Statistical method for revealing form-function relations in biological networks

Andrew Mugler[a,1,2], Boris Grinshpun[b], Riley Franks[c], and Chris H. Wiggins[b,d]

[a]Department of Physics; [b]Department of Applied Physics and Applied Mathematics; [d]Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10027; and [c]Department of Applied and Computational Mathematics, California Institute of Technology, Pasadena, CA 91125

Edited* by Leslie Greengard, New York University, New York, NY, and approved November 12, 2010 (received for review June 25, 2010)

Over the past decade, a number of researchers in systems biology have sought to relate the function of biological systems to their network-level descriptions—lists of the most important players and the pairwise interactions between them. Both for large networks (in which statistical analysis is often framed in terms of the abundance of repeated small subgraphs) and for small networks which can be analyzed in greater detail (or even synthesized in vivo and subjected to experiment), revealing the relationship between the topology of small subgraphs and their biological function has been a central goal. We here seek to pose this revelation as a statistical task, illustrated using a particular setup which has been constructed experimentally and for which parameterized models of transcriptional regulation have been studied extensively. The question "how does function follow form" is here mathematized by identifying which topological attributes correlate with the diverse possible information-processing tasks which a transcriptional regulatory network can realize. The resulting method reveals one form-function relationship which had earlier been predicted based on analytic results, and reveals a second for which we can provide an analytic interpretation. Resulting source code is distributed via http://formfunction.sourceforge.net.

form and function | information theory | dynamical systems | systems biology

The observation that form constrains function in biological systems has historical roots far older than systems biology. Century-old examples include those made in D'Arcy Thompson's "On Growth and Form" (1) and the observation that the quick responses necessary for reflex actions such as heat- and pain-avoidance could be manifest only by a dedicated input-output relay circuit from fingers to brain and back (2). Advances in synthetic biology, often requiring design of systems for which only topology can be specified without control over precise parameter values, has motivated a reintroduction of such topological thinking in biological systems (3–5). A second source of such inquiry is high-throughput systems biology, in which technological advances provide topologies of large biological networks without precise knowledge of their interactions, dynamics, or possible naturally occuring inputs (6–8). Such limitations thwart our desire to learn form-function relations from data or to derive them from plausible first-principles modeling. Our goal here is to illustrate how reframing the question as one of computation and statistical analysis allows a clear, quantitative, interpretable approach.

Setup

Mathematical progress requires clear definitions of terms, including, here, "form," "function," and "follow." To define the first two we must choose a specific experimental setup; we here choose one which has been experimentally realized repeatedly: that of a small, synthetic transcriptional regulatory network with "inducible promoters," meaning that the efficacy of the transcription factors may be diminished by introducing small interfering molecules (5, 9–11) (Fig. 1*A*). A common "output" responding to the "input" presence of such small molecules is the abundance of inducible green fluorescent protein (GFP), which provides an optical readout of one of the regulated genes. The "form," then, will be defined by the topology of such a small regulatory network, distinguishing between up- and down-regulatory edges in the network.† "Function" will be defined by the realizable input-output relations of a device with two binary inputs (corresponding to presence or absence of two species of interfering small molecule) and one real-valued output (the transcriptional level of the output gene).

Among other published experiments which correspond to this setup is that of Guet et al. (9). We remind the reader of two par-

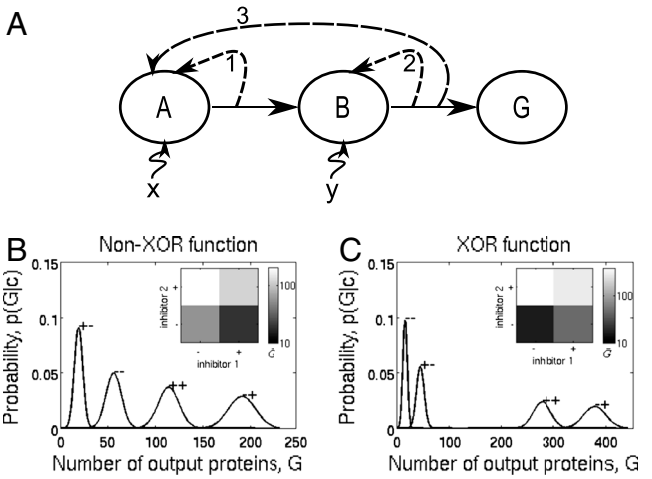

Fig. 1. Network set and input-output functions. (*A*) Transcription factor *A* regulates the expression of transcription factor *B*, which regulates the expression of fluorescent protein *G*. The efficacies of *A* and *B* are reduced by the presence of chemical inhibitors, labeled by scaling factors *x* and *y* (Eq. 1), respectively. We distinguish between up- and down-regulation and consider all ways in which regulatory edges 1, 2, and 3 may appear, for a total of 160 networks. (*B* and *C*) Examples of non-XOR (*B*) and XOR (*C*) functions (see *Results: Nonparametric Analysis*), as defined by the ranking of conditional probability distributions $p(G|c)$, where $c \in \{--, -+, +-, ++\}$ describes whether each inhibitor is present (+) or absent (–). Insets show mean protein number $\bar{G}$ in each of the four states. The functions in *B* and *C* are both realized by the particular network in *A* in which edge 3 is absent and all remaining edges are down-regulating.

