Observability of complex systems

Yang-Yu Liu[a,b,c,d,e], Jean-Jacques Slotine[f,g,h], and Albert-László Barabási[a,b,c,d,e,i,1]

[a]Center for Complex Network Research and Departments of [b]Physics, [c]Computer Science, and [d]Biology, Northeastern University, Boston, MA 02115; [e]Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, MA 02115; [f]Nonlinear Systems Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139; Departments of [g]Mechanical Engineering and [h]Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139; and [i]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115

Edited by Giorgio Parisi, University of Rome, Rome, Italy, and approved December 26, 2012 (received for review September 6, 2012)

A quantitative description of a complex system is inherently limited by our ability to estimate the system's internal state from experimentally accessible outputs. Although the simultaneous measurement of all internal variables, like all metabolite concentrations in a cell, offers a complete description of a system's state, in practice experimental access is limited to only a subset of variables, or sensors. A system is called observable if we can reconstruct the system's complete internal state from its outputs. Here, we adopt a graphical approach derived from the dynamical laws that govern a system to determine the sensors that are necessary to reconstruct the full internal state of a complex system. We apply this approach to biochemical reaction systems, finding that the identified sensors are not only necessary but also sufficient for observability. The developed approach can also identify the optimal sensors for target or partial observability, helping us reconstruct selected state variables from appropriately chosen outputs, a prerequisite for optimal biomarker design. Given the fundamental role observability plays in complex systems, these results offer avenues to systematically explore the dynamics of a wide range of natural, technological and socioeconomic systems.

algebraic observability | biochemical reactions | control theory

The internal variables of a complex system are rarely independent of each other, as the interactions between the system's components induce systematic interdependencies between them. Hence, a well-selected subset of variables can contain sufficient information about the rest of the variables, allowing us to reconstruct the system's complete internal state, making the system observable. To address observability in quantitative terms, we focus on systems whose dynamics can be described by the generic state-space form

$$\dot{\mathbf{x}}(t)=\mathbf{f}(t,\mathbf{x}(t),\mathbf{u}(t)), \qquad \textbf{[1]}$$

where $\mathbf{x}(t)\in\mathbb{R}^N$ represents the complete internal state of the system (e.g., the concentrations of all metabolites in a cell), and the input vector $\mathbf{u}(t)\in\mathbb{R}^K$ captures the influence of the environment. Observing the system means that we monitor a set of variables $\mathbf{y}(t)\in\mathbb{R}^M$ that depend on the time t, the system's internal state $\mathbf{x}(t)$, and the external input $\mathbf{u}(t)$,

$$\mathbf{y}(t)=\mathbf{h}(t,\mathbf{x}(t),\mathbf{u}(t)). \qquad \textbf{[2]}$$

Observability requires us to establish a relationship between the outputs $\mathbf{y}(t)$, the state vector $\mathbf{x}(t)$, and the inputs $\mathbf{u}(t)$ in a manner that we can uniquely infer the system's complete initial state $\mathbf{x}(0)$. The observability criteria can be formulated algebraically for dynamical systems consisting of polynomial or rational expressions (1, 2) stating that [**1**] is observable if the Jacobian matrix $\mathcal{J}=[J_{ij}]_{NM\times N}$ has full rank,

$$\text{rank }\mathcal{J}=N, \qquad \textbf{[3]}$$

where $J_{ij}=\frac{\partial L_f^{\left\lfloor \frac{i-1}{M}\right\rfloor} h_{(i-1)\%\,M+1}}{\partial x_j}$, the Lie derivatives $L_f:=\frac{\partial}{\partial t}+\sum_{i=1}^N f_i\frac{\partial}{\partial x_i}+\sum_{j\in\mathbb{N}}\sum_{l=1}^K u_l^{(j+1)}\frac{\partial}{\partial u_l^{(j)}}$, $\lfloor x\rfloor$ is the largest integer not greater than x, and $\%$ is the modulo operation (*SI Text*, section I). For a linear time-invariant dynamic system (3, 4), $\dot{\mathbf{x}}(t)=\mathbf{A}\,\mathbf{x}(t)+\mathbf{B}\,\mathbf{u}(t)$ and $\mathbf{y}(t)=\mathbf{C}\,\mathbf{x}(t)$, $\mathcal{J}$ reduces to the observability matrix $\mathcal{O}=[\mathbf{C}^{\mathrm{T}},(\mathbf{C}\,\mathbf{A})^{\mathrm{T}},\cdots,(\mathbf{C}\,\mathbf{A}^{N-1})^{\mathrm{T}}]^{\mathrm{T}}$.

