A New Formulation of Gene Networks

Gemunu H. Gunaratne\textsuperscript{1\star}, Preethi H. Gunaratne\textsuperscript{2,3}, Lars Seemann\textsuperscript{1}, and Andrei Török\textsuperscript{4}

\textsuperscript{1}Department of Physics, University of Houston, Houston, TX 77204

\textsuperscript{2}Department of Biology and Biochemistry, University of Houston, Houston, TX 77204

\textsuperscript{3}Human Genome Sequencing Center and Department of Pathology, Baylor College of Medicine, Houston, TX 77030

\textsuperscript{4}Department of Mathematics, University of Houston, Houston, TX 77204

\textsuperscript{\star} Corresponding author. Email: gemunu@uh.edu
Abstract

Most difficulties associated with implementing gene therapy, which could prevent the onset of hereditary diseases in genetically susceptible individuals, are due to the complexity of the underlying regulatory networks. The forms of nonlinear interactions between the hundreds of genes, proteins, and metabolites in these networks are not known with sufficient accuracy to compute changes necessary to move their steady states. An alternative approach to analyzing a biological process is to limit consideration to genes on the network. Steady state solutions of these influence networks can be obtained from DNA microarray experiments on mutants; in principle, effective interaction between their nodes can be inferred from many such solutions. However, since they contain a large number of nodes, the computation of influence networks requires a prohibitively large set of microarray experiments. Furthermore, since error rates associated with gene expression measurements are large, the inversions required to deduce a network increase the uncertainty to a level that makes verifiable predictions impossible. Here, we propose a reduction of influence networks that retains the ability to address several critical questions on the underlying system. The construction relies on the identification of a small group of genes that can be used to affect changes in other nodes of the network. The reduced network, which we refer to as the effective empirical subnetwork (EES), can be computed using results from a small number of microarray experiments on genetically perturbed systems. It can be used to make predictions on the expression profiles of other mutants, and to compute how to implement pre-specified changes in the steady state of the underlying biological process. These assertions are verified in a synthetic influence network. We also use previously published experimental data to compute the EES associated with an oxygen deprivation network of E. coli, and use it to predict gene expression levels on a double mutant. The predictions are significantly different from the experimental results for less than 30% of genes.
I. INTRODUCTION

Living systems are typically able to maintain their physiological state under environmental changes and isolated genetic mutations [1]. This robustness, referred to as homeostasis [2] or canalization [3, 4], is achieved through feedback within highly connected regulatory networks of genes, proteins and metabolites [5–10]. For example, an action that reduces the expression of one gene may cause coordinate changes in other nodes to leave the physiological state unaffected. If a genetic mutation blocks one pathway, other avenues on the associated network may take its place. Unfortunately, this systemic stability often makes it difficult to eliminate defects in a biological network, as evidenced by the surprising lack of efficacy of many drugs that were designed to act on single molecular targets [11, 12]. The coupling can also lead to side effects from medications. For example, anti-inflammatory COX-2 inhibitors (e.g., Vioxx) cause adverse cardiovascular effects due to a concomitant imbalance of the lipids prostacyclin and thromboxane A2, which lie on the same network [13]. Clearly, the most effective and least detrimental changes in a biological process are implemented by altering the system in its entirety. This task requires predictive mathematical models which can be constructed from experimental data. In this paper, we propose an approach for such a construction.

There are hundreds of genes, proteins, and other molecular participants associated with most biological processes. Gene regulatory networks model all interactions between these nodes. However, the forms of these dependencies, as well as kinetic parameters such as reaction rates and diffusion constants are, at best, only known approximately [14]. It is unlikely that gene regulatory networks which are sufficiently accurate to make quantitative predictions on the underlying biological processes will be available in the near future [15, 16].

Many approaches to reduce the complexity of regulatory networks have been proposed [5]. Small modules or network motifs [17, 18] associated with specific tasks have been identified. Boolean variables [19] can reduce the complexity, although the coarse-graining will limit predictability to qualitative characteristics such as bifurcations. Limiting consideration to genes provides a reduction that can still address quantitative questions on the biological process [14, 20]. Regulatory interactions between nodes of these influence networks include direct physical actions as well as indirect actions mediated by proteins or other molecules.

Consider an influence network containing $N$ genes $G = \{G_1, G_2, \ldots, G_N\}$; denote the
expression level of the $K^{th}$ gene by $X_K$ and the state of the network by $X \equiv \{X_1, X_2, \ldots X_N\}$.