Author contributions: A.M., B.G., R.F., and C.H.W. designed research; A.M., B.G., and R.F. performed research; A.M. contributed new reagents/analytic tools; A.M. and B.G. analyzed data; and A.M. and C.H.W. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

†We use → to indicate up-regulation, ⊣ to indicate down-regulation, and → to indicate regulation whose sign is not specified; additionally we use ⇝ to indicate inhibition by a small molecule.

[1]To whom correspondence should be addressed. E-mail: mugler@amolf.nl.

[2]Present address: Foundation for Fundamental Research on Matter (FOM), Institute for Atomic and Molecular Physics (AMOLF), Science Park 104, 1098 XG, Amsterdam, The Netherlands.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1008898108/-/DCSupplemental.

ticularly noteworthy observations of Guet et al.: (*i*) that many of their experimentally-realized small networks were "broken" in the sense that the output remained constant over the different possible inputs; and (*ii*) that often the same topology can realize different input-output relations (or even be broken, i.e., can realize both a particular function as well as a lack of function entirely). Within a mathematical model, such behavior follows straightforwardly from considering the behavior of a given dynamical system at different points in the space of quantitative parameters (12). Revealing such degeneracy of functions by exploring the parameter space given a topology (and given an algebraic expression modeling the regulatory interactions among the genes) may be recast as one of optimizing—locally in parameter space—the mutual information between input and output over this space (13–15). Mutual information (MI) as a cost function is advantageous both biologically [in that many natural systems including transcriptional regulatory networks are known to operate near their information-optimal constraints (16–18)] and mathematically (in that by optimizing MI we can identify parameter settings which are functional without demanding in advance the particular input-output functions we seek). That is, we optimize for functionality rather than for a specific predetermined function.

MI between input and output is defined as a functional of the joint distribution $p(c,G)$—the probability of a (here, categorically-valued) setting of the chemistry and the (here, real-valued) concentration of the output gene (see *Method*). The noise in the relationship between input and output in biological networks has many underlying sources; however, one source is intrinsic: the finite copy number of the constituent proteins introduces a "shot noise," much discussed in the systems biology literature (19–23). Particular additional sources of noise may thwart information processing as well in specific systems; hoping to remain as general as possible, we will here consider only intrinsic noise. In this respect, we aim to describe the functional response(s) of single cells, as opposed to an averaged response of a pooled population of cells.

Having defined form and function, we must define "follow." Here we will need a measure of, for any topological feature of the networks, the extent to which the value of the feature (e.g., length of a cycle, number of down-regulations, etc.) does or does not correlate with the input-output relations the network can perform. Because we wish to correlate two categorical (rather than real-valued) features, MI is again useful, in that we will rank topological features by the information between their value and the particular input-output relations that networks with that feature are found to perform.

To summarize: "function" is given by the possible input-output relations of which a given topology is capable; these are found by numerical optimization over parameters of the MI between input and output, with the probabilistic description of the transcriptional output set by intrinsic noise; "form" is mathematized by enumerating a list of topological attributes descriptive of our small networks. The question "how does function follow form" is here recast as a ranking of topological features based on the correlation (here, given by MI) between topological feature value and the particular input-output relations realizable for a given topology.

Method

The method is illustrated in Fig. 2, which makes clear how the (general) search for form-function relations may be composed into separate tasks, the implementation of which is particular to one particular experimental setup. For example, the particular elements of the network set Ω_n follow from system-specific choices made below, but the design of the task (Fig. 2) can be applied to relating form and function more generally. For a different experimental context, one or more of the "modules" in the design of the algorithm may need to be replaced, but the design itself we anticipate will be useful to revealing form-function relations in a wide variety of contexts.

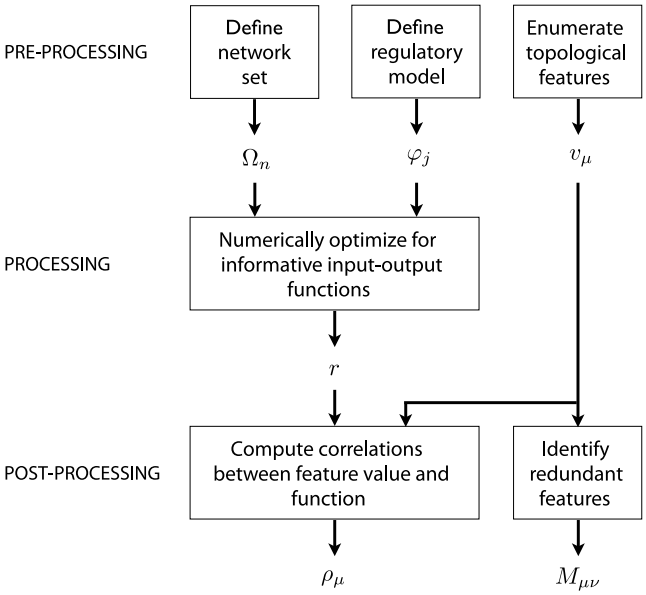

Fig. 2. Outline of the procedure used to ask "how does function follow form." Variables are defined in text.

