg binds the promoter of NhaC3 and a perturbation in TFBg function results in significant downregulation of this gene
TFBf influences transcription of 181 genes	TFBf binds the promoters of 62 of these genes ($p < 10^{-4}$)
TFBb influences transcription of 64 genes	TFBb binds the promoters of 29 of these genes ($p < 10^{-6}$)
Trh4 influences transcription of glutamine synthetase (GlnA)	Trh4 binds the promoters of several glutamate metabolism genes including GlnA
VNG1179C influences transcription of the primary Cu-efflux mechanism	The VNG1179C knockout strain is Cu sensitive due to lack of transcriptional activation of YvgX
A secondary Cu-efflux pump (ZntA) is transcriptionally activated when the primary mechanism is suppressed	Transcription of zntA was upregulated at steady state in the VNG1179C knockout strain
SirR influences transcription of Mn and PO_4 transport genes	A sirR knockout resulted in perturbed regulation of Mn and PO_4 transport genes and poor growth under Mn stress
Siderophore biosynthesis is upregulated under Mn stress	Transcript levels of siderophore biosynthesis genes are significantly increased relative to the wild-type levels when the SirR knockout strain is subjected to Mn stress
The unknown function protein VNG1459H is associated with the phototrophy process	Unique peptides from this protein were detected only upon enriching the membrane complexes responsible for phototrophy
VNG0019H is a transcriptional repressor of the B subunit of DNA gyrase	Transcription of DNA gyrase B was upregulated in the VNG0019H deletion strain

See Supplemental Data for details.

glutamine synthetase and nitrogen source import and degradation (Figure 3B). The accessory processes (cobalamin biosynthesis, Co, and PO_4 transport) provide cofactors for these core functions, and, therefore, their expression is also modeled within the same subcircuit. Finally, the regulatory influences (activation or repression) connecting biclusters also provide insight into the nature of operational relationships among physiological processes, such as the inverse regulatory relationship (via TBPe AND KaiC and CspD1 AND PhoU) of glutamine synthesis-associated processes (*bc*45, *bc*61, *bc*75, *bc*76, *bc*163, and *bc*174) to those associated with its breakdown (*bc*20, *bc*28, and *bc*48) (Figures 3 and S4).

Next, to experimentally verify some predicted regulatory influences in this model we first tested the regulatory influence of Trh4 on *glnA* by localizing all of the genome-wide binding sites for this TF using ChIP-chip analysis. Consistent with EGRIN we observed a direct physical association of Trh4 with the *glnA* promoter (Figure S6A). In fact, we discovered that Trh4 also binds upstream to several other genes of glutamate metabolism including succinate semialdehyde dehydrogenase (GabD); carbamoyl phosphate synthase (CarB); glutamine-hydrolyzing NAD^+ synthase (NadE); and carbamate kinase (ArcC), implicating Trh4 as a key regulator of nitrogen assimilation (Figures S6B and S6C and Table S8). We also experimentally validated the predicted coregulation of PO_4 and Mn transport by SirR. Specifically, deleting *sirR* results in perturbed Mn-dependent transcriptional control of Mn and PO_4 transport genes, which is manifested by poor survival under Mn stress (Kaur et al., 2006). Thus, the three approaches for constructing regulatory networks, i.e., via statistical learning, physical mapping, and genetic analysis, provide complementary information that mechanistically characterizes the regulation of cellular physiology.

Thus, the regulatory influences within EGRIN provide operational relationships among biclusters, i.e., disparate coregulated segments of physiology. These are not merely correlations among changes in these functions, rather they are quantitative and temporal relationships. In this particular example, this is demonstrated in the capacity of EGRIN to recapitulate transcriptional changes in all 9 biclusters despite differences in how they relate to each other in different environments (Figures 3C and 3D)—this would not be possible if the influences were simple correlations. Therefore, a reasonable conclusion from this observation is that the dynamic relationships among the different processes have been captured in the EGRIN model. This architecture of intercoordination of different biological processes is bound to be unique to every organism (Kirschner, 2005) and can only be learned through an integrated systems approach employing environmental and genetic perturbations as described herein.

EGRIN Accurately Predicts Transcriptional Responses of over 1900 Genes to Completely Novel TF and EF Perturbations

An important test of our global regulatory network model and by proxy our understanding of *H. salinarum NRC-1*'s response to the environment is represented in the