To simplify the observability analysis, we assume that we can monitor a selected subset of state variables, i.e., $\mathbf{y}(t)=(\cdots,x_i(t),\cdots)^{\mathrm{T}}$, which we call sensors. Observability of complex systems can then be posed as follows: Identify the minimum set of sensors from whose measurements we can determine all other state variables. Whereas [**3**] offers a formal answer to the observability issue in the context of small engineered systems, it has notable practical limitations for natural and complex systems. First, it can only confirm (or deny) if a specific sensor set can be used to observe a system, without telling us how to select it. Second, a brute-force search for a minimum sensor set requires us to inspect via [**3**] of about 2^N sensor combinations, a computationally prohibitive task for large complex systems. Third, the rank test of the Jacobian matrix via symbolic computation is computationally limited to small systems (5). Hence, the fundamental and the practically useful question of identifying the minimum set of sensors through which we can observe a large complex system remains unsolved.

To resolve these limitations, one can exploit the dynamic interdependence of the system's components through a graphical representation, a common approach used in structured system theory (6–10). This procedure consists of the following steps:

i) Inference diagram: We draw a directed link $x_i\rightarrow x_j$ if x_j appears in x_i's differential equation (i.e., if $\frac{\partial f_i}{\partial x_j}$ is not identically zero), implying that one can collect information on x_j by monitoring x_i as a function of time. Because the constructed network captures the information flow in inferring the state of individual variables, we call it an inference diagram (Fig. 1*C*). By flipping the direction of each edge, the procedure recovers the system digraph encountered in structured systems theory (11–13).

ii) Strongly connected component (SCC) decomposition: We decompose the inference diagram into a unique set of maximal SCCs, which are the largest subgraphs chosen such that there is a directed path from each node to every other node in the subgraph (14). The SCCs of the inference diagram of Fig. 1*C* are surrounded by dashed circles. Note that each node in a SCC contains information pertaining to all other nodes within the SCC.

iii) Sensor node selection: We call root SCCs those SCCs that have no incoming edges (shaded circles in Fig. 1*C*). We choose at least one node from each root SCC to ensure observability of the whole system. For example, the inference

Author contributions: Y.-Y.L., J.-J.S., and A.-L.B. designed research; Y.-Y.L., J.-J.S., and A.-L.B. performed research; Y.-Y.L. contributed analytic tools; Y.-Y.L. analyzed data; and Y.-Y.L. and A.-L.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: alb@neu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1215508110/-/DCSupplemental.

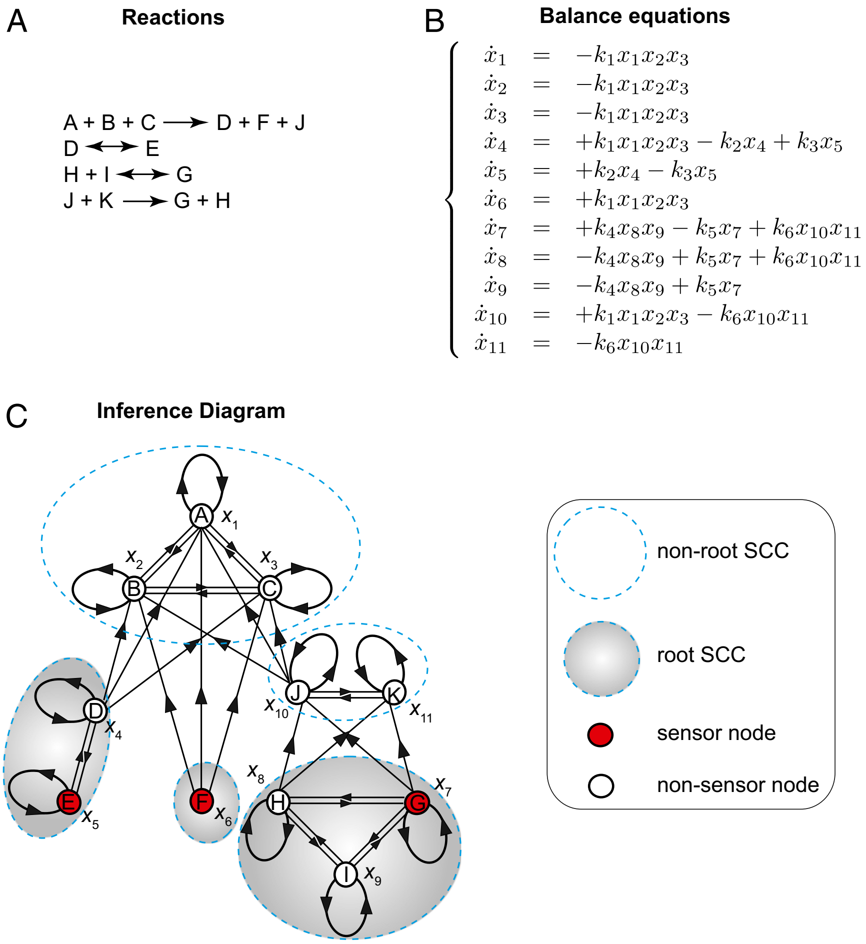
A
Reactions
A + B + C → D + F + J
D ↔ E
H + I ↔ G
J + K → G + H
B
Balance equations
$\dot{x}_1 = -k_1x_1x_2x_3$
$\dot{x}_2 = -k_1x_1x_2x_3$
$\dot{x}_3 = -k_1x_1x_2x_3$
$\dot{x}_4 = +k_1x_1x_2x_3 - k_2x_4 + k_3x_5$
$\dot{x}_5 = +k_2x_4 - k_3x_5$
$\dot{x}_6 = +k_1x_1x_2x_3$
$\dot{x}_7 = +k_4x_8x_9 - k_5x_7 + k_6x_{10}x_{11}$
$\dot{x}_8 = -k_4x_8x_9 + k_5x_7 + k_6x_{10}x_{11}$
$\dot{x}_9 = -k_4x_8x_9 + k_5x_7$
$\dot{x}_{10} = +k_1x_1x_2x_3 - k_6x_{10}x_{11}$
$\dot{x}_{11} = -k_6x_{10}x_{11}$
C
Inference Diagram
A x_1
B x_2
C x_3
D x_4
E x_5
F x_6
G x_7
H x_8
I x_9
J x_{10}
K x_{11}
non-root SCC
root SCC
sensor node
non-sensor node