Below we describe the method in detail as applied to our particular experimental setup; further information on networks, gene regulation, linear noise analysis, information theory, and optimization is available in the supporting information *SI Appendix*.

Networks and Features. We consider the simplest set of networks with two chemical inputs and one genetic output (Fig. 1*A*). Each of two chemical inhibitors is either present or absent, giving four possible input states. When present, the chemicals respectively inhibit two transcription factors, the second of which regulates the fluorescent output. We consider all ways in which each of the transcription factors may regulate itself and the other (with the constraint that neither is unregulated) and distinguish between up- and down-regulation, giving a total of 160 networks (the combinatorial accounting is presented in *SI Appendix*). The topology of each network $n \in \{1,2,\ldots,160\}$ constrains the model parameters ϑ to lie within the particular feasibility set Ω_n. Having defined the set of networks, we enumerate topological features and their values v_μ (here, $\mu \in \{1,2,\ldots,17\}$; see Table 1).

Table 1. Topological features ranked by correlation measure ρ

	Topological feature	ρ
1.	Sign of forward edges	0.92366
2.	Number of up-regulating edges	0.18249
3.	Number of down-regulating edges	0.18098
4.	Sign of autoregulation of species *B*	0.03228
5.	Number of positive feedback cycles	0.01133
6.	Number of negative feedback cycles	0.01109
7.	Nesting structure of feedback cycles	0.00418
8.	Type of interaction of edges into *B*	0.00288
9.	Number of edges	0.00184
10.	Number of feedback cycles	0.00184
11.	Sign of $A-B$ feedback cycle	0.00173
12.	Number of additive interactions	0.00141
13.	Type of interaction of edges into *A*	0.00085
14.	Number of nested feedback cycles	0.00081
15.	Sign of autoregulation of species *A*	0.00058
16.	Number of multiplicative interactions	0.00048
17.	Sign of edge $B \to A$	0.00021

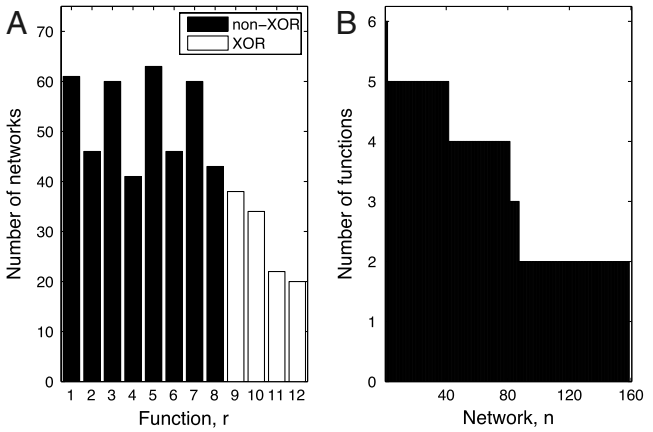

Fig. 3. Nonparametric analysis of network functionality. (*A*) Histogram showing how many networks can perform each input-output function. Functions are numbered along the horizontal axis as in Fig. 5. (*B*) Histogram showing how many input-output functions may be realized by each network. The order of networks along the horizontal axis is determined by ranking according to number of functions realized.

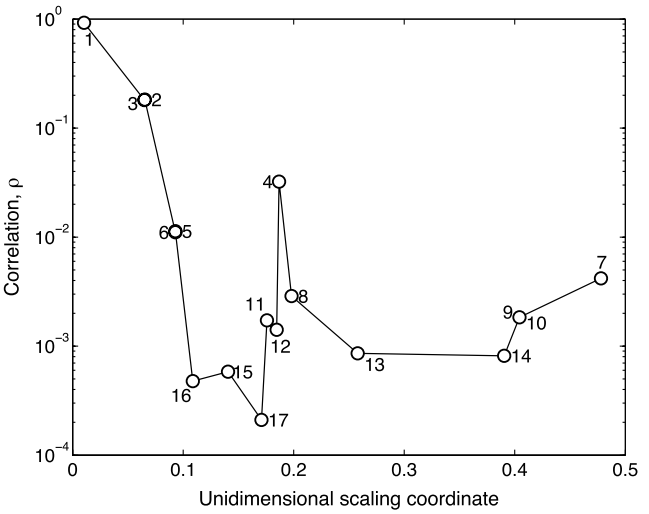

Fig. 4. Identifying feature redundancy. Correlation measure ρ is plotted against a unidimensional scaling coordinate which groups similar features together (i.e., the components of the eigenvector corresponding to the largest-magnitude eigenvalue of the feature adjacency matrix $M_{\mu\nu}$). Features are numbered by rank (Table 1).