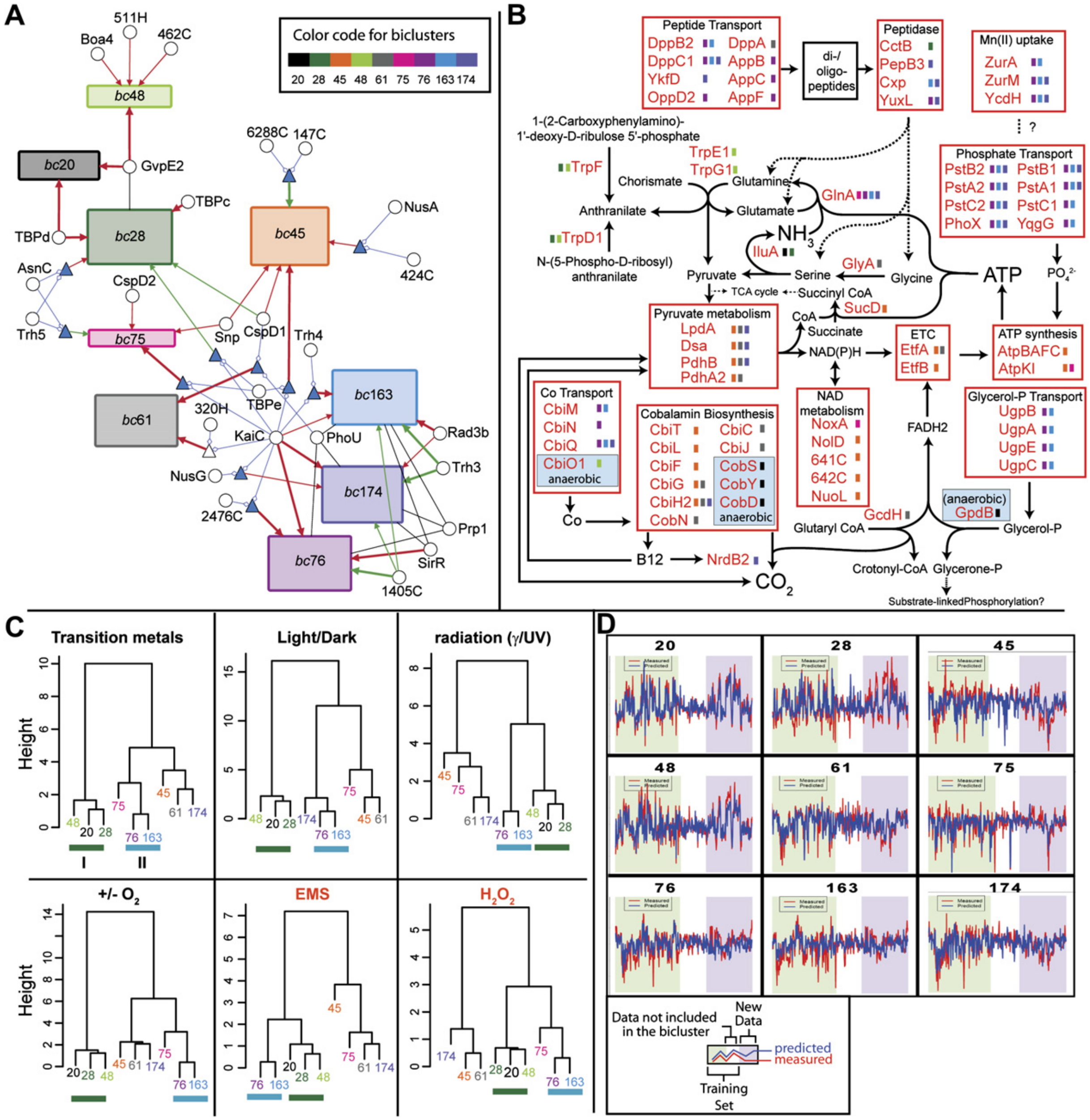

Figure 3. Regulatory Influences in EGRIN Model Intercoordination of Metabolic Processes in Diverse Environments

(A) Components of pyruvate metabolism, ATP synthesis, glutamate-glutamine metabolism, and accessory processes for transport of raw materials and synthesis of cofactors are distributed across 9 biclusters (boxes) containing altogether 162 genes. Functions associated with each bicluster, *cis*-regulatory motifs, and properties of biclusters are provided in the Supplemental Data (Table S7 and Figure S5). The expression of genes in these 9 biclusters is modeled by gene-regulatory influences (red: activate, green: repress, black: possible autoregulators coclustered with the regulated genes) from 27 TFs (circular nodes) that operate individually or in combination through AND gates (connected by blue edges).

(B) Metabolic pathways were reconstructed on the basis of known and putative functions of genes in the 9 biclusters. Memberships of various enzymes or enzyme subunits in each of the 9 biclusters in (A) are indicated with color-coded bars next to each step in the metabolic pathway (see key in panel A for interpreting this color code).

(C) The dendogram represents relationships among the 9 biclusters based on the similarities among the averaged expression profiles of their member genes. The differences in how the biclusters (cellular processes) relate to one another in varying environments are illustrated by highlighting relationships between two bicluster groups: I (*bc*20, *bc*28, *bc*48) and II (*bc*76 and *bc*163).

