OPEN ACCESS Freely available online

PLoS one

# Using Effective Subnetworks to Predict Selected Properties of Gene Networks

**Gemunu H. Gunaratne[1]*, Preethi H. Gunaratne[2,3], Lars Seemann[1], Andrei Török[4]**

**1** Department of Physics, University of Houston, Houston, Texas, United States of America, **2** Department of Biology and Biochemistry, University of Houston, Houston, Texas, United States of America, **3** Human Genome Sequencing Center and Department of Pathology, Baylor College of Medicine, Houston, Texas, United States of America, **4** Department of Mathematics, University of Houston, Houston, Texas, United States of America

## Abstract

***Background:*** Difficulties associated with implementing gene therapy are caused by the complexity of the underlying regulatory networks. The forms of interactions between the hundreds of genes, proteins, and metabolites in these networks are not known very accurately. An alternative approach is to limit consideration to genes on the network. Steady state measurements of these *influence networks* can be obtained from DNA microarray experiments. However, since they contain a large number of nodes, the computation of influence networks requires a prohibitively large set of microarray experiments. Furthermore, error estimates of the network make verifiable predictions impossible.

***Methodology/Principal Findings:*** Here, we propose an alternative approach. Rather than attempting to derive an accurate model of the network, we ask what questions can be addressed using lower dimensional, highly simplified models. More importantly, is it possible to use such robust features in applications? We first identify a small group of genes that can be used to affect changes in other nodes of the network. The reduced effective empirical subnetwork (EES) can be computed using steady state measurements on a small number of genetically perturbed systems. We show that the EES can be used to make predictions on 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.

***Conclusions/Significance:*** The constraints imposed by gene expression levels of mutants can be used to address a selected set of questions about a gene network.

**Citation:** Gunaratne GH, Gunaratne PH, Seemann L, Török A (2010) Using Effective Subnetworks to Predict Selected Properties of Gene Networks. PLoS ONE 5(10): e13080. doi:10.1371/journal.pone.0013080

**Editor:** Johannes Jaeger, Centre for Genomic Regulation (CRG), Universitat Pompeu Fabra, Spain

**Received** May 4, 2010; **Accepted** August 30, 2010; **Published** October 8, 2010



**Funding:** This work was partially funded by grant 0607345 from the National Science Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: gemunu@uh.edu

## 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 characteris-

tics such as bifurcations. In *gene influence networks* [14,20], a gene, its transcript, and protein are represented by a single node, which is quantified by the expression level of the mRNA. Regulatory interactions between nodes of an influence network include actions mediated by other components in the network.

Consider an influence network containing $N$ genes $\mathbf{G}=\{G_1,G_2,\ldots,G_N\}$; denote the expression level of the $K^{th}$ gene by $X_K$ and the state of the network by $\mathbf{X}\equiv\{X_1,X_2,\ldots X_N\}$. The influence network can be modeled by a set of ordinary differential equations $\mathbf{F}:\mathbb{R}^N\rightarrow\mathbb{R}^N$

$$\begin{aligned}\dot{X}_1 &= F_1(\mathbf{X})\\ \dot{X}_2 &= F_2(\mathbf{X})\\ \ldots\quad &\cdot\quad \ldots\\ \dot{X}_N &= F_N(\mathbf{X}).\end{aligned}\qquad(1)$$

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

We propose an alternative approach. Rather than attempting to construct an accurate model of a gene network, we ask what questions on the network can be addressed (perhaps approximately) using low-dimensional and highly simplified effective models constructed from empirical data. What data would be needed for the construction? Will issues addressed through the approach be useful in applications?

We first note 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 may also include microRNAs within the cluster, each of which affect many genes through post-transcriptional regulation [28,29], even though their fold induction on each gene is small.

Suppose we have partitioned the $N$ genes of the influence network into clusters, and identified a small number of genes/microRNAs from each cluster that can be used to control the expression levels of the remaining genes. Denote the set of these nodes by $\mathbf{S}$. The number $n$ of nodes in $\mathbf{S}$ is much smaller than $N$. We will represent their expression levels by $\mathbf{x}=\{x_1,x_2,\ldots,x_n\}$, 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 $\mathbf{X}=\{\mathbf{x}_{int},\mathbf{X}_{ext}\}$ where we will refer to $\mathbf{x}_{int}=\{x_1,x_2,\ldots,x_n\}$ and $\mathbf{X}_{ext}=\{X_{n+1},\ldots,X_N\}$, as "internal" and "external" variables respectively.