Fig. 1. Graphical approach. (*A*) Chemical reaction system with 11 species (A,B,. . .,J,K) involved in four reactions. Because two reactions are reversible, we have six elementary reactions. (*B*) Balance equations of the chemical reaction system shown in *A*. Concentrations of the 11 species are denoted by x_1, x_2,. . .,x_{10}, x_{11}, respectively. Rate constants of the six elementary reactions are given by k_1, k_2,. . .,k_6, respectively. Balance equations are derived using the mass-action kinetics. (*C*) Inference diagram is constructed by drawing a directed link ($x_i \rightarrow x_j$) if x_j appears in the right-hand side of x_i's balance equation shown in *B*. SCCs, which are the largest subgraphs chosen such that there is a directed path from each node to every other node in the subgraph, are marked with dashed circle; root SCCs, which have no incoming links, are shaded in gray. A potential minimum set of sensor nodes, whose measurements allow us to reconstruct the state of all other variables (metabolite concentrations), is shown in red.

diagram of Fig. 1*C* contains three root SCCs, hence, we need at least three sensors to observe the system.

Results

The graphical approach (GA) described above reduces observability, a dynamical problem of a nonlinear system with many unknowns, to a property of the static map of the inference diagram, which is accurately mapped for an increasing number of complex systems (15, 16). This leads us to our first result: we find that monitoring the root SCCs identified by the GA are necessary for observing any nonlinear dynamic system. In other words, we prove that the number of root SCCs yields a strict lower bound of the size of the minimum sensor set (*SI Text,* section II, A). Consequently any state observer, a dynamical device that aims to estimate the system's internal state, will fail if it does not monitor these sensors.

If the dynamics [**1**] is linear, we can use the maximum matching (MM) algorithm to predict not only the necessary, but also the minimum sensor set sufficient for observability (17). Numerical simulations on model networks indicate that for linear systems the sensor set predicted by MM is noticeably larger than the necessary sensor set predicted by GA (Fig. 2*B*). The reason is that any symmetries in the state variables that leave the inputs, outputs, and all their derivatives invariant will make the system unobservable. Indeed, a dynamical system with internal symmetries can have infinitely many temporal trajectories that cannot be distinguished from each other by monitoring the outputs (5). For example, a dynamical system defined by the equations $\dot{x}_1 = x_2x_4 + u, \dot{x}_2 = x_2x_3, \dot{x}_3 = \dot{x}_4 = 0$ is predicted by GA to be observable by monitoring $y = x_1$. However, the system has a family of symmetries σ_λ: $\{x_1, x_2, x_3, x_4\} \rightarrow \{x_1, \lambda x_2, x_3, x_4/\lambda\}$, so that the input u and the output y and all their derivatives are independent of λ (18). This means that we cannot distinguish whether the system is in state $(x_1, x_2, x_3, x_4)^{\mathrm{T}}$ or its symmetric counterpart $(x_1, \lambda x_2, x_3, x_4/\lambda)^{\mathrm{T}}$, because they are both consistent with the same input–output behavior. Hence, we cannot uncover the system's internal state by monitoring x_1 only. For linear systems, the symmetries correspond to topological features detectable from the inference diagram (Fig. 2*A*).