(D) The incorporation of weighted regulatory influences with an associated time constant into EGRIN enables the architecture of the network to change with the environment. As a consequence of this, despite environment-specified differences in relationships among cellular processes (C) the same set of regulatory influences acting on each bicluster accurately models the averaged transcriptional changes of its constituent genes

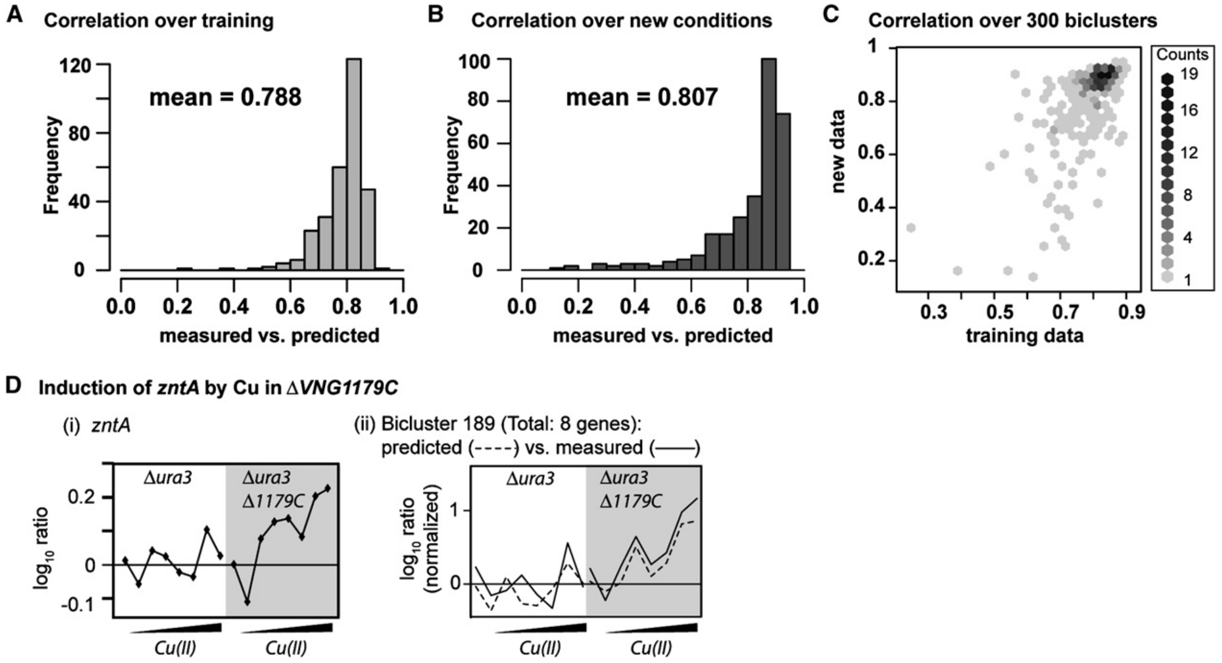

Figure 4. Prediction of Transcriptional Changes in New Environments

Histogram of Pearson correlations of predicted and measured mRNA levels of individual biclusters over the 266 experiments in the training set (A) and the 131 newly collected experiments (B) are shown. (C) shows a comparison of correlations between predicted and measured mRNA levels for all 300 biclusters in training set and new data. (D) Transcription of the broad specificity metal ion efflux pump ZntA is upregulated under Cu stress in the *ΔVNG1179C* strain background in which the primary efflux pump is transcriptionally inactivated (*Δura3* is the parent strain in which knockouts are constructed). This altered transcriptional response of ZntA to Cu was accurately modeled by the regulatory influences on *bc*189, which contains this gene along with 7 other genes.

accuracy with which this model predicts transcriptional responses to new EF and/or TF perturbations. To perform such a test we compared the predicted transcriptomes (relative transcript levels of 1929 genes) generated by the EGRIN to the measured transcriptomes in 147 new experiments that spanned (1) new combinatorial perturbations in EFs that individually were part of the original training set; (2) EF perturbations that were not part of the training set, such as with the oxidative stress agent hydrogen peroxide and the chemical mutagen ethyl methyl sulfonate; and (3) new combinations of TF and EF perturbations, including Mn stress response of a knockout strain of SirR, an Mn-responsive TF; Cu stress response of a knockout strain of VNG1179C, a Cu-responsive TF (Kaur et al., 2006); and responses of GTF-perturbed strains during cell growth in batch cultures (Table S2) (Facciotti et al., 2007). Since the first time points in each time course experiment are only used for predicting transcriptional changes in subsequent time points, the actual number of experiments for which we made predictions was 131 (147 experiments—16 first time point experiments). We observed excellent concordance (i.e., a mean Pearson correlation of ~0.8) between the predicted and measured mRNA levels over these 147 new studies (Figures 4A–4C) with an error equivalent to that found over the original training set (Figure S7). The network even predicted novel responses to new combinations of EF and TF perturbations, such as the steady-state transcriptional upregulation of ZntA, a broad-specificity metal ion efflux pump for Zn, Ni, Cu, and Co (Kaur et al., 2006), under increased Cu stress upon genetic disruption of the primary Cu-specific efflux system (Figure 4D; see Supplemental Data for additional examples and details).