The influence network can be modeled by a set of ordinary differential equations $F : \mathbb{R}^N \rightarrow \mathbb{R}^N$

$$
\dot{X}_1 = F_1(X) \\
\dot{X}_2 = F_2(X) \\
\ldots \\
\dot{X}_N = F_N(X).
$$

(1)

Steady states of influence networks can be obtained from DNA microarray experiments [5]. However, most influence networks contain hundreds of genes; thus, even if $F_K(X)$ are assumed to have a simple form [21], a prohibitively large number of microarray experiments will need to be conducted in order to compute $F$. Moreover, gene expression levels in microarray experiments have large ($\sim 10\%$) error bars; when $N$ is large, the inversions needed to compute $F$ will exacerbate the uncertainty to a level that makes its predictions inapplicable. One possibility is to only extract partial information on these networks through inference algorithms such as Network Identification by Multiple Regression [14], and Microarray Network Identification [22].

We propose an alternative reduction which relies on the fact that genes in an influence network can be partitioned into strongly coupled subgroups or clusters. This partition can be made either using co-expression under genetic perturbations [23–25], or through the use of the Gene Ontology (GO) database (http://geneontology.org). Our main assumption is that the behavior of all nodes within a cluster can be controlled by imposing suitable changes in a small, specially chosen, subset of its members. The set could include genes that translate to transcription factors, and would hence influence many other genes [26, 27]. It could also include micro-RNAs within the cluster, each of which affect many genes through post-transcriptional regulation [28, 29].

Suppose we have partitioned the $N$ genes of the influence network into clusters, and identified a small number of genes/micro RNAs from each cluster that can be used to control the expression levels of the remaining genes. Denote the set of these nodes by $S$. The number $n$ of nodes in $S$ is much smaller than $N$. We will represent their expression levels by $x = \{x_1, x_2, \ldots, x_n\}$, and re-index the variables so that the remaining expression levels are $\{X_{n+1}, \ldots, X_N\}$. With the new ordering, we write the state of the network as
The set \( X = \{ x_{int}, X_{ext} \} \) where we will refer to \( x_{int} = \{ x_1, x_2, \ldots, x_n \} \) and \( X_{ext} = \{ X_{n+1}, \ldots, X_N \} \), as “internal” and “external” variables respectively.

We limit consideration only to steady states of the system as internal variables are modified. When external perturbations are made on genes within \( S \), expression levels of the remaining genes at equilibrium are determined by Eqn. (1). These steady states lie on an \( n \)-dimensional surface in \( \mathbb{R}^N \), which we denote by \( S_S \). Figure 1(A) shows a schematic (2-dimensional) solution surface for the synthetic network introduced in the Results Section.

We make the following observations on solutions of the system \( \dot{X} = F(X) \). First, we assume that the unconstrained system has a unique stable equilibrium which we denote by \( P(0) = \{ p_{int}(0), P_{ext}(0) \} = \{ p_{1}(0), \ldots, p_{n}(0), p_{n+1}(0), \ldots, p_{N}(0) \} \). It satisfies the \( N \) equations \( F(P(0)) = 0 \). The point \( P_0 \) representing it lies on \( S_S \). Next, consider the single knockout mutant \( \Delta G_m \) (assumed to be viable) obtained by knocking out the \( m \)th gene. Since \( x_m \) is set externally, the \( m \)th equation of (1) is no longer valid. The solution for the expression levels is obtained by solving the remaining \((N-1)\) equations. We denote this equilibrium by \( P^{(m)} = \{ p_{int}^{(m)}, P_{ext}^{(m)} \} \) with \( p_{m}^{(m)} = 0 \), and represent it by \( P_m \). Since the equilibrium is associated with changes made within the set \( S \), \( P_m \) lies on \( S_S \). Consequently, \( P_0 \) as well as \( P_m \) for \( m = 1, 2, \ldots, n \) lie on \( S_S \), see Figure 1(B).

Our goal is to construct a system, referred to as the “effective empirical subnetwork” (EES), that can be computed using the gene expression levels of mutants discussed above, and whose equilibria are close to \( S_S \). Observe that \( P_0 \) and the \( n \) points \( P_m \) define a unique \( n \)-dimensional plane in \( \mathbb{R}^N \), which we denote by \( H_S \). Figure 1(B) shows the surface for the example above. The EES is the (linear) system whose solutions lie on \( H_S \) as changes are made within \( S \). Due to the constraints imposed (i.e., that they pass through \( P_0 \), and \( P_m \), \( m = 1, 2, \ldots, n \)), the two surfaces \( S_S \) and \( H_S \) are close to each other in the region shown.