In this paper, we limit consideration to networks with steady state solutions. When external perturbations are made on genes within $\mathbf{S}$, 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 $\mathcal{S}_S$. 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 $\dot{\mathbf{X}}=\mathbf{F}(\mathbf{X})$. First, we assume that the unconstrained system has a unique stable steady state which we denote by $\mathbf{P}^{(0)}=\left\{\mathbf{p}^{(0)}_{int},\mathbf{P}^{(0)}_{ext}\right\}=\left\{p^{(0)}_1,\ldots,p^{(0)}_n,P^{(0)}_{n+1},\ldots,P^{(0)}_N\right\}$. It satisfies the $N$ equations $\mathbf{F}\left(\mathbf{P}^{(0)}\right)=0$. The point $\mathcal{P}_0$ representing it lies on $\mathcal{S}_S$. 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 $\mathbf{P}^{(m)}=\left\{\mathbf{p}^{(m)}_{int},\mathbf{P}^{(m)}_{ext}\right\}$ with $p^{(m)}_m=0$, and represent it by $\mathcal{P}_m$. Since the equilibrium is associated with changes made within the set $\mathbf{S}$, $\mathcal{P}_m$ lies on $\mathcal{S}_S$.

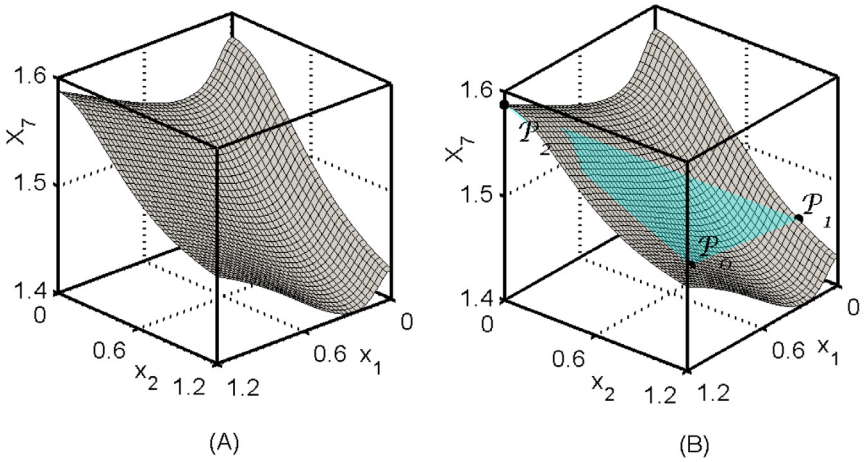


**Figure 1. Example of an $n$-dimensional solution surface $\mathcal{S}_S$ of (1).** The example is chosen from the synthetic network introduced in the Results Section. (a) The surface shown represents the expression levels of the external variable $X_7$ as the internal variables $x_1$ and $x_2$ are modified. (b) The point $\mathcal{P}_0$ representing expression levels of the wildtype and points $\mathcal{P}_m$, $m=1,2$ 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) $\mathcal{H}_S$ passing through $\mathcal{P}_0$, and $\mathcal{P}_m$, $m=1,2$. As can be seen, due to restrictions imposed on the EES, the surfaces $\mathcal{S}_S$ and $\mathcal{H}_S$ are close.
doi:10.1371/journal.pone.0013080.g001

Consequently, $\mathcal{P}_0$ as well as $\mathcal{P}_m$ for $m=1,2,\ldots,n$ lie on $\mathcal{S}_S$, 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 $\mathcal{S}_S$. Observe that $\mathcal{P}_0$ and the $n$ points $\mathcal{P}_m$ define a unique $n$-dimensional plane in $\mathbb{R}^N$, which we denote by $\mathcal{H}_S$. Figure 1(b) shows the surface for the example above. The EES describes the set $\mathcal{H}_S$ as parametrized by the internal variables. Since both $\mathcal{S}_S$ and $\mathcal{H}_S$ contain the points $\mathcal{P}_0$, and $\mathcal{P}_m$, $m=1,2,\ldots,n$), we expect them to be close in the region of interest.

Observe that the $EES:\mathbb{R}^n\rightarrow\mathbb{R}^N$ is a linear function determined by $\mathcal{P}_0$, and $\mathcal{P}_m$, $(m=1,2,\ldots,n)$, but is otherwise independent of $\mathbf{F}$. In particular, each $X_K$ is a linear function of $x_1,\ldots,x_n$. Since $\mathcal{P}_0$ lies on $\mathcal{H}_S$

$$X_K-P_K^{(0)}=\sum_{i=1}^{n}a_{Ki}\left(x_i-p_i^{(0)}\right), \tag{2}$$

for each $K=(n+1),\ldots,N$. The coefficients $a_{Ki}$ can be evaluated by noting that $\mathcal{P}_1,\mathcal{P}_2,\ldots,\mathcal{P}_n$ lie on $\mathcal{H}_S$. There is one additional complication, that we illustrate using the following example. Suppose we consider a mutant where only $x_1$ is externally set. The remaining expression levels of the steady state of this mutant are solved using 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. Thus, we expect there to be relationships between the internal variables as well. As we show in the Methods Section, these dependencies can be assumed to take the form

$$x_k-p_k^{(0)}=\sum_{i\neq k}a_{ki}\left(x_i-p_i^{(0)}\right), \tag{3}$$

for $k=1,2,\ldots,n$.