We speculate that two nonexclusive properties of biological systems and the environment explain why the EGRIN model predicts gene expression changes in new experiments. First, even single perturbations in the training set actually represent multiple perturbations from the molecular and cellular perspective. This occurs because of the physicochemical relationships among EFs, which cause a change in one EF to alter others. For example, intense sunlight raises temperature to increase salinity via evaporation that in turn reduces dissolved oxygen content. The cells sense these complex changes in multiple EFs to elicit the appropriate response that deals both with the primary perturbation as well as the resulting

even for responses to new EF perturbations (for example, responses to EMS and H_2O_2). Each of the nine graphs shows profiles of predicted versus measured transcript level changes in each individual bicluster in environmental responses that were part of the training set as well as 147 completely new experiments.

secondary changes. A good example of this is the differential regulation of all three metal ion efflux pumps (ZntA, YvgX, and Cpx) in experiments wherein metal ion composition was not intentionally perturbed (data not shown). Therefore, from an informational perspective, each EF perturbation experiment actually provides the information regarding cellular responses to changes in multiple EFs. This information has been incorporated into the EGRIN and can partly explain our ability to predict cellular responses to new conditions. Second, since biological networks (metabolic and gene regulatory networks) are highly interconnected, cellular responses elicited from the primary perturbation propagate throughout the cellular networks via shared metabolites and common regulatory elements. This design may have evolved to deal with the anticipated secondary environmental perturbations noted above or simply as a consequence of the interconnectedness of biological networks. EGRIN is, thus, a model of the control of physiological responses to both primary EF or TF perturbation as well as secondary changes in other related factors.

Insights into the Unique Lifestyle of a Halophilic Archaeon

A biological network such as EGRIN is an essential resource for characterizing processes critical to the interaction of an organism with its changing environment through hypothesis formulation and testing. For instance, we have experimentally validated circuits that manage Cu and Mn stress; discovered a hierarchy of regulation among two alternate mechanisms for Cu efflux; delineated specialized functions of TF family members; and also assigned functions to proteins with no characterized primary sequence orthologs. All of these and many additional examples are summarized in Table 1, and the details are provided in Supplemental Data. While it is obvious that these are all new insights into the biology of *H. salinarum NRC-1*, it is important to note that several of these have a broader impact on furthering our general understanding of gene regulation in Archaea, the most poorly studied of the three domains of life. From a more general standpoint, they also collectively highlight how a systems approach can help design specific experimentally testable hypotheses within the broader context of the global architecture of transcription regulation. The ability to gather this level of information regarding a poorly characterized organism from a single study is significant and unprecedented.

Herein, we describe one example that highlights how EGRIN has helped discover functional promoter interactions of a TFB family of GTFs (Facciotti et al., 2007) in the context of an important physiological property of *H. salinarum NRC-1* that enables its growth in high salinity. Briefly, to withstand high salinity *H. salinarum NRC-1*, like most halophilic archaea, maintains a high potassium/sodium (~4 M K^+ and ~1 M Na^+) content in its cytoplasm, which is in inverse proportion to the high Na^+/K^+ content in its environment (~2.7 mM K^+ and ~4.3 M Na^+). This type of adaptation to hypersaline conditions is believed to be energetically favorable relative to alternate strategies of synthesis and/or accumulation of organic osmolytes such as glycine-betaine. Active Na^+ extrusion and K^+ uptake are, therefore, central to this process and mediated through coupling the transport of these ions to an electrochemical proton (H^+) gradient or at the expense of ATP (Oren, 1999). The *H. salinarum NRC-1* genome encodes at least five putative Na^+/H^+ antiporters (COG1757), perhaps to buffer loss of a function central to its survival. Despite this redundancy, we were able to formulate a specific hypothesis regarding transcription regulation of the most abundantly expressed paralogs (NhaC3) (Table S9).

If we consider the ChIP-chip data alone this gene appears to be potentially under the direct control of up to five different TFBs (TFBb, c, d, f, and g) (Figure 5A). However, according to EGRIN, among these five TFBs, a perturbation in TFBg should have strongest influence on the transcription of this gene (Figure 5B). We tested this hypothesis by investigating the consequence of perturbing each of the seven TFBs (see Experimental Procedures for details) on the transcription of *nhaC3* during growth. Indeed, only a perturbation in TFBg resulted in significant downregulation of this active Na^+ extrusion pump during all stages of growth (Figure 5C). Although we cannot rule out that the other TFBs (and possibly additional regulators) can as well mediate transcriptional control of this gene in other environmental settings, it is clear from this example that the physical map of protein-DNA interactions alone is insufficient to construct functional biological circuits. More importantly, in a specific set of environmental conditions identified by cMonkey we can now further characterize the regulation of this pump relative to other aspects of physiology, such as phototrophy, which is also directly influenced by TFBg and known to establish a H^+ gradient that drives the extrusion of Na^+ (Lanyi, 1980). As EGRIN is refined through additional rounds of experiment and analysis we expect that such circuits will be characterized to an extent that will eventually enable the engineering of halophilicity into other organisms, such as into crops for agriculture in arid climates.