Regulatory Model. The mean protein numbers of the two transcription factors $\bar{A}$ and $\bar{B}$ and the fluorescent output $\bar{G}$ are described by the deterministic dynamics

$$\frac{1}{R_A}\frac{d\bar{A}}{dt} = \varphi_A(a,b) - \bar{A}, \qquad \frac{1}{R_B}\frac{d\bar{B}}{dt} = \varphi_B(a,b) - \bar{B},$$
$$\frac{1}{R_G}\frac{d\bar{G}}{dt} = \varphi_G(b) - \bar{G}, \qquad [1]$$

where $a = \{\bar{A}/x,\bar{A}\}$ when the first inhibitor is {present, absent}, $b = \{\bar{B}/y,\bar{B}\}$ when the second inhibitor is {present, absent}, and the R_j are degradation rates ($j \in \{A,B,G\}$). The parameters $x > 1$ and $y > 1$ model the effect of the interfering small molecules by reducing the effective concentrations of the proteins. There are thus a total of four chemical input states denoted $c \in \{--, -+, +-, ++\}$, each state describing whether each of the two inhibitors is present (+) or absent (−). The terms φ_j describe the transcriptional regulation of each species by its parent(s) and are formulated under a statistical mechanical model (24–26). The statistical mechanical approach to modeling transcription is principled, compact, and in the case of combinatorial regulation (24) captures the diversity of multidimensional responses observed in experimental systems (27–29). Full algebraic forms of the φ_j are dependent on topology, including, in the case of combinatorial regulation, whether the transcription factor interaction is additive or multiplicative (see *SI Appendix*).

The stochastic description of each network is set by intrinsic noise. We obtain probability distributions over protein numbers using the linear noise approximation (LNA) (20, 30, 31), because, in contrast to simulation techniques (19), the LNA does not rely on sampling and is thus much more computationally efficient (making many-parameter optimization feasible). Under the LNA the steady-state distribution over each species' protein number is a Gaussian expansion around the deterministic mean given by the steady state of Eq. **1**. The covariance matrix Ξ under the LNA is determined from model parameters by (numerically) solving the Lyapunov equation $J\Xi + \Xi J^T + D = 0$, where J is the Jacobian of the system in Eq. **1** and $D = \mathrm{diag}\{R_A(\varphi_A + \bar{A}), R_B(\varphi_B + \bar{B}), R_G(\varphi_G + \bar{G})\}$ is an effective diffusion matrix. Of particular importance are the distributions $p(G|c)$, the stochastic response of the output species G given that the system is in each of the four input states c. The input-output MI may be computed directly from this quantity, $I[p(c,G)] = \sum_c \int dG\, p(G|c)p(c)\log_2[p(G|c)/\sum_{c'} p(G|c')p(c')]$, with the provision that the input states are equally likely, $p(c) = 1/4$.

Input-Output Functions. The possible input-output responses of each network are found by locally optimizing the input-output MI in parameter space. The optimization is done numerically using MATLAB's fminsearch() and initialized by sampling uniform-randomly in the logs of the parameters (specific bounds from which initial parameters are sampled are given in *SI Appendix*: *Table S1*). The optimization is performed at constrained average protein number $N \equiv (\bar{A} + \bar{B} + \bar{G})/3$ and average time scale separation $T \equiv [(R_A + R_B)/2]/R_G$ by maximizing the quantity $L \equiv I - \eta N - \kappa T$ for values of the Lagrange multipliers η and κ which give biologically plausible values for N and T for single cells (R_G is fixed).

Optimization of MI has the effect of increasing the separation among the distributions $p(G|c)$ (see Fig. 1 *B* and *C*). To reflect the fact that many observed regulatory networks are known to operate near their information-optimal limits (16–18), we use in the statistical analysis only those optimal solutions whose MI lies above a cutoff value. Choosing a cutoff larger than 1.5 bit (which corresponds to two fully separated distributions and two fully overlapping distributions) ensures that the means of the distributions are fully resolved, and thus allows one to define the function performed by the network as the ranking r of the means of the distributions $p(G|c)$ along the G axis. For the results in this study we use a cutoff of 1.55 bit. The method can be easily extended to include less informative locally optimal functions, e.g., binary logic gates such as an AND function, by using a lower MI cutoff and generalizing the definition of function as ranking (14). Fig. 1 *B* and *C* shows examples of two different functions performed by the same network that are local optima in MI at different points in parameter space; they correspond to $r = 1$ and $r = 9$ on the horizontal axis of Fig. 5, respectively.

To correct for repeated sampling of the same local optimum at close but numerically distinct points in the real-valued parameter space, nearest-neighbor optima performing the same function are merged. This choice enforces that the distribution of optimal parameters is sampled uniformly. The robustness of subsequent results to this choice is tested numerically (see *Results*).