We have thus implemented two significant simplifications. The original system $\mathbf{F}:\mathbb{R}^N\rightarrow\mathbb{R}^N$ contained a large number ($N\sim$ several hundred) of nonlinearly coupled variables. In contrast, the $EES:\mathbb{R}^n\rightarrow\mathbb{R}^N$ has a small number ($n\sim10$) of internal variables, is linear, and can be constructed using the steady state solutions of the original system (wild-type) and $n$ single knockout mutants. Clearly, the EES is not an accurate representation of the original network. The issue is whether there are questions about $\mathbf{F}$ that can be addressed using the EES. As we show below, this is indeed the case due to geometrical constraints imposed on the solution surface. Specifically, the EES can be used to predict, approximately, the expression levels of all nodes in $\mathbf{F}$, when external changes are made within $\mathbf{S}$; *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 equilibrium of the system can be moved from its initial state $\mathcal{P}_0$ to a pre-specified set of expression levels defined by a point $\mathcal{P}_{aim}$, see Figure 2. Since $\mathcal{P}_{aim}$ will, in general, not lie on the solution surface, it cannot be reached through changes within $\mathbf{S}$. Instead, we can use the EES to compute $\mathcal{P}_{lin}$, which is the closest point to $\mathcal{P}_{aim}$ on the plane $\mathcal{H}_S$, see Figure 2. Since the surfaces $\mathcal{S}_S$ and $\mathcal{H}_S$ are close, 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.

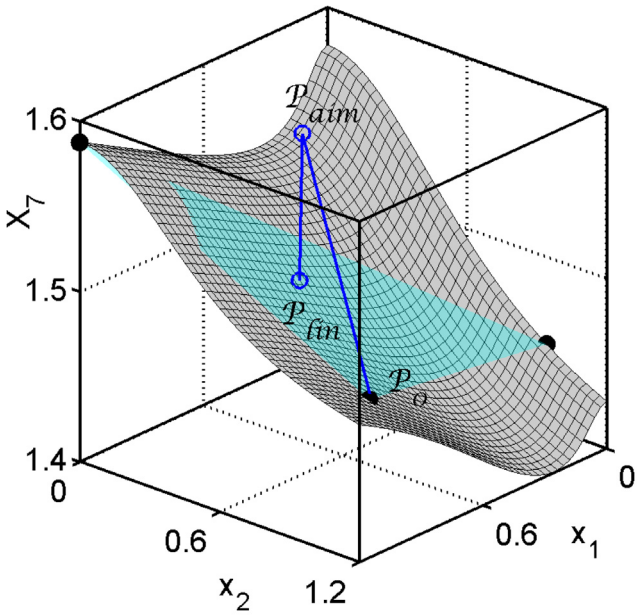


**Figure 2. Moving the equilibrium from $\mathcal{P}_0$ to $\mathcal{P}_{aim}$ by implementing changes within S.** In general this is not possible because interactions between nodes force the equilibrium to remain on $\mathcal{S}_S$. However, it is possible to compute $\mathcal{P}_{lin}$, the point closest to $\mathcal{P}_{aim}$ that can be reached by the EES. Due to the proximity of $\mathcal{S}_S$ and $\mathcal{H}_S$, the point $\mathcal{P}_{sys}$ obtained by projecting $\mathcal{P}_{lin}$ to $\mathcal{S}_S$ is close to $\mathcal{P}_{lin}$. Thus, it is possible to pre-determine if the movement of the equilibrium forced by the changes made in **S** are acceptable.
doi:10.1371/journal.pone.0013080.g002

## Results

### A Synthetic Influence Network

In the model system, $F_K(\mathbf{X})$ is a linear combination of sigmoidal Hill functions; specifically,

$$F_K(\mathbf{X})=X_K\sum_{I=1}^{N}a_{KI}\left[H(X_I;c_{KI})-H\left(P_I^{(0)};c_{KI}\right)\right], \tag{4}$$

where $H(X;c)=X^h/\left(X^h+c^h\right)$ 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 $\mathbf{X}$ of the system. The action is activating if $a_{KI}>0$ and inhibiting if $a_{KI}<0$. The system is constructed so that $\mathbf{P}^{(0)}$ is a steady state of Eqns. (1). Numerically, we find that model systems defined by Eqns. (1) and (4), have at most one stable equilibrium. We suspect that this is due to restrictions imposed by the fact that the sign of each partial derivative $\partial F_K/\partial X_I$ is independent of the state of the system.