Conclusions

Our choice of *H. salinarum NRC-1* has helped highlight the power of a systems approach for rapidly discovering new biology in largely uncharacterized organisms. By observing the consequences of systematically perturbing this organism with both genetic and environmental perturbations we were able to construct statistically significant and meaningful associations among most genes encoded in the genome of this organism. However, transcriptional control of ~20% of all genes is not represented within the biclusters in the EGRIN model. While this could be due to technical limitations in measuring transcript level changes of these genes, or absence of their differential regulation in response to perturbations used in our studies, an important point to consider is that our model does not yet account for a plethora of regulatory mechanisms such as epigenetic modifications, small RNAs,

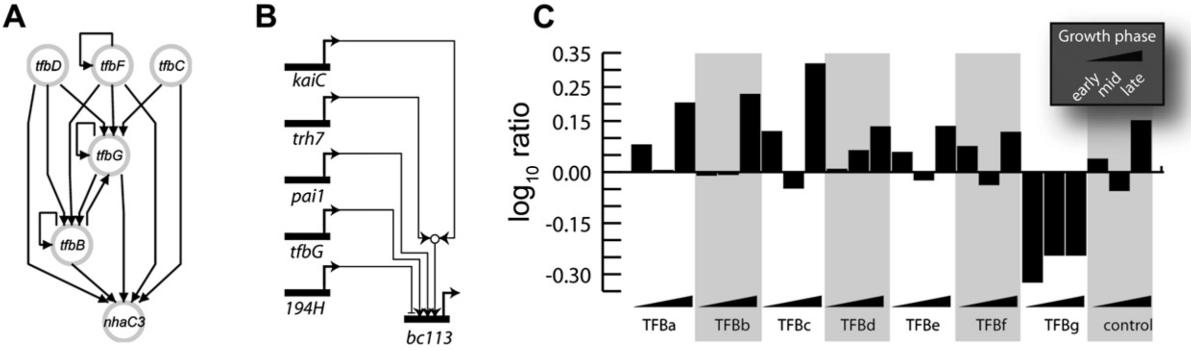

Figure 5. Perturbation in TFBg Function Results in Altered Regulation of the Na^+/H^+ Antiporter NhaC3

(A) The protein-DNA interaction map for five TFBs generated using ChIP-chip indicates the relative distribution of their binding sites upstream to their own promoters and the promoter for *nhaC3*.

(B) According to cMonkey *nhaC3* is coregulated with genes in five biclusters within EGRIN (*bc*2, *bc*3, *bc*12, *bc*16, *bc*50, and *bc*113). The average expression changes of genes in four of these biclusters are modeled by corresponding changes in TFBg transcript levels; the circuit diagram shows the Inferelator model for one of these biclusters (*bc*113).

(C) *nhaC3* transcript levels during different phases of growth in five strains, each carrying a plasmid-borne copy of the respective cmyc-tagged *tfb* gene (see Experimental Procedures for details). Functional binding of each of these engineered TFBs to cognate promoters was confirmed using ChIP-chip (Facciotti et al., 2007). Considering that all of these strains also have a chromosomal wild-type copy of the engineered TFB, the downregulation of *nhaC3* in the cmyc-TFBg strain suggests that this is a dominant-negative mutant.

posttranslational protein modifications, and metabolite-based feedback. The challenges associated with investigating these important control mechanisms at a global level are now being overcome through technological innovations. Our approach to regulatory network inference is extensible to incorporate these new data types and model their associated control mechanisms to eventually completely model the entire regulatory circuit in this archaeon. What is powerful about this approach is that it took under 6 years to move from genome sequence to this level of understanding for a relatively poorly studied organism. Indeed, it would be significantly quicker to implement the same approach with a newly sequenced organism given that much of the scientific methods including experimental procedures, algorithms, and software have been delineated through our study.

This does bring up an obvious question of whether the potential for enormous complexity of a biological system will ever allow the construction of a complete model of a cell. In this regard it has been favorably suggested, at least in the context of metabolism, that despite this potential for complexity, a cell usually functions in one of few dominant modes or states (Barrett et al., 2005). We speculate that this natural property of a biological system simplifies the problem to inferring gene regulatory models for its transitions among relatively few states. In addition, as discussed earlier, the extensive connectivity within EF and biological networks makes it tractable to effectively construct a comprehensive model of cellular responses to changes in multiple EFs from a modest number of well-designed systematic perturbation experiments (Faith et al., 2007; Hayete et al., 2007). We believe that this type of a model will hold true for environmental responses of all organisms and, more importantly, that it should be possible to construct such models solely from EF perturbation experiments. This will be especially valuable in context of organisms that currently lack tools for genetic analysis.

EXPERIMENTAL PROCEDURES

Genome Reannotation

A significant fraction (38%) of ~2400 genes in *H. salinarum NRC-1* could not be assigned any function using primary sequence-based approaches (Ng et al., 2000). We overcame this hurdle by incorporating functional relationships among proteins from comparative genomics (Bowers et al., 2004) as well as protein structure predictions to detect similarities at a three-dimensional level to proteins and protein domains in the Protein Data Bank (PDB) (Sussman et al., 1998). Using this approach nearly 90% of all predicted genes had some meaningful association with either a characterized protein, a protein family, or a structural fold (Bonneau et al., 2004). Further, through analysis of protein family signature or predicted structural matches we cataloged a list of 128 putative TFs (14 general transcription factors [6 TATA-binding proteins (TBPs), 7 transcription factor B (TFBs), and 1 transcription factor E alpha-subunit ortholog] and 114 putative sequence-specific DNA-binding proteins).