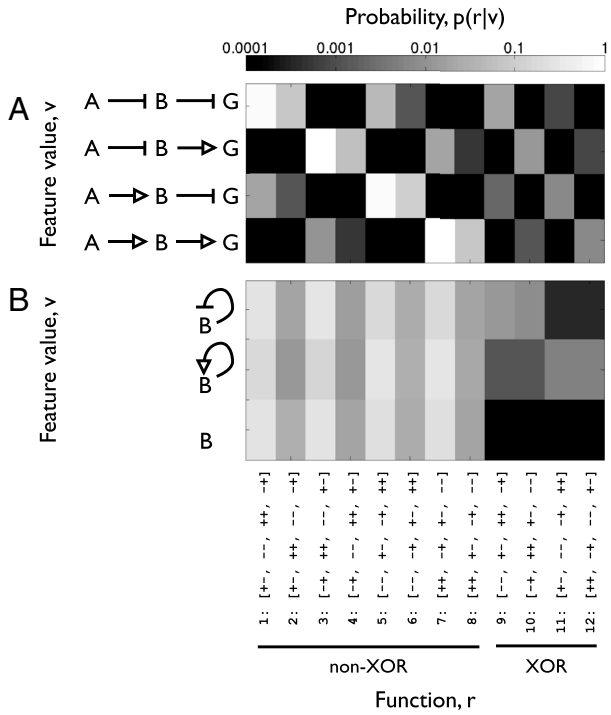

Fig. 5. Conditional distributions showing the probability of a particular input-output function given the value of a topological feature, for two features: (*A*) the signs (up- or down-regulating) of the forward regulations, and (*B*) the sign (up-regulating, down-regulating, or absent) of the autoregulation of species *B*.

Correlating Feature Value and Function. For each topological feature μ, the correlation between feature value v_μ and function r is computed from the joint probability distribution $p(v_\mu,r)$. This distribution is defined by the optimization data and the factorization

$$p(v_\mu,r) = \sum_{\vartheta,n} p(v_\mu,r,\vartheta,n) = \sum_{\vartheta,n} p(v_\mu,r|\vartheta,n)p(\vartheta,n)$$
$$= \sum_{\vartheta,n} p(v_\mu|n)p(r|\vartheta,n)p(\vartheta|n)p(n), \quad [2]$$

where ϑ in Ω_n runs over all points in parameter space at which an optimum is found. Here $p(v_\mu|n)$ is $\{0,1\}$-valued, set by whether network n has value v for feature μ; and $p(r|n,\vartheta)$ is $\{0,1\}$-valued, set by whether network n performs function r at point ϑ, according to the optimization data. The distributions $p(\vartheta|n)$ and $p(n)$ are assumed to be "flat," i.e., $p(\vartheta|n) = 1/|\vartheta|_n$ and $p(n) = 1/|n|$, where $|\vartheta|_n$ is the number of distinct local optima in parameter space for network n, and $|n|$ is the number of networks; the robustness of subsequent results to weakening either of these assumptions is tested numerically (see *Results*).

The correlation between feature value and function is computed as their MI, normalized by the entropy of $p(v_\mu)$, to yield a statistic

$$\rho_\mu \equiv \frac{I[p(v_\mu,r)]}{H[p(v_\mu)]} \quad [3]$$

that ranges from 0 (when the function provides no information about the feature value) to 1 (when only one feature value is consistent with each realizable function).

Results

Nonparametric Analysis. We first present an analysis which requires no assumptions about what is a "flat" distribution in parameter space, i.e., we simply enumerate how many networks can perform each input-output function, and how many input-output functions are performed by each network (Fig. 3). This analysis recovers an intuitive result (Fig. 3*A*): that "XOR" functions, in which the sign of the influence of one input depends on the value of the other, are more difficult to realize (i.e., they are observed in fewer networks). The analysis also reveals that each network can perform at least two functions (Fig. 3*B*). These functions are the two consistent with the signs of the forward regulatory edges $A \to B$ and $B \to G$, as described in detail in *Forward Regulation*. Because the topology $A \to B \to G$ is that obtained in the parametric limit when the feedback edges are of negligible strength, it is clear that these functions must be realizable; Fig. 3*B* shows further that these functions are sufficiently informative to be observed as information-optima.

Topological Features and Robustness. Table 1 ranks the topological features by ρ, which measures how uniquely form determines function. Recall that in computing ρ from $p(v_\mu,r)$ we assume that the distributions $p(n)$ and $p(\vartheta|n)$ are both uniform (Eq. **2**); we find that the ranking in Table 1 is robust to deviations of both distributions from uniformity, as demonstrated by the following numerical experiments.

The uniformity of $p(n)$ is perturbed by artificially setting $p(n) \propto (u_n)^\epsilon$, where u_n is a vector of random numbers and ϵ tunes the entropy of the distribution, i.e., $\epsilon = 0$ recovers the maximum-entropy (uniform) distribution, while $\epsilon \to \infty$ produces the zero-entropy distribution $p(n) = 1 \Leftrightarrow n = \arg\max(u_n)$. We find that the ranking of the top four features is preserved under ∼15% perturbations in the entropy, and that the ranking of the top three features is preserved under ∼30% perturbations (see *SI Appendix*: *Fig. S2A*). This result demonstrates that the feature ranking is considerably robust to perturbations in the uniformity of $p(n)$.