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 solution 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 $\mathbf{P}^{(0)}$ 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 $\mathbf{P}^{(0)}$ are $J_{KI}=a_{KI}H'(P^{(0)};c_{KI})$; thus $a_{KI}$ can be computed for a given set

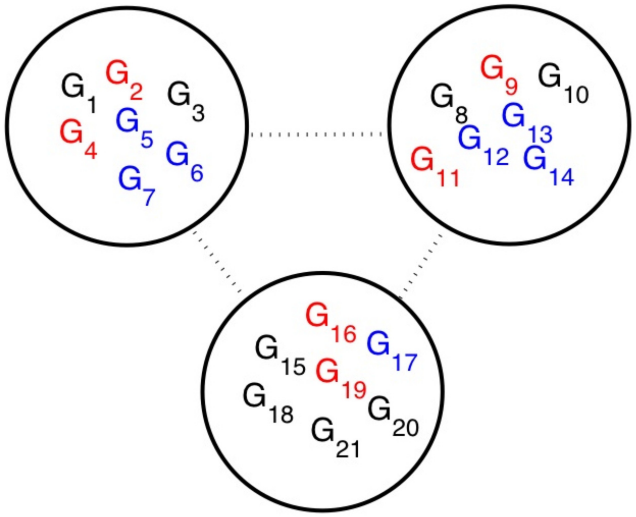


**Figure 3. A schematic of the synthetic network.** The 21 genes in the system consists of 3 groups, each with 7 genes. Genes within a cluster are coupled by interactions whose intensity is chosen randomly. Genes between clusters are weakly coupled. The "mutants" $\Delta G_K$ shown in black are not viable; i.e., the corresponding set of equations do not have a solution. Genes shown in red are used to construct the effective empirical subnetwork.
doi:10.1371/journal.pone.0013080.g003

of $J_{KI}$'s. Since we require $\mathbf{P}^{(0)}$ 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, $\mathbf{S}=\{G_2,G_4,G_9,G_{11},G_{16},G_{19}\}$ (genes marked in red in Figure 3). The variables $X_K$, $K=7,8,\ldots,21$ are re-indexed as described before. The $EES:\mathbb{R}^6\rightarrow\mathbb{R}^{21}$ is computed using the expression levels of all 21 genes at $\mathcal{P}_0$, $\mathcal{P}_1$, $\mathcal{P}_2$, $\mathcal{P}_3$, $\mathcal{P}_4$, $\mathcal{P}_5$ and $\mathcal{P}_6$.

In order to illustrate the proximity of $\mathcal{H}_S$ to $\mathcal{S}_S$, 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 $\mathcal{P}_0$; 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 $\mathbf{S}$, and the 15 nodes outside of $\mathbf{S}$ are compared. We differentiate between these two groups.

Results for the first group (genes in $\mathbf{S}$) 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 $\mathbf{S}$) 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 $\mathbf{P}^{(0)}=\left\{\mathbf{p}_{int}^{(0)},\mathbf{P}_{ext}^{(0)}\right\}$, with

$$\mathbf{p}_{int}^{(0)}=\{0.89,0.97,1.17,0.63,0.51,0.51\},$$
$$\mathbf{P}_{ext}^{(0)}=\{1.47,0.74,0.83,1.24,0.58,1.03,0.85,1.17,$$
$$0.96,1.39,1.40,0.53,1.21,0.68,1.15\}.$$

We want to find out how the expression levels of genes in $\mathbf{S}$ need to be changed so that the system moves to, or 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 $\mathcal{P}_{aim}$ (see Figure 2) given by $\mathbf{P}^{(aim)}=\{\mathbf{p}_{int}^{(aim)},\mathbf{P}_{ext}^{(aim)}\}$, where

$$\mathbf{p}_{int}^{(aim)}=\{0.8,0.6,1.3,0.7,0.6,0.6\},$$
$$\mathbf{P}_{ext}^{(aim)}=\{1.6,0.8,0.9,1.3,0.5,1.1,0.9,1.2,0.9,1.2,1.5,0.4,1.3,0.6,1.1\}.$$

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

$$\mathbf{p}_{int}^{(lin)}=\{0.87,0.75,1.30,0.75,0.57,0.45\},$$
$$\mathbf{P}_{ext}^{(lin)}=\{1.49,0.73,0.84,1.08,0.56,1.03,0.89,1.18,$$
$$0.87,1.22,1.48,0.46,1.22,0.67,1.15\}.$$

Finally, we use the model system Eqns. (1) and (4) to compute the external variables when internal variables are fixed at $\mathbf{P}_{int}^{(lin)}$. It is found to be $\mathbf{P}^{(sys)}=\{\mathbf{p}_{int}^{(sys)},\mathbf{P}_{ext}^{(sys)}\}$, where $\mathbf{p}_{int}^{(sys)}=\mathbf{p}_{int}^{(lin)}$, and

$$\mathbf{P}_{ext}^{(sys)}=\{1.48,0.75,0.85,1.11,0.58,1.00,0.91,1.21,$$
$$0.95,1.25,1.39,0.54,1.21,0.67,1.14\}.$$