Genetic and Environmental Perturbations

We compiled a list of EFs (oxygen, sunlight, transition metals [Mn, Fe, Co, Ni, Cu, and Zn], UV radiation, and desiccation/rehydration [simulated with gamma radiation]) that are major forces in the natural habitat of *H. salinarum NRC-1*. Growth rate and survival characteristics in varying concentrations or exposures of these EFs were characterized to design the appropriate environmental perturbations (Baliga et al., 2004; Kaur et al., 2006; Kottemann et al., 2005). Genetic perturbations were also designed with either single gene in-frame deletions or non-native expression of 32 genes including sensors, signal transducers, response regulators, and enzymes functions that implicated them as potentially important regulators of responses to these EFs (Table S1).

Transcriptome Analyses

Two hundred and sixty-six microarray experiments were used for the construction of the network (see below), and 147 microarray experiments were used to validate predictions from the network. Roughly two-thirds of the 266 microarray-based experiments probed environmental responses to different doses of the same EF and/or temporal responses during acclimation to constant stress or recovery from a transitory perturbation. The remaining studies investigated perturbed responses in genetically perturbed strains (Table S2) (Baliga et al., 2004), gamma radiation (Whitehead et al., 2006), transition metals (Kaur et al., 2006), oxygen (Schmid et al., 2007), etc. All experiments that passed the statistical tests (Figure S1) (Ideker et al., 2000) were archived along with a digital log of growth conditions, genotypes, quantity and quality of perturbation, and time information. This meta-data information was used in the network inference procedure described in step 5. The description for 147 new experiments is discussed in the text and in Table S2.

Discovery of Coregulated Genes with cMonkey

The cMonkey algorithm iteratively scanned genes and/or conditions to identify groups of genes that are putatively coregulated in certain environmental conditions (biclusters) (Reiss et al., 2006). The probability of adding a gene to a bicluster was prioritized by two additional types of information: (1) its computationally predicted functional associations with genes in a given bicluster and (2) match(es) in its promoter to conserved *cis*-regulatory motif signatures detected by cMonkey in the putative gene promoters within a particular bicluster. Both of these constraints biased the composition of a bicluster to contain genes that have a greater likelihood of biochemically functioning together (e.g., genes in the same biochemical pathway and/or sharing a common motif in their promoter regions are more likely to be influenced by the same EF and/or TF). Finally, a critical attribute of this procedure is that it allows genes to belong to multiple biclusters to be consistent with known properties of biological systems in which genes can participate in multiple physiological functions depending on the condition or state of the cell.

Construction of EGRIN with Inferelator

Using the Inferelator algorithm (Bonneau et al., 2006), we discovered instances wherein individual or combinatorial changes in the concentrations of certain TFs (Table S4) and/or EFs (archived in the meta-information from step 3) temporally preceded average transcriptional changes within a given bicluster or a gene. Briefly, the Inferelator (1) selects parsimonious models (i.e., minimum number of regulatory influences for each bicluster) that are predictive (Thorsson et al., 2005); (2) explicitly includes the time dimension to discover causal influences; and (3) models combinatorial logic, i.e., interactions between EFs and TFs and between pairs of TFs. The collection of the complete set of regulatory influences connects all biclusters and genes into an integrated EGRIN.

Data Visualization, Exploration, and Analysis with Gaggle

We used the Gaggle (Shannon et al., 2006) software and database interoperability framework to interactively explore EGRIN in the context of (1) underlying experimental data in a local database (SBEAMS [http://www.sbeams.org]), (2) protein signatures (COG [Tatusov et al., 2000], PFam [Bateman et al., 2000]), (3) metabolic pathways (KEGG [Kanehisa, 2002]), (4) functional associations from an evolutionary standpoint, and (5) co-occurrence in scientific literature (STRING [von Mering et al., 2005]).

Data Accessibility

Microarray and ChIP-chip data in this manuscript have been submitted to NCBI GEO public repository. Proteomics data are available from the Peptide Atlas website (http://www.peptideatlas.org/). The data, algorithms, software, biclusters, and gene regulatory influence circuits are also accessible at http://baliga.systemsbiology.net/egrin.php.

Supplemental Data

Supplemental Data include Supplemental Results, Supplemental Experimental Procedures, eight figures, and nine tables and can be found with this article online at http://www.cell.com/cgi/content/full/131/7/1354/DC1/.