The uniformity of $p(\vartheta|n)$ is perturbed similarly, and we find that the ranking of the top seven features is preserved under ∼40% perturbations in the entropy of $p(\vartheta)$ (see *SI Appendix*: *Fig. S2B*). In this case we also have an independent entropy scale, given by the fact that we may decompose $p(\vartheta|n)$ as $p(\vartheta|n) = \sum_{\vartheta_0} p(\vartheta|\vartheta_0,n)p(\vartheta_0|n)$, where ϑ_0 is the parameter setting that initializes an optimization and $p(\vartheta|\vartheta_0,n)$ is determined by the optimization itself. If we assume uniformity of $p(\vartheta_0|n)$, instead of $p(\vartheta|n)$, then $p(\vartheta|n)$ is computable from the numbers of times the optimization converges repeatedly on each local optimum ϑ. The entropy in this case is 13% different from that of the uniform distribution, and the ranking of ρ is almost entirely unchanged (*SI Appendix*: *Fig. S2B*). This observation demonstrates that the results are not sensitive to whether one takes the distribution of initial parameters or the distribution of optimal parameters to be uniform.

Nonredundant Features. Many topological features are not independent; for example, the feature "number of up-regulating edges" is highly correlated with "number of down-regulating edges." To interpret which features are associated with which sets of realizable functions, it is useful to group nearly identical features together and use only the feature which is most informative about function (highest in ρ) as the exemplar among each group. To quantify redundancy among features, we compute the MI between each pair of features and normalize by the minimum entropy to produce a weighted adjacency matrix $M_{\mu\nu} = I[p(v_\mu,v_\nu)]/\min\{H[p(v_\mu)],H[p(v_\nu)]\}$, which we then use as the basis for unidimensional scaling (32) (see *SI Appendix*).

Fig. 4 plots features ρ values against the unidimensional scaling coordinate, revealing two distinct groups of highly informative features. The first, which includes the features ranked 1, 2, 3, 5, and 6, is dominated by feature 1: the signs (up- or down-regulating) of the forward regulatory edges $A \to B$ and $B \to G$. The second, which includes the features ranked 4, 8, 11, and 12, is

dominated by feature 4: the sign (up-regulating, down-regulating, or absent) of the autoregulation of species B. The high information content of each of these two features is revealed visually by inspection of the conditional distribution $p(r|v_\mu) = p(v_\mu,r)/p(v_\mu)$ (Fig. 5), as described in detail in the next sections. The functional importance of both of these features is supported by analytic results; for the first feature these analytic predictions were made in earlier work (14) and are recalled here, while for the second feature we here derive the supporting analytic results.

Forward Regulation. The topological feature that is most informative of network function is feature 1: the signs of the forward regulatory edges $A \to B$ and $B \to G$. Inspection of the conditional distribution $p(r|v)$ in Fig. 5*A* reveals a rich, highly organized (and thus highly informative) structure which we here interpret.

The most prominent aspect of the probability matrix in Fig. 5*A* is the high-probability double-diagonal spanning the eight non-XOR functions (i.e., functions 1 and 2 are most often performed by networks with the first feature value, functions 3 and 4 the second feature value, functions 5 and 6 the third feature value, and functions 7 and 8 the fourth feature value). These are the functions one would expect by looking at the forward edges alone, i.e., in the absence of feedback. For example, in networks with the last feature value, $A \to B \to G$, inhibition of A and of B will both reduce the expression of G, such that the state in which both small molecules are present ($++$) produces the lowest-ranked output, and conversely, the state in which both small molecules are absent ($--$) produces the highest-ranked output; functions 7 and 8 are the two that satisfy these criteria. In previous work (14) we termed these functions "direct," and we showed analytically that networks are limited to direct functions even when feedback is added, so long as each species is regulated by at most one other species. This fact is validated here numerically: a plot of $p(r|v)$ for only those networks in our set in which each species is regulated by one other species shows nonzero entries only for the direct functions (*SI Appendix*: *Fig. S4*).

Among all networks, including those with combinatorial feedback (i.e., two edges impinging on one node), we see that direct functions still dominate, indicated by the bright double-diagonal in Fig. 5*A*. Networks with combinatorial feedback perform other functions as well, but more rarely; examples include those functions in Fig. 5*A* above and below the double-diagonal and XOR functions 9–12. The performance of these additional functions remains well organized by feature value, which makes the signs of the forward edges a highly informative feature.

Autoregulation of *B*. Other than feature 1 (and the features highly correlated with feature 1), the most informative feature is feature 4: the autoregulation of species B. Inspection of the conditional distribution $p(r|v)$ for this feature (Fig. 5*B*) reveals that the information content lies in the ability to perform XOR functions. Specifically, networks in which B is autoregulated are observed to perform XOR functions, while networks in which B is not autoregulated are not observed to perform XOR functions. Indeed, autoregulation has been observed to enhance the functional response to multiple inputs in a related study in the context of Boolean logic gates (33).

XOR functions are those in which the sign of the influence of one input depends on the value of the other; mathematically they satisfy one or both of two properties:

$$\text{XOR property I: } \operatorname{sign}(d\bar{G}/dx) \text{ depends on } y,$$

$$\text{XOR property II: } \operatorname{sign}(d\bar{G}/dy) \text{ depends on } x.$$

The four observed XOR functions satisfy property I (e.g., Fig. 1*C* inset); no functions satisfying property II are observed (Fig. 5*B*). Analytic support for these facts is obtained by calculating $d\bar{G}/dx$ and $d\bar{G}/dy$, respectively. Algebraic expressions below are given for the simplified case $R_A = R_B = R_G$.