The Euclidean distances between the points are $d(\mathcal{P}_0,\mathcal{P}_{aim})=0.55$, $d(\mathcal{P}_0,\mathcal{P}_{lin})=0.40$, $d(\mathcal{P}_{aim},\mathcal{P}_{lin})=0.38$, and $d(\mathcal{P}_{lin},\mathcal{P}_{sys})=0.15$. Thus, we attempted to move the equilibrium from $\mathcal{P}_0$ to a point $\mathcal{P}_{aim}$ that was a distance 0.55 away, but were only able to move it on $\mathcal{H}_S$ to a point $\mathcal{P}_{lin}$, which is a distance 0.40 away from $\mathcal{P}_{aim}$. However, $\mathcal{P}_{lin}$ is only a distance 0.15 from the point $\mathcal{P}_{sys}$, which is the solution of the original system when expression levels of the internal variables are set to $\mathbf{p}_{int}^{(lin)}$. We have found that $\mathcal{P}_{lin}$ and $\mathcal{P}_{sys}$ are close in studies of several model systems and for many points $\mathcal{P}_{aim}$. Thus, the EES can be used to pre-determine, approximately, the equilibrium of the original network when changes made within $\mathbf{S}$.

## Transcriptional Regulatory Network in *E.coli*

The EES can be constructed using microarray data from the wildtype and single knockout mutants of genes in $\mathbf{S}$. It can then be used to predict gene expression levels of other mutants. This 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.

It should be noted that gene expression levels in *E.coli* are unlikely to be in a steady state; rather, the expression levels reported in Ref. [32] are averages from a group of cells in various stages in the cell cycle. The analysis in this Section assumes that the computation of the EES and its predictions are valid for these averages. Preliminary results from our current work on systems exhibiting circadian rhythms validate this assumption.

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 Supporting Information S1.

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 **G**, see Supporting Information S1. Since the EES is linear, we rescale the expression levels of each gene by its (mean) value in the wildtype. Table 1 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 1, 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 1) to zero.

The component of the EES for the internal variables is

$$\mathbf{B}_{int}^{(E.coli)} = \begin{pmatrix} 1.00 & -0.62 & -0.66 & 0.72 & -0.07 \\ -0.21 & 1.00 & 0.15 & -0.26 & -0.20 \\ 0.25 & -0.30 & 1.00 & 0.23 & 0.13 \\ -0.24 & 0.16 & -0.01 & 1.00 & 0.20 \\ -0.01 & 0.06 & -0.26 & -0.08 & 1.00 \end{pmatrix}. \quad (5)$$

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 **S**, predicted using the EES, are $[appY]_{EES} = -0.23$, $[oxyR]_{EES} = 0.91$, and $[soxS]_{EES} = 0.78$. 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 (`ttest` in MATLAB, The Mathworks, Inc.), and it is found that the null hypothesis, that experimental data comes from a (normal) distribution with mean equal to the computed gene expression level, is rejected at the 5% level only for *appY*.

Next, we implement the analysis for genes outside of **S**. The null hypothesis cannot be rejected at the 5% level for 213 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 Supporting Information S1. 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 Supporting Information S1, we highlight the genes for which the null hypothesis 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.

## Discussion

An accurate model of the gene regulatory network associated with a hereditary disease can be used to compute the most effective and least detrimental treatment to prevent its onset. Unfortunately these networks contain hundreds of genes, proteins, and other molecules whose interactions are only partially known [5,8–10]. It is unlikely that detailed models of such networks will be available in the near future. The question raised in the paper is whether information needed to move the steady state of a network can be deduced from an analysis of highly simplified, empirically determined models. The data used for analysis is obtained from microarray experiments.

Our approach is as follows. We first identify a (relatively) small set **S** of $n$ nodes (internal variables) in the influence network which

**Table 1.** Normalized gene expression levels in the wildtype and mutants.

| | *Wildtype* | $\Delta appY$ | $\Delta arcA$ | $\Delta fnr$ | $\Delta oxyR$ | $\Delta soxS$ | $\Delta arcA\Delta fnr$ |
|---|---|---|---|---|---|---|---|
| *appY* | 1.00±0.34 | (0.03±0.01) | 0.31±0.13 | 0.35±0.02 | 1.66±0.76 | 0.72±0.14 | 0.34±0.04 |
| *arcA* | 1.00±0.20 | 0.72±0.05 | (0.12±0.02) | 0.93±0.08 | 0.86±0.01 | 0.77±0.11 | (0.08±0.02) |
| *fnr* | 1.00±0.17 | 1.21±0.02 | 0.88±0.02 | (0.07±0.02) | 1.04±0.07 | 1.09±0.03 | (0.07±0.01) |
| *oxyR* | 1.00±0.02 | 0.80±0.19 | 0.99±0.22 | 0.91±0.12 | (0.09±0.03) | 1.18±0.19 | 0.80±0.04 |
| *soxS* | 1.00±0.08 | 1.05±0.21 | 1.02±0.14 | 0.73±0.13 | 0.94±0.21 | (0.04±0.01) | 0.76±0.10 |