ACKNOWLEDGMENTS

This work was supported by grants from NIH (P50GM076547), DoE (MAGGIE: DE-FG02-07ER64327), and Institute for Systems Biology institutional support to N.S.B.; NSF (EF-0313754 and EIA-0220153), DoD (DAAD13-03-O-0057), and DoE (DE-FG02-04ER63807) to N.S.B. and L.H.; USAMRAA (US Army Medical Research Acquisition Activity) W81XWH-04-01-0307 to R.B.; NSF (MCB-0425825) and NASA (NNG05GN58G) to J.D.R. and N.S.B.; and NIH (R01GM067152) to C.H.J. Postdoctoral fellowships are acknowledged from NSF to M.T.F. (0400598) and K.W. and from NIH to A.K.S. We also thank Nathan Price and David Galas for critical reading of the manuscript and their helpful comments.

Received: August 10, 2007
Revised: September 27, 2007
Accepted: October 31, 2007
Published: December 27, 2007

REFERENCES

Baliga, N.S., Bjork, S.J., Bonneau, R., Pan, M., Iloanusi, C., Kottemann, M.C.H., Hood, L., and DiRuggiero, J. (2004). Systems level insights into the stress response to UV radiation in the halophilic archaeon halobacterium NRC-1. Genome Res. *14*, 1025–1035.

Baliga, N.S., and DasSarma, S. (1999). Saturation mutagenesis of the TATA box and upstream activator sequence in the haloarchaeal bop gene promoter. J. Bacteriol. *181*, 2513–2518.

Baliga, N.S., Kennedy, S.P., Ng, W.V., Hood, L., and DasSarma, S. (2001). Genomic and genetic dissection of an archaeal regulon. Proc. Natl. Acad. Sci. USA *98*, 2521–2525.

Baliga, N.S., Pan, M., Goo, Y.A., Yi, E.C., Goodlett, D.R., Dimitrov, K., Shannon, P., Aebersold, R., Ng, W.V., and Hood, L. (2002). Coordinate regulation of energy transduction modules in Halobacterium sp. analyzed by a global systems approach. Proc. Natl. Acad. Sci. USA *99*, 14913–14918.

Barrett, C.L., Herring, C.D., Reed, J.L., and Palsson, B.O. (2005). The global transcriptional regulatory network for metabolism in Escherichia coli exhibits few dominant functional states. Proc. Natl. Acad. Sci. USA *102*, 19103–19108.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., and Sonnhammer, E.L. (2000). The Pfam protein families database. Nucleic Acids Res. *28*, 263–266.

Bonneau, R., Baliga, N.S., Deutsch, E.W., Shannon, P., and Hood, L. (2004). Comprehensive de novo structure prediction in a systems-biology context for the archaea Halobacterium sp. NRC-1. Genome Biol. *5*, R52.

Bonneau, R., Reiss, D.J., Shannon, P., Facciotti, M., Hood, L., Baliga, N.S., and Thorsson, V. (2006). The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. Genome Biol. *7*, R36.

Bowers, P.M., Pellegrini, M., Thompson, M.J., Fierro, J., Yeates, T.O., and Eisenberg, D. (2004). Prolinks: a database of protein functional linkages derived from coevolution. Genome Biol. *5*, R35.

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.

Eichenberger, P., Fujita, M., Jensen, S.T., Conlon, E.M., Rudner, D.Z., Wang, S.T., Ferguson, C., Haga, K., Sato, T., Liu, J.S., et al. (2004). The program of gene transcription for a single differentiating cell type during sporulation in Bacillus subtilis. PLoS Biol. *2*, e328. 10.1371/journal.pbio.0020328.

Facciotti, M.T., Bonneau, R., Hood, L., and Baliga, N.S. (2004). Systems biology experimental design - considerations for building predictive gene regulatory network models for prokaryotic systems. Curr. Genomics *5*, 527–544.

Facciotti, M.T., Reiss, D.J., Pan, M., Kaur, A., Vuthoori, M., Bonneau, R., Shannon, P., Srivastava, A., Donohoe, S.M., Hood, L.E., et al. (2007). General transcription factor specified global gene regulation in archaea. Proc. Natl. Acad. Sci. USA *104*, 4630–4635.

Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J., and Gardner, T.S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. *5*, e8. 10.1371/journal.pbio.0050008.

Grant, W. (2004). Life at low water activity. Phil. Trans. Roy. Soc. B Biol. Sci. *359*, 1249–1267.

Hayete, B., Gardner, T.S., and Collins, J.J. (2007). Size matters: network inference tackles the genome scale. Mol. Syst. Biol. *3*, 77.

Hofacker, A., Schmitz, K.M., Cichonczyk, A., Sartorius-Neef, S., and Pfeifer, F. (2004). GvpE- and GvpD-mediated transcription regulation of the p-gvp genes encoding gas vesicles in Halobacterium salinarum. Microbiology *150*, 1829–1838.

Ideker, T., Thorsson, V., Siegel, A.F., and Hood, L.E. (2000). Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. J. Comput. Biol. *7*, 805–817.

Kanehisa, M. (2002). The KEGG database. Novartis Found. Symp. *247*, 91–101.