To understand why autoregulation of B is necessary for XOR functions satisfying property I, we calculate $d\bar{G}/dx$ analytically. We obtain (see *SI Appendix*)

$$\frac{d\bar{G}}{dx} = \frac{1}{-\Delta}\frac{\partial a}{\partial x}\frac{\partial \varphi_B}{\partial a}\frac{\partial \varphi_G}{\partial \bar{B}}, \qquad [4]$$

where $\Delta = (\partial\varphi_A/\partial\bar{B})(\partial\varphi_B/\partial\bar{A}) - [(\partial\varphi_A/\partial\bar{A}) - 1][(\partial\varphi_B/\partial\bar{B}) - 1]$ is the determinant of the Jacobian of the dynamical system in Eq. **1** and is always negative for stable fixed points. Eq. **4** has an intuitive form when considering the direct path from x to G (Fig. 1*A*): the term $\partial a/\partial x = -\bar{A}/x^2 < 0$ corresponds to the inhibitory signal $x \rightsquigarrow A$, the term $\partial\varphi_B/\partial a$ corresponds to the regulatory edge $A \to B$, and the term $\partial\varphi_G/\partial\bar{B}$ corresponds to the regulatory edge $B \to G$. Because G has only one regulatory input, $\varphi_G(b)$ is monotonic, making $\partial\varphi_G/\partial\bar{B} = (d\varphi_G/db)(\partial b/\partial\bar{B}) = (d\varphi_G/db)/y$ of unique sign. The same is true for $\partial\varphi_B/\partial a$ when B has only one regulatory input (i.e., when B is not autoregulated). However when B has more than one regulatory input (i.e., when B is autoregulated), the sign of $\partial\varphi_B/\partial a$ can depend on y, allowing XOR property I. Specifically, under our regulatory model, when B is autoregulated, $\partial\varphi_B/\partial a$ is the product of a positive term and a term quadratic in $b = \bar{B}/y$ that has positive roots for some parameter settings (see *SI Appendix*). This analysis suggests inspection of the parameters themselves obtained via optimization; doing so, we observe that the vast majority of observed XOR functions result from optimal parameter values for which there exists a positive root in the range $0 < \bar{B}/y < \sim 100$, which is the range of protein numbers in which our optimal solutions lie (Fig. 1 *B* and *C*). To summarize, nonmonotonicity in the regulation of species B, which can occur only when B is autoregulated, produces the observed XOR functions.

To understand why XOR functions satisfying property II are not observed, we calculate $d\bar{G}/dy$ analytically. We obtain (see *SI Appendix*)

$$\frac{d\bar{G}}{dy} = \frac{1}{-\Delta}\left(1 - \frac{\partial\varphi_A}{\partial\bar{A}}\right)\frac{\partial b}{\partial y}\frac{d\varphi_G}{db}, \qquad [5]$$

where as before the determinant Δ is always negative. The last two terms correspond to edges along the direct path from y to G, i.e., $y \rightsquigarrow B$ and $B \to G$ respectively, and are of unique sign; the term in parentheses describes the effect of the upstream species A and feedback. In all optimal solutions the term in parentheses is observed to be positive, despite wide variations in the orders of magnitude of each of the optimal parameters across solutions. This observation is largely explained by a stability analysis: for four of the six possible topological classes of networks (those in which A is singly, not doubly, regulated; Fig. 1*A*), stability of a fixed point of Eq. **1** implies that the term in parentheses is positive; for the other two topological classes, stability implies that this term is greater than a quantity that is zero for some parameter settings and of unknown sign for others (see *SI Appendix*). In this last case it is unclear whether negative values of this term are analytically forbidden or simply exceedingly unlikely given the regulatory model and the space of optimal solutions. Empirically this term is always positive, and type-II XOR functions are not observed.

The necessity of both forward regulation and autoregulation of species B for the performance of XOR functions highlights the importance of combinatorial regulation in functional versatility. As previously mentioned, networks without combinatorial regulation are limited to a particular class of functions which does not include XOR functions (14). Moreover, in the original experi-

ment of Guet et al. (9), each species was singly regulated, and accordingly no XOR functions were observed.

Discussion

Both in order to assign functional significance to observed small network topologies in nature, and to design synthetic networks which will execute a desired function or set of functions, it is useful to develop a systematic approach for revealing the extent to which the form of a small network guides or constrains its functions. Resorting to hypothesized functions may be appealing in terms of interpretability, but this strategy risks overemphasizing those functions which one is looking for, or overlooking an unexpected function entirely.

The statistical analysis, along with the analytic results presented above, illustrate how the search for form-function relations can be posed as an algorithmic approach leading to interpretable mechanisms. While we have illustrated the analysis for a particular, experimentally-realized setup, the approach itself, subdivided into a set of distinct modules in Fig. 2, we anticipate will be applicable to a wide variety of biophysical contexts. Similarly, we have chosen a framework from which much can be discovered by analytic study of the deterministic dynamical system; other experimental setups may require vastly different analytic explanations, but the idea of using statistical methods to highlight the features of paramount importance should be implementable as illustrated. We look forward to exploring the extent to which form does or does not follow function—and how—in related biophysical and biochemical models of small information-processing networks in biology.