Rescaled expression levels of *appY*, *arcA*, *fnr*, *oxyR*, and *soxS* in the wildtype *E.coli*, single knockout mutants, and the double knockout mutant $\Delta arcA\Delta fnr$ 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.
doi:10.1371/journal.pone.0013080.t001

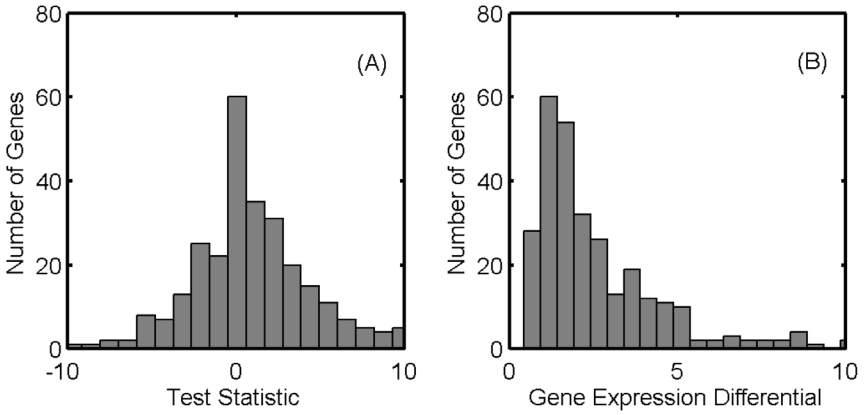


**Figure 4. Comparison of EES predictions with experimental gene expression levels of $\Delta arcA\Delta fnr$.** (a) The histogram of the test statistic of the Student's distribution (two degrees of freedom) for the 299 genes chosen for the study. (b) The good agreement in (a) is not due to lack of variation 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]).
doi:10.1371/journal.pone.0013080.g004

can be used to affect the remaining genes. (Mathematically, for each external variable $X_I$, we require one or more of $\partial F_I/\partial x_k$, where $x_k$ are the internal variables, to be non-vanishing.) Next, we limit consideration only to steady states of the network as internal variables are modified. Finally, this solution surface $\mathcal{S}_S$ is approximated using the unique $n$-dimensional plane $\mathcal{H}_S$ defined by the gene expression levels of the wildtype and the $n$ single knockout mutants in $\mathbf{S}$; the model system whose solutions lie on the plane is the EES.

Some genes may be critical in the sense that the organism may not be viable when they are knocked out. There were similar nodes (shown in black in Figure 3) in our synthetic model. In our approach they cannot be used as internal nodes. However, if they need to be utilized, the EES can be computed using heterozygous mutants or those where the expression level is up/down regulated to a value other than zero through, for example, transfection [33,34].

We emphasize that, due to the reduced dimensionality and its linearity, we do not expect the EES to be an accurate model of the original system. However, because of the geometrical constraints, it is possible to use the EES to (approximately) compute answers to a very limited set of questions about the system. Specifically, they are questions on gene expression levels when external changes are made within $\mathbf{S}$. As an example, the EES can predict gene expression levels in double knockout mutants. We tested the predictions using previously published data on a double knockout mutant in an oxygen deprivation network of *E.coli*. (Here, as in most cases, the underlying 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. The predictions were significantly different from the experimentally obtained expression levels for less than 30% of genes.

Interestingly, the EES can be used to compute how expression levels of genes within $\mathbf{S}$ need to be changed so that the equilibrium of the entire network is moved from its initial state $\mathcal{P}_0$ to, or as close as possible to, a pre-specified position $\mathcal{P}_{aim}$. 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 $\mathcal{P}_{aim}$ to the surface $\mathcal{H}_S$. If $\mathcal{P}_{aim}$ is far from $\mathcal{H}_S$, then the set of internal variables need to be expanded in order to find acceptable solutions.