Observe that \( EES : \mathbb{R}^n \rightarrow \mathbb{R}^N \) is a linear function determined by \( P_0 \), and \( P_m \), \( m = 1, 2, \ldots, n \), but is otherwise independent of \( F \). In particular, each \( X_K \) is a linear function of \( x_1, \ldots, x_n \). Since \( P_0 \) lies on \( H_S \)

\[
X_K - P_{K}^{(0)} = \sum_{i=1}^{n} a_{Ki} \left( x_i - p_{i}^{(0)} \right),
\]

for each \( K = (n+1), \ldots, N \). The coefficients \( a_{Ki} \), \( 1 \leq i \leq n \), can be determined from the values of \( X_K \) in \( P_1, P_2, \ldots, P_n \). There is one additional consideration which we illustrate using the following example. Suppose we consider a mutant where only \( x_1 \) is controlled.
through an external action. The remaining expression levels of the steady state of this mutant are obtained by solving the last \((N - 1)\) components of Eqn. (1). In particular, the expression levels \(x_2, x_3, \ldots, x_n\) of the internal variables in this mutant depend on \(x_1\). In general, the internal variables themselves depend on the gene expression levels whose values are externally imposed. As we show in the Methods Section, these dependencies can be assumed to take the form

\[
x_k - p_{k(0)} = \sum_{i \neq k} a_{ki} (x_i - p_{i(0)}),
\]

for \(k = 1, 2, \ldots, n\).

The reductions made in constructing the EES restrict its applications to steady states of the network when changes are made within \(S\). It can be used to predict gene expression levels of mutants obtained by modifying multiple nodes in \(S\); e.g., double knockout mutants. The validity of the EES construction can be tested by comparing its predictions with microarray data from such mutants. More significantly, the EES can be used to compute how the steady state of the system can be moved from its initial state \(P_0\) to a pre-specified set of expression levels defined by a point \(P_{aim}\), see Figure 2. Since \(P_{aim}\) will, in general, not lie on the solution surface, it cannot be reached through changes within \(S\). Instead, we can use the EES to compute \(P_{lin}\), which is the closest point to \(P_{aim}\) on the plane \(H_S\), see Figure 2. Since the surfaces \(S_S\) and \(H_S\) are close, the changes imposed on the system by the external actions are expected to be close to those computed from the EES. Below, we verify this proximity in a synthetic influence network.

II. RESULTS

A. A Synthetic Influence Network

The model system is constructed using \(F_K(X)\) to be a linear combination of sigmoidal Hill functions; specifically,

\[
F_K(X) = X_K \sum_{I=1}^{N} a_{KI} \left[ H(X_I; c_{KI}) - H \left( P_{I(0)}; c_{KI} \right) \right],
\]

where \(H(X; c) = X^h / (X^h + c^h)\) is the Hill function and the Hill index \(h\) is chosen to be 2. The action of the \(I^{th}\) gene on the \(K^{th}\) one is characterized by parameters \(a_{KI}\) and \(c_{KI}\), which are assumed to be independent of the state \(X\) of the system. The action is activating
if $a_{KI} > 0$ and inhibiting if $a_{KI} < 0$. Note that the system is constructed so that $P^{(0)}$ is an equilibrium of Eqns. (1). Numerically, it appears that model systems defined by Eqns. (1) and (4), have at most one stable steady state.

In order to compute the solutions to the knockout mutant $\Delta G_k$, we set $x_k = 0$, and solve the remaining equations of (1) as a nonlinear least squares problem. When the normalized residue fails to fall below $10^{-10}$, it is assumed that the corresponding steady state does not exist.

We report on a model system containing 21 nodes and shown schematically in Figure 3. We start with the three subnetworks, each of size 7. The vector $P^{(0)}$ for each of these subgroups consists of random entries between 0.5 and 1.5, and the matrix $(c_{KI})$ contains random values between 0 and 2. The entries of the Jacobian of the system given by Eqns. (1) and (4) at $P^{(0)}$ are $J_{KI} = a_{KI} H'(P^{(0)}; c_{KI})$; thus $a_{KI}$ can be computed for a given set of $J_{KI}$’s. Since we require $P^{(0)}$ to be stable, we insist that all eigenvalues of the Jacobian be negative. This is guaranteed by starting with a diagonal matrix with negative entries and performing a (random) orthonormal transformation. Once three such subnetworks are computed, their nodes are coupled by sparse, weak interactions. Each node in a subnetwork is coupled to only one from each of the other subnetworks, and the mean coupling strength is chosen to be 0.1 of the average coupling within subgroups.