Kaur, A., Pan, M., Meislin, M., Facciotti, M.T., El-Geweley, R., and Baliga, N.S. (2006). A systems view of haloarchaeal strategies to withstand stress from transition metals. Genome Res. *16*, 841–854.

Kirschner, M.W. (2005). The meaning of systems biology. Cell *121*, 503–504.

Kitano, H. (2002). Systems biology: A brief overview. Science *295*, 1662–1664.

Kottemann, M., Kish, A., Iloanusi, C., Bjork, S., and Diruggiero, J. (2005). Physiological responses of the halophilic archaeon Halobacterium sp. strain NRC1 to desiccation and gamma irradiation. Extremophiles *9*, 219–227.

Lanyi, J.K. (1980). Light-driven primary sodium ion transport in Halobacterium halobium membranes. J. Supramol. Struct. *13*, 83–92.

Laub, M.T., McAdams, H.H., Feldblyum, T., Fraser, C.M., and Shapiro, L. (2000). Global analysis of the genetic network controlling a bacterial cell cycle. Science *290*, 2144–2148.

Liu, Y., Zhou, J., Omelchenko, M.V., Beliaev, A.S., Venkateswaran, A., Stair, J., Wu, L., Thompson, D.K., Xu, D., Rogozin, I.B., et al. (2003). Transcriptome dynamics of Deinococcus radiodurans recovering from ionizing radiation. Proc. Natl. Acad. Sci. USA *100*, 4191–4196.

Masuda, N., and Church, G.M. (2003). Regulatory network of acid resistance genes in Escherichia coli. Mol. Microbiol. *48*, 699–712.

Muller, J.A., and DasSarma, S. (2005). Genomic analysis of anaerobic respiration in the archaeon Halobacterium sp. strain NRC-1: dimethyl sulfoxide and trimethylamine N-oxide as terminal electron acceptors. J. Bacteriol. *187*, 1659–1667.

Ng, W.V., Kennedy, S.P., Mahairas, G.G., Berquist, B., Pan, M., Shukla, H.D., Lasky, S.R., Baliga, N.S., Thorsson, V., Sbrogna, J., et al. (2000). From the cover: genome sequence of halobacterium species NRC-1. Proc. Natl. Acad. Sci. USA *97*, 12176–12181.

Oren, A. (1999). Bioenergetic aspects of halophilism. Microbiol. Mol. Biol. Rev. *63*, 334–348.

Peel, J.L. (1962). Vitamin B12 derivatives and the CO2-pyruvate exchange reaction: a reappraisal. J. Biol. Chem. *237*, PC263–PC265.

Reiss, D.J., Baliga, N.S., and Bonneau, R. (2006). Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. BMC Bioinformatics *7*, 280.

Ruepp, A., and Soppa, J. (1996). Fermentative arginine degradation in Halobacterium salinarium (formerly Halobacterium halobium): genes, gene products, and transcripts of the arcRACB gene cluster. J. Bacteriol. *178*, 4942–4947.

Schmid, A.K., Reiss, D.J., Kaur, A., Pan, M., King, N., Van, P.T., Hohmann, L., Martin, D.B., and Baliga, N.S. (2007). The anatomy of microbial cell state transitions in response to oxygen. Genome Res. *17*, 1399–1413.

Shannon, P., Reiss, D.J., Bonneau, R., and Baliga, N.S. (2006). Gaggle: An open-source software system for integrating bioinformatics software and data sources. BMC Bioinformatics *7*, 176.

Sintchak, M.D., Arjara, G., Kellogg, B.A., Stubbe, J., and Drennan, C.L. (2002). The crystal structure of class II ribonucleotide reductase reveals how an allosterically regulated monomer mimics a dimer. Nat. Struct. Biol. *9*, 293–300.

Sussman, J.L., Lin, D., Jiang, J., Manning, N.O., Prilusky, J., Ritter, O., and Abola, E.E. (1998). Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. Acta Crystallogr. D Biol. Crystallogr. *54*, 1078–1084.

Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. *28*, 33–36.

Thorsson, V., Hornquist, M., Siegel, A.F., and Hood, L. (2005). Reverse engineering galactose regulation in yeast through model selection. Stat. Appl. Genet. Mol. Biol. *4*, 28.

von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A., and Bork, P. (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. *33*, D433–D437.

Whitehead, K., Kish, A., Pan, M., Kaur, A., Reiss, D.J., King, N., Hohmann, L., DiRuggiero, J., and Baliga, N.S. (2006). An integrated systems approach for understanding cellular responses to gamma radiation. Mol. Syst. Biol. *2*, 47.

Accession Numbers

The microarray data discussed in this publication have been deposited in NCBIs Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) and are accessible through GEO Series accession numbers GSE1040, GSE4890, GSE4891, GSE4892, GSE4893, GSE4894, GSE4895, GSE4896, GSE4897, GSE4898,GSE4899, GSE4900, GSE5557, GSE5929, GSE6776, GSE7609, GSE7610, GSE7611, GSE7612, GSE7613, GSE7709, GSE7710, GSE7711, GSE7712, GSE7713, GSE7740. See Table S2 for more details.