Before concluding, we briefly address a few issues; the first is the observation that, in the parameter range considered, the model system given by Eqns. (1) and (4) have at most one stable steady state. Even though we required $\mathbf{P}^{(0)}$ to be stable (by an appropriate choice of eigenvalues of the linearization), non-linear systems can, in general, be expected to have additional solutions. However, our model has a special feature: the signs of the partial derivatives $\partial F_K/\partial X_J$ are independent of the state of the system. The analogous biological statement is that, if nodes $J$ and $K$ are isolated, the action of node $J$ on node $K$ increases in magnitude as $X_J$ increases. Is this condition, combined with the choice of eigenvalues, sufficient to guarantee a uniques stable solution? We are currently studying this question. It should be noted that the uniqueness of solutions has been proven for several other classes of monotonic nonlinear systems [35–37].

The second issue involves the partitioning of genes into clusters and the choice of internal variables. Internal variables in the oxygen deprivation network of *E.coli* were already determined from the experiments reported in Ref. [32]. We used the GO classification to identify nodes belonging to the network. Different approaches can be used to partition genes into clusters when biological classifications are not available. For example, one could use topological (*e.g.*, persistent homology [38,39]) or graph theoretic (*e.g.*, spectral clustering [23], community clustering [24]) methods. Integrated genomic analysis, which successfully identified subtypes of gliobastoma [40], can also be used in clustering genes through the use of heat maps [41,42]. The choice of internal variables requires biological input. Mathematically, the requirement is that each node in the cluster can be affected by suitable changes in internal variables. As we mentioned, genes that translate to transcription factors, or microRNAs [28,29] within the cluster, could act as internal nodes.

Third, can one estimate the proximity of $\mathcal{H}_S$ to the solution surface $\mathcal{S}_S$? Differences in gene expression levels of double knockout mutants are one measure of the proximity. Alternatively, we could use the corresponding differences in heterozygous single knockouts (whose expression levels are roughly half of the wildtype) and the predictions of the EES.

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 studies on this issue.

## 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 $\mathcal{H}_S$. 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 $\mathbf{p}^{(m)}$ 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 $a_{Ki}$'s.

We note, however, that the internal variables themselves are interrelated. For example, in the single knockout mutant $\Delta G_1$, all expression levels (other than $x_1$) are determined by solving the last $(N-1)$ equations of (1). Thus, we need to derive relationships between the internal variables. Consider for example, the dependence of $x_n$ on the remaining internal variables. In order to find its form, let us reduce the set of internal variables to $\{x_1,x_2,\ldots,x_{n-1}\}$; $x_n$ is now an external variable. Hence, with the approximations used in the paper, $x_n$ is a linear combination of the remaining internal variables. Since $\mathbf{P}^{(0)}$ is one solution of the system

$$x_k-p_k^{(0)}=\sum_{i=1}^{n-1}a_{ki}\left(x_i-p_i^{(0)}\right). \qquad (6)$$

Similar relationships are obtained for the other internal variables.

## Supporting Information

**Supporting Information S1** Tables showing (1) The set G of 299 genes chosen to study the oxygen deprivation network of E.coli. These genes have the common biological function 0006355 "regulation of transcription, DNA-dependent" (2) Mean values of the 299 genes in the wildtype and the mutants. (3) Standard deviation of 299 genes in the wildtype and mutants. (4) The coefficients of the Effective Empirical Network. (5) Comparison between the predicted and experimental gene expression levels for the double knockout of fnr and arcA. The experimental data are normalized by the corresponding mean value of the wildtype replicates (item (2)).
Found at: doi:10.1371/journal.pone.0013080.s001 (0.20 MB XLS)

## Acknowledgments

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

## Author Contributions

Conceived and designed the experiments: GHG PHG. Performed the experiments: GHG LS. Analyzed the data: LS AT. Contributed reagents/materials/analysis tools: PHG. Wrote the paper: GHG AT.