The EES is to be constructed using the expression levels of single knockout mutants. As illustrated in Figure 3, mutants $\Delta G_1$, $\Delta G_3$, $\Delta G_8$, $\Delta G_{10}$, $\Delta G_{15}$, $\Delta G_{18}$, $\Delta G_{20}$, and $\Delta G_{21}$ in our example are not viable; i.e., when the corresponding $X$ is set to zero, the system (1) does not have a solution. The subset on which to construct the EES can contain any of the other nodes. In the work reported here, $S = \{G_2, G_4, G_9, G_{11}, G_{16}, G_{19}\}$ (genes marked in red in Figure 3). The variables $X_K$, $K = 7, 8, \ldots, 21$ are re-indexed as described before. The EES : $\mathbb{R}^6 \to \mathbb{R}^{21}$ is computed using the expression levels of all 21 genes at $P_0, P_1, P_2, P_3, P_4, P_5$ and $P_6$.

In order to illustrate the proximity of $H_S$ to $S_S$, we use the following example, see Figure 1. Since we need to reduce the dimensionality for visualization, we fix the expression levels of (the re-indexed genes) $G_3$, $G_4$, $G_5$, and $G_6$ at their values at $P_0$; for our model, \{ $x_3, x_4, x_5, x_6$ \} = \{ 1.1716, 0.6279, 0.5140, 0.5128 \}. For each pair of values for $(x_1, x_2)$, we solve the model system (1) for the remaining 15 expression levels. These solutions lie on 2-dimensional surface in $\mathbb{R}^{17}$. The gray surface of Figure 1 is $X_7(x_1, x_2)$. The 2-dimensional
plane $\mathcal{H}_S$ of the EES contains points $\mathcal{P}_0$, $\mathcal{P}_1$, and $\mathcal{P}_2$.

Next, we compare expression levels of double knockout mutants predicted by the EES with the corresponding solutions of the model system (1). The 14 viable double knockout mutants of the system are $\Delta G_1 \Delta G_2$, $\Delta G_1 \Delta G_3$, $\Delta G_1 \Delta G_4$, $\Delta G_1 \Delta G_5$, $\Delta G_1 \Delta G_6$, $\Delta G_2 \Delta G_3$, $\Delta G_2 \Delta G_4$, $\Delta G_2 \Delta G_5$, $\Delta G_2 \Delta G_6$, $\Delta G_3 \Delta G_5$, $\Delta G_3 \Delta G_6$, $\Delta G_4 \Delta G_5$, $\Delta G_4 \Delta G_6$, and $\Delta G_5 \Delta G_6$. In each case, the expression levels of the 4 remaining nodes in $\mathcal{S}$, and the 15 nodes outside of $\mathcal{S}$ are compared. We differentiate between these two groups.

Results for the first group (genes in $\mathcal{S}$) are as follows. Of the 56 comparisons, 46 EES predictions are within 1% of the expression levels computed from (1), 3 others are between $1 - 5\%$, and 3 between $5 - 10\%$. Results for the second group (genes outside of $\mathcal{S}$) are as follows. Of the 210 expression levels to be compared, 170 EES predictions are within 1% of the expression levels computed from (1), 30 more are between $1 - 5\%$, and 7 others are between $5 - 10\%$.

We finally demonstrate how the equilibrium of the system can be moved (near) to a pre-specified set of expression levels. The original equilibrium of our example is $\mathbf{P}^{(0)} = \left\{ \mathbf{P}_{\text{int}}^{(0)}, \mathbf{P}_{\text{ext}}^{(0)} \right\}$, with

$\mathbf{P}_{\text{int}}^{(0)} = \{0.89, 0.97, 1.17, 0.63, 0.51, 0.51\}$,

$\mathbf{P}_{\text{ext}}^{(0)} = \{1.47, 0.74, 0.83, 1.24, 0.58, 1.03, 0.85, 1.17, 0.96, 1.39, 1.40, 0.53, 1.21, 0.68, 1.15\}.$