ACKNOWLEDGMENTS. The authors thank Nicolas Buchler and William Bialek for useful conversations. A.M. was supported by National Science Foundation (NSF) Grant DGE-0742450; C.H.W. was supported by National Institutes of Health (NIH) Grants 1U54CA121852-01A1 and 5PN2EY016586-03.

1. Thompson D (1917) *On growth and form* (Cambridge University Press, Cambridge, United Kingdom).
2. Sherrington SC (1906) *The integrative action of the nervous system* (Yale University Press, New Haven, CT).
3. Stricker J, et al. (2008) A fast, robust and tunable synthetic gene oscillator. *Nature* 456:516–519.
4. Gardner TS, Cantor CR, Collins JJ (2000) Construction of a genetic toggle switch in *Escherichia coli*. *Nature* 403:339–342.
5. Elowitz MB, Leibler S (2000) A synthetic oscillatory network of transcriptional regulators. *Nature* 403:335–338.
6. Cusick ME, Klitgord N, Vidal M, Hill DE (2005) Interactome: gateway into systems biology. *Hum Mol Genet* 2:R171–181.
7. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270:467–470.
8. Giot L, et al. (2003) A protein interaction map of Drosophila melanogaster. *Science* 302:1727–1736.
9. Guet CC, Elowitz MB, Hsing W, Leibler S (2002) Combinatorial synthesis of genetic networks. *Science* 296:1466–1470.
10. Basu S, Mehreja R, Thiberge S, Chen MT, Weiss R (2004) Spatiotemporal control of gene expression with pulse-generating networks. *Proc Natl Acad Sci USA* 101:6355–6360.
11. Mangan S, Zaslaver A, Alon U (2003) The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *J Mol Biol* 334:197–204.
12. Strogatz SH (2000) *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering* (Westview Press, Cambridge, MA).
13. Ziv E, Nemenman I, Wiggins CH (2007) Optimal signal processing in small stochastic biochemical networks. *PLoS ONE* 2:e1077.
14. Mugler A, Ziv E, Nemenman I, Wiggins CH (2008) Serially regulated biological networks fully realise a constrained set of functions. *IET Syst Biol* 2:313–322.
15. Mugler A, Ziv E, Nemenman I, Wiggins CH (2009) Quantifying evolvability in small biological networks. *IET Syst Biol* 3:379–387.
16. Tkačik G, Callan CG, Bialek W (2008) Information flow and optimization in transcriptional regulation. *Proc Natl Acad Sci USA* 105:12265–12270.
17. Mehta P, Goyal S, Long T, Bassler BL, Wingreen NS (2009) Information processing and signal integration in bacterial quorum sensing. *Mol Syst Biol* 5:325.
18. Walczak AM, Tkačik G, Bialek W (2010) Optimizing information flow in small genetic networks. II. Feed-forward interactions. *Phys Rev E* 81:41905.
19. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* 81:2340–2361.
20. Paulsson J (2004) Summing up the noise in gene networks. *Nature* 427:415–418.
21. Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. *Science* 297:1183–1186.
22. Walczak AM, Mugler A, Wiggins CH (2009) A stochastic spectral analysis of transcriptional regulatory cascades. *Proc Natl Acad Sci USA* 106:6529–6534.
23. Mugler A, Walczak AM, Wiggins CH (2009) Spectral solutions to stochastic models of gene expression with bursts and regulation. *Phys Rev E* 80:041921.
24. Buchler NE, Gerland U, Hwa T (2003) On schemes of combinatorial transcription logic. *Proc Natl Acad Sci USA* 100:5136–5141.
25. Bintu L, et al. (2005) Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev* 15:116–124.
26. Bintu L, et al. (2005) Transcriptional regulation by the numbers: applications. *Curr Opin Genet Dev* 15:125–135.
27. Kaplan S, Bren A, Zaslaver A, Dekel E, Alon U (2008) Diverse two-dimensional input functions control bacterial sugar genes. *Mol Cell* 29:786–792.
28. Cox RS, Surette MG, Elowitz MB (2007) Programming gene expression with combinatorial promoters. *Mol Syst Biol* 3:145.
29. Mayo AE, Setty Y, Shavit S, Zaslaver A, Alon U (2006) Plasticity of the cis-regulatory input function of a gene. *PloS Biol* 4:e45.
30. Elf J, Ehrenberg M (2003) Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome Res* 13:2475–2484.
31. van Kampen NG (1992) *Stochastic processes in physics and chemistry* (North-Holland, Amsterdam).
32. Borg I, Groenen PJF (2005) *Modern multidimensional scaling: theory and applications* (Springer-Verlag, New York, NY), 2nd ed.
33. Hermsen R, Ursem B, ten Wolde PR (2010) Combinatorial gene regulation using auto-regulation. *PLoS Comput Biol* 6:e1000813.