## References


1. Wagner A (2005) Robustness and Evolvability in Living Systems. Princeton University Press, Princeton, New Jersey.
2. Bernard C (1927) An Introduction to the Study of Experimental Medicine. Macmillan, New York: Macmillan, New York.
3. Waddington C (1959) Canalization of Development and Genetic Assimilation of Acquired Characters. Nature 183: 1654–1655.
4. Waddington C (1960) Experiments on Canalizing Selection. Genetical Research 1: 140–150.
5. Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R (2009) Gene regulatory network inference: Data integration in dynamic models-A review. Biosystems 96: 86–103.
6. Barabasi A, Oltvai Z (2004) Network biology: Understanding the Cell's Functional Organization. Nature Reviews Genetics 5: 101–U15.
7. Ravasz E, Somera A, Mongru D, Oltvai Z, Barabasi A (2002) Hierarchical Organization of Modularity in Metabolic Networks. Science 297: 1551–1555.
8. Oltvai Z, Barabasi A (2002) Life's Complexity Pyramid. Science 298: 763–764.
9. Kitano H (2007) Towards a Theory of Biological Robustness. Molecular Systems Biology 3: 1–7.
10. Hauser K, Abdollahi A, Huber PE (2009) Inverse system perturbations as a new methodology for identifying transcriptomic signaling participants in balanced biological processes. Cell Cycle 8: 2718–2722.
11. Zimmermann GR, Lehar J, Keith CT (2007) Multi-target therapeutics: when the whole is greater than the sum of the parts. Drug Discovery Today 12: 34–42.
12. Frantz S (2005) Playing Dirty. Nature 437: 942–943.
13. Yang K, Bai H, Ouyang Q, Lai L, Tang C (2008) Finding multiple target optimal intervention in disease-related molecular network. Molecular Systems Biology 4: 228.
14. Gardner TS, di Bernardo D, Lorenz D, Collins JJ (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. Science 301: 102–105.
15. Bornholdt S (2005) Less is More in Modeling Large Genetic Networks. Science 310: 449+.
16. Covert MW, Palsson BO (2003) Constraints-based models: Regulation of gene expression reduces the steady-state solution space. Journal of Theoretical Biology 221: 309–325.
17. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, et al. (2002) Network motifs: Simple building blocks of complex networks. Science 298: 824–827.
18. Shen-Orr S, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of Escherichia coli. Nature Genetics 31: 64–68.
19. Bornholdt S (2008) Boolean network models of cellular regulation: prospects and limitations. Journal of the Royal Society Interface 5: S85–S94.
20. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, et al. (2007) Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. PLOS Biology 5: 54–66.
21. Someren EPv, Wessels LFA, Reinders MJT (2000) Linear modeling of genetic networks from experimental data. In: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology. AAAI Press. pp 355–366.
22. di Bernardo D, Thompson M, Gardner T, Chobot S, Eastwood E, et al. (2005) Chemogenomic profiling on a genomewide scale using reverse-engineered gene networks. Nature Biotechnology 23: 377–383.
23. von Luxburg U (2007) A tutorial on spectral clustering. Statistics and Computing 17: 395–416.
24. Newman MEJ (2004) Fast algorithm for detecting community structure in networks. Physical Review E 69: 066133.
25. Sturn A, Quackenbush J, Trajanoski Z (2002) Genesis: Cluster Analysis of Microarray Data. Bioinformatics 18: 207–208.
26. Babu M, Luscombe N, Aravind L, Gerstein M, Teichmann S (2004) Structure and Evolution of Transcriptional Regulatory Networks. Current Opinion In Structural Biology 14: 283–291.
27. Gill G (2001) Regulation of the initiayion of eukaryotic transcription. Essays in Biochemistry 37: 33–43.
28. Ambros V (2004) The functions of animal microRNAs. Nature 431: 350–355.
29. Bartel D (2004) MicroRNAs: Genomics, biogenesis, mechanism, and function. Cell 116: 281–297.
30. Herrgard MJ, Covert MW, Palsson BO (2003) Reconciling gene expression data with known genome-scale regulatory network structures. Genome Research 13: 2423–2434.
31. Salmon KA, Hung S, Steffen NR, Krupp R, Baldi P, et al. (2005) Global gene expression profiling in escherichia coli k12 - effects of oxygen availability and arca. Journal of Biological Chemistry 280: 15084–15096.
32. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO (2004) Integrating high-throughput and computational data elucidates bacterial networks. Nature 429: 92–96.
33. Tsukakoshi M, Kurata S, Nomiya Y, Ikawa Y, Kasuya T (1984) A Novel Method of DNA Transfection by Laser Microbeam Cell Surgery. Applied Physics B-Photophysics and Laser Chemistry 35: 135–140.
34. Bertram J (2006) MATra - Magnet Assisted Transfection: Combining Nanotechnology and Magnetic Forces to Improve Intracellular Delivery of Nucleic Acids. Current Pharmaceutical Biotechnology 7: 277–285.

35. Feinberg M (1995) The existence and uniqueness of steady states for a class of chemical reaction networks. Arch Rational Mech Anal 132: 311–370.
36. Hirsch MW, Smith H (2005) Monotone dynamical systems. In: Handbook of differential equations: ordinary differential equations Vol. II, Elsevier B. V., Amsterdam. pp 239–357.
37. Angeli D, Sontag ED (2008) Translation-invariant monotone systems, and a global convergence result for enzymatic futile cycles. Nonlinear Anal Real World Appl 9: 128–140.
38. Carlsson G, Zomorodian A (2009) The Theory of Multidimensional Persistence. Discrete & Computational Geometry 42: 71–93.
39. Carlsson G (2009) Topology and Data. Bulletin of the American Mathematical Society 46: 255–308.
40. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, et al. (2010) Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 17: 98–110.
41. Eisen M, Spellman P, Brown P, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States Of America 95: 14863–14868.
42. Spellman P, Sherlock G, Zhang M, Iyer V, Anders K, et al. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Molecular Biology of the Cell 9: 3273–3297.