We wish to find out how the expression levels of genes in $\mathcal{S}$ need to be changed so that the system moves as close as possible to a pre-specified set of expression levels for all genes. As an example, we attempt to change the equilibrium of the system to $\mathcal{P}_{\text{aim}}$ (see Figure 2) given by $\mathbf{P}^{(\text{aim})} = \left\{ \mathbf{P}_{\text{int}}^{(\text{aim})}, \mathbf{P}_{\text{ext}}^{(\text{aim})} \right\}$, where

$\mathbf{P}_{\text{int}}^{(\text{aim})} = \{0.8, 0.6, 1.3, 0.7, 0.6, 0.6\}$

$\mathbf{P}_{\text{ext}}^{(\text{aim})} = \{1.6, 0.8, 0.9, 1.3, 0.5, 1.1, 0.9, 1.2, 0.9, 1.2, 1.5, 0.4, 1.3, 0.6, 1.1\}.$

Since we have computed the EES, we can calculate the projection $\mathcal{P}_{\text{lin}}$ of $\mathcal{P}_{\text{aim}}$ on $\mathcal{H}_S$. It is given by $\mathbf{P}^{(\text{lin})} = \left\{ \mathbf{P}_{\text{int}}^{(\text{lin})}, \mathbf{P}_{\text{ext}}^{(\text{lin})} \right\}$, where

$\mathbf{P}_{\text{int}}^{(\text{lin})} = \{0.87, 0.75, 1.30, 0.75, 0.57, 0.45\}$

$\mathbf{P}_{\text{ext}}^{(\text{lin})} = \{1.49, 0.73, 0.84, 1.08, 0.56, 1.03, 0.89, 1.18, 0.87, 1.22, 1.48, 0.46, 1.22, 0.67, 1.15\}.$

Finally, we use the model system Eqns. (1) and (4) to compute the external variables when internal variables are fixed at $\mathbf{P}_{\text{int}}^{(\text{lin})}$. It is found to be $\mathbf{P}^{(\text{sys})} = \left\{ \mathbf{P}_{\text{int}}^{(\text{sys})}, \mathbf{P}_{\text{ext}}^{(\text{sys})} \right\}$, where $\mathbf{P}_{\text{int}}^{(\text{sys})} =
The Euclidean distances between the points are $d(P_0, P_{aim}) = 0.55$, $d(P_0, P_{lin}) = 0.40$, $d(P_{aim}, P_{lin}) = 0.38$, and $d(P_{lin}, P_{sys}) = 0.15$. Thus, we attempted to move the equilibrium from $P_0$ to a point $P_{aim}$ that was a distance 0.55 away, but were only able to move it on $\mathcal{H}_S$ to a point $P_{lin}$, which is a distance 0.40 away from $P_{aim}$. However, $P_{lin}$ is only a distance 0.15 from the point $P_{sys}$, which is the solution of the original system when expression levels of the internal variables are set to $p_{int}^{(lin)}$. We have found that $P_{lin}$ and $P_{sys}$ are close in studies of several model systems and for many points $P_{aim}$. Thus, the EES can be used to pre-determine, approximately, the steady state of the original network when changes made within $S$.

### B. Transcriptional Regulatory Network in *E.coli*

The EES can be constructed using microarray data from the wildtype and single knockout mutants of genes in $S$. It can then be used to predict gene expression levels of other mutants. This last observation is of interest due to the availability of previously published data on an oxygen deprivation network in *E.coli* [30, 31]. Ref. [32] reports gene expression levels in the wildtype and in single knockout mutants of key transcriptional regulators in the oxygen response, namely $\Delta arcA$, $\Delta appY$, $\Delta fnr$, $\Delta oxyR$, and $\Delta soxS$, as well as in the double knockout mutant $\Delta arcA \Delta fnr$, in aerobic and anaerobic glucose minimal medium conditions. Since the oxygen deprivation network is not fully active under aerobic conditions, we focus on the behavior of *E.coli* under anaerobic conditions.

We construct $G$ as follows. In the Gene Ontology classification assigned by Affymetrix, the five genes $arcA$, $appY$, $fnr$, $oxyR$, and $soxS$ have a common term “GO:0006355, Regulation of transcription, DNA-dependent.” Moreover, this is the only common classification for the five genes. We choose $G$ to be the set of all genes carrying this term. The full list of 299 genes is given in Supplementary Materials.

The data set GSE1121 of the GEO site (www.ncbi.nlm.nih.gov) [32] provides gene expression levels for four replicates of the wildtype and three each for the mutants. The replicates are used to estimate the mean and standard deviation for the expression levels of each gene.
in \( \mathbf{G} \), see Supplementary Materials. Since the EES is linear, we rescale the expression levels of each gene by its (mean) value in the wildtype. Table I gives these rescaled expression levels for the internal variables \([arcA]\), \([appY]\), \([fnr]\), \([oxyR]\), and \([soxS]\) under anaerobic glucose minimal medium conditions.

Note that error estimates for the expression levels of several genes is large. This is the reason that a reduced network is essential in order to make useful predictions. Second, as seen from Table I, reported expression levels of the gene \(G_k\) in the mutant \(\Delta G_k\) is non-zero. Presumably, what is measured are non-functional analogs of the corresponding genes. In calculating the EES, we set these expression levels (shown in parentheses in Table I) to zero.

The component of the EES for the internal variables is

\[
\mathbf{B}^{(\text{E.coli})}_{\text{int}} = \begin{pmatrix}
1.00 & -0.62 & -0.66 & 0.72 & -0.07 \\
-0.21 & 1.00 & 0.15 & -0.26 & -0.20 \\
0.25 & -0.30 & 1.00 & 0.23 & 0.13 \\
-0.24 & 0.16 & -0.01 & 1.00 & 0.20 \\
-0.01 & 0.06 & -0.26 & -0.08 & 1.00
\end{pmatrix}.
\]

(5)

The next step is to compute the EES predictions for \([appY]\), \([oxyR]\), and \([soxS]\) in the double knockout \(\Delta arcA\Delta fnr\). This is done using the matrix (5) and setting \([arcA]\) and \([fnr]\) to zero. Expression levels of the remaining genes in \(\mathbf{S}\), predicted using the EES, are \([appY]_{\text{EES}} = -0.23\), \([oxyR]_{\text{EES}} = 0.91\), and \([soxS]_{\text{EES}} = 0.78\). We need to determine, at a 5% level of confidence, if these predicted values are consistent with those from the replicates of the double mutant. The comparison is made using the \(t\)-test (\texttt{ttest} in MATLAB, The Mathworks, Inc.), and it is found that the null hypothesis, that the computed gene expression levels equal the mean of the experimental data, is rejected at the 5% level for \(appY\) and \(oxyR\).

Next, we implement the analysis for genes outside of \(\mathbf{S}\). The null hypothesis cannot be rejected at the 5% level for 210 of the 294 genes. The three experimental values of the expression level of each gene in the double knockout, the corresponding predictions of the EES, and the test statistic \(t\) are given in Supplementary Materials. Since the Student’s distribution associated with the comparison has two degrees of freedom, the null hypothesis is rejected when \(|t| > 4.303\). The histogram of the test statistic for the 299 genes is shown in Figure 4(A). In Supplementary Materials, we highlight the genes for which the null hypothe-
sis is rejected. We emphasize that, unlike in many prior studies whose assertions are limited to whether genes in mutants are up/down regulated, our predictions are quantitative.

The proximity of the predicted and experimental values is not due to a lack of variability in the expression levels of genes in $G$. We verify this by computing the differential expression of genes in the mutants. Figure 4(B) shows the histogram of the largest deviations from the wildtype, normalized by the standard deviation (between replicates) in the wildtype. Expression levels of over half the genes deviate by more than 2 standard deviations.

III. DISCUSSION

The determination of the most effective and least detrimental treatment to prevent the onset of hereditary diseases requires knowledge of the underlying gene regulatory networks. Unfortunately these networks contain hundreds of genes, proteins, and other molecules whose interactions are only known partially [5, 8–10]. The necessary changes can only be computed with simplified, yet predictive models of the regulatory networks. The goal of our work is to outline a new approach to construct such models.

Our approach builds on the influence network, whose nodes are the genes in the regulatory network. DNA microarray experiments on systematic genetic perturbations can be used to find multiple solutions of an influence network [10]. However, making accurate quantitative predictions from influence networks is difficult because they contain a large number of nodes; specifically, in inverting the expression level data to compute an influence network, the error estimates are magnified to a level where making useful predictions becomes impossible.

The reduction we propose has two components. We first identify a (relatively) small set $S$ of $n$ nodes (internal variables) in the influence network which can be used to affect the remaining genes. Next, we limit consideration to steady states of the network as the internal variables are modified. Finally, this solution surface $S_S$ is approximated using the unique $n$-dimensional plane $H_S$ defined by the gene expression levels of the wildtype and the $n$ single knockout mutants in $S$; the model system whose solutions lie on the plane is the EES.

In general, the computation of the EES can be done with mutants whose expression levels are up/down regulated externally to values other than zero (homozygous knockout mutants). As an example, one could use gene expression levels of heterozygous mutants to construct the EES. In principle, we could also up/down regulate the expression levels of
genes chosen for $S$, say, using transfection [33, 34]. However, to the best of our knowledge, consequences of such external actions on the gene expression levels within a cell are not known precisely. If and when they are known, the information can be easily incorporated into the construction of the EES.

The EES can only be used to address a limited set of questions, specifically issues regarding the expression levels of genes in $G$ in the steady state when those in $S$ are varied externally. For example, it can be used to predict gene expression levels in double knockout mutants. We tested our algorithms using previously published data on a double knockout mutant in an oxygen deprivation network of $E.coli$. (Here, as in most cases, the form of the network is unknown.) We identified the group of 299 genes to be studied using the Gene Ontology database. The EES was computed using the expression levels of five single knockout mutants, and used to predict their expression levels in the double mutant. These prediction were significantly different from the experimentally obtained expression levels in less than 30% of genes.

Interestingly, the EES can be used to compute how expression levels of genes within $S$ need to be changed so that the steady state of the entire network is moved from its initial state $P_0$ to, or as close as possible to, a pre-specified position $P_{aim}$. We showed through an example that the solution computed using the EES is close to that of the full network. However, the efficacy of the move depends on the proximity of $P_{aim}$ to the surface $H_S$. If $P_{aim}$ is far from $H_S$, then the set of internal variables need to be expanded in order to find acceptable solutions.

We believe that approaches similar to those outlined here can prove useful in treating complex genetic diseases by helping identify optimal combinations of up/down regulation of genes (or optimal combinations of single target drugs) that have minimal side effects and are most effective in moving the equilibrium of the network in its entirety to a preferred state. We hope our work motivates additional studies on these issues.

IV. MATERIALS AND METHODS

Construction of the EES: As illustrated in Figures 1, the EES is constructed so that, as internal variables are modified, the solutions of the system lie on the $n$-dimensional plane $H_S$. Thus the external variables are linear in $x_k$’s, and consequently, have the form given
by Eqns. (2). We need to compute the coefficients $a_{Ki}$ for $K = (n + 1), (n + 2), \ldots, N$ and $i = 1, \ldots, n$. This is done by noting that the expression levels $P^{(m)}$ of each of the $n$ mutants $\Delta G_1, \Delta G_2, \ldots, \Delta G_n$ satisfies Eqn. (2), thus providing the conditions necessary to compute these $a_{Ki}$’s.

Moreover, the internal variables themselves are inter-related. For example, in a single knockout mutant, the expression levels of the remaining genes are solved using $(N - 1)$ equations of (1). To find the dependence of $x_1$ on the remaining internal variables, let us reduce the set of internal variables to $\{x_2, x_3, \ldots, x_n\}$; $x_1$ is now an external variable. Hence, with the method introduced above, $x_1$ is a linear combination of the remaining internal variables. Since $P^{(0)}$ is one solution of the system

$$x_1 - p^{(0)}_1 = \sum_{i=2}^{n} a_{1i} \left( x_i - p^{(0)}_i \right).$$

The values of $x_1$ in $P_i$, $2 \leq i \leq n$, determine the coefficients $a_{1i}$. Similar relationships are obtained for the other internal variables; this is Eqn. (3).

V. ACKNOWLEDGEMENTS

The authors would like to thank Tim Cooper, John Miller, and Gregg Roman for discussions.


**Figure Legends**

**FIG. 1:** (A) An example of an $n$-dimensional surface $S_S$ of solutions of the system (1) when the expression levels of internal variables are modified. The example is chosen from the synthetic network introduced in the Results Section. The surface shown represents the expression levels of the external variable $X_7$ as the internal variables $x_1$ and $x_2$ are modified. (B) The point $P_0$ representing expression levels of the wildtype and points $P_m$, $m = 1, 2$ representing expression levels of single knockout mutants $\Delta G_m$ lie on this surface. The EES is defined so that its solutions lie on the unique 2-dimensional plane (blue) $H_S$ passing through $P_0$, and $P_m$, $m = 1, 2$. As can be seen, due to restrictions imposed on the EES, the surfaces $S_S$ and $H_S$ are close.

**FIG. 2:** Suppose, we are required to move the equilibrium of the system from $P_0$ to $P_{aim}$ by implementing changes within $S$. In general this is not possible because interactions between nodes force the steady state to remain on $S_S$. However, it is possible to compute $P_{lin}$, the point closest to $P_{aim}$ that can be reached by the EES. Due to the proximity of $S_S$ and $H_S$, the point $P_{sys}$ obtained by projecting $P_{lin}$ to $S_S$ is close to $P_{lin}$. Thus, it is possible to pre-determine if the new steady state of the system, forced by changes made in $S$, is acceptable.

**FIG. 3:** A schematic of the synthetic network used in our calculations. The 21 genes in the system consists of 3 groups, each with 7 genes. Genes within a cluster are coupled by interactions whose intensity is chosen randomly. Genes between clusters are weakly coupled. The “mutants” $\Delta G_K$ shown in black are not viable; i.e., the corresponding set of equations do not have a solution. Genes shown in red are used to construct the effective empirical subnetwork.
FIG. 4: (A) The comparison of the predictions of the EES and the experimental gene expression levels of the double knockout mutant $\Delta arcA\Delta fnr$. The Figure shows the histogram of the test statistic of the Student’s distribution (two degrees of freedom) for the 299 genes chosen for the study. (B) The good agreement in (A) is not due to lack of variation in the gene expression levels between the wildtype and mutants. The histogram shows the largest differential expression level of mutants, normalized by the standard deviation for the wildtype (computed from the four replicates given in the data set GSE1121 of the GEO database [32]).
### Tables

<table>
<thead>
<tr>
<th></th>
<th>Wildtype</th>
<th>∆appY</th>
<th>∆arcA</th>
<th>∆fnr</th>
<th>∆oxyR</th>
<th>∆soxS</th>
<th>∆arcA∆fnr</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>appY</em></td>
<td>1.00 ± 0.34</td>
<td>(0.03 ± 0.01)</td>
<td>0.31 ± 0.13</td>
<td>0.35 ± 0.02</td>
<td>1.66 ± 0.76</td>
<td>0.72 ± 0.14</td>
<td>0.34 ± 0.04</td>
</tr>
<tr>
<td><em>arcA</em></td>
<td>1.00 ± 0.20</td>
<td>0.72 ± 0.05</td>
<td>(0.12 ± 0.02)</td>
<td>0.93 ± 0.08</td>
<td>0.86 ± 0.01</td>
<td>0.77 ± 0.11</td>
<td>(0.08 ± 0.02)</td>
</tr>
<tr>
<td><em>fnr</em></td>
<td>1.00 ± 0.17</td>
<td>1.21 ± 0.02</td>
<td>0.88 ± 0.02</td>
<td>(0.07 ± 0.02)</td>
<td>1.04 ± 0.07</td>
<td>1.09 ± 0.03</td>
<td>(0.07 ± 0.01)</td>
</tr>
<tr>
<td><em>oxyR</em></td>
<td>1.00 ± 0.02</td>
<td>0.80 ± 0.19</td>
<td>0.99 ± 0.22</td>
<td>0.91 ± 0.12</td>
<td>(0.09 ± 0.03)</td>
<td>1.18 ± 0.19</td>
<td>0.80 ± 0.04</td>
</tr>
<tr>
<td><em>soxS</em></td>
<td>1.00 ± 0.08</td>
<td>1.05 ± 0.21</td>
<td>1.02 ± 0.14</td>
<td>0.73 ± 0.13</td>
<td>0.94 ± 0.21</td>
<td>(0.04 ± 0.01)</td>
<td>0.76 ± 0.10</td>
</tr>
</tbody>
</table>

**TABLE I:** Rescaled expression levels of *appY*, *arcA*, *fnr*, *oxyR*, and *soxS* in the wildtype *E.coli*, single knockout mutants, and the double knockout mutant ∆arcA∆fnr under anaerobic glucose minimal medium conditions. The data have been rescaled by the mean value of the expression levels in wildtype cells. The mean and standard errors are calculated from the replicates given in the data set GSE1121 of the GEO site www.ncbi.nlm.nih.gov. The values in parentheses are set to zero in computing the EES.