In summary, we find that for an arbitrary network topology with linear dynamics, the minimum sensor set predicted by the GA is generally not sufficient for full observability. However, the vast majority of systems of practical interest are not linear. Next we offer a rather surprising result, showing that for several much-studied nonlinear dynamical systems the symmetries in state variables are extremely rare; therefore the sensor set predicted by GA is not only necessary but also sufficient for observability. Hence, for these systems we can provide the full solution to the observability problem, which is our second and main result.

Biochemical Reaction Networks. We apply GA to biochemical reaction networks, which, with their well-characterized wiring diagram and dynamics but largely unknown parameters (kinetic constants), represent an appropriate prototype of complex systems. Consider a biochemical reaction system of N species $\{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_N\}$ involved in R reactions $\{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_R\}$ with $\mathcal{R}_j : \sum_{i=1}^{N} \alpha_{ji}\mathcal{S}_i \rightarrow \sum_{i=1}^{N} \beta_{ji}\mathcal{S}_i$, where $\alpha_{ji} \geq 0$ and $\beta_{ji} \geq 0$ are the stoichiometry coefficients. Under the continuum hypothesis and

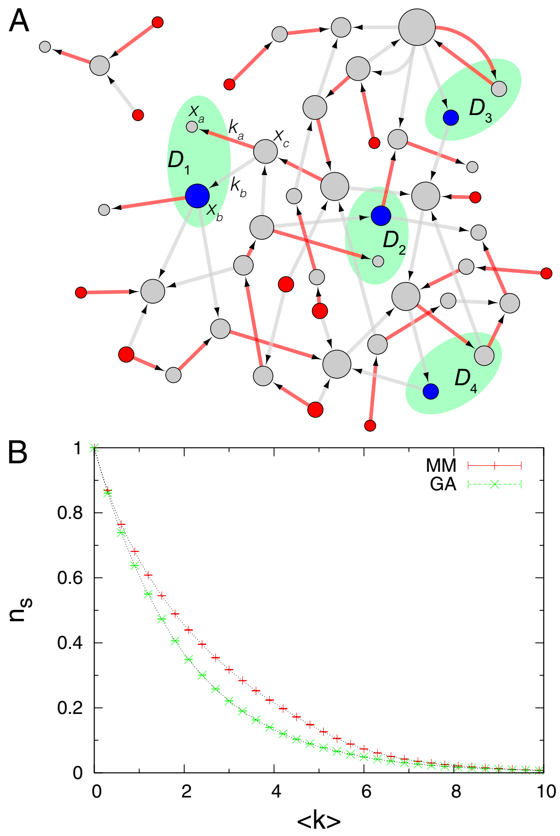

Fig. 2. Sensor nodes in linear systems. For linear systems $\dot{\mathbf{x}}(t)=\mathbf{A}\ \mathbf{x}(t)+\mathbf{B}\ \mathbf{u}(t)$ and $\mathbf{y}(t)=\mathbf{C}\ \mathbf{x}(t)$, the minimum set of sensors sufficient for full observability can be calculated exactly using the maximum matching (MM) algorithm (17), whereas the necessary sensor set is provided by the GA. (*A*) Erdös–Rényi (ER) network with mean degree $\langle k\rangle\sim 3.5$. The necessary sensor set predicted by GA is shown in red; the additional nodes that we also need to be monitor to obtain full observability are shown in blue. Hence, red and blue nodes together form the sufficient sensor set. Dilations are highlighted in green. Dilation occurs if there is a subset S of the nodes (i.e., the state variables) such that $|T(S)|<|S|$, where the neighborhood set $T(S)$ of a set S is defined to be the set of all nodes j where a directed edge exists from j to a node in S (6). Such dilations can be identified via MM algorithm. If the blue nodes are not monitored then their dilations will cause symmetries that leave the outputs and derivatives of outputs invariant. For example, in *A*, if x_b is not monitored, the subset $S_1=\{x_a, x_b\}$ will cause a dilation D_1 and a family of symmetries $\sigma_\lambda:\{\cdots,x_a,x_b,\cdots\}\to\{\cdots,x_a-k_b\lambda,x_b+k_a\lambda,\cdots\}$ that leave the outputs (and their derivatives) invariant because $\sigma_\lambda(\dot{x}_c)=\sigma_\lambda(k_ax_a+k_bx_b)=k_a(x_a-k_b\lambda)+k_b(x_b+k_a\lambda)=k_ax_a+k_bx_b=\dot{x}_c$. (*B*) n_s representing the fraction of sensors, predicted by GA (green "x") or MM (red "+"), as a function of $\langle k\rangle$ for ER random networks of size $n=10^4$. The results are averaged over 10 realizations with error bars defined as SEM. The difference between the two curves indicates that for such linear systems GA underestimates the necessary sensor set. (Similar results are also obtained for linear systems with scale-free random network topology; refs. 15, 32.) We find, however, that for nonlinear dynamics, the GA-identified nodes can be both sufficient and necessary for observability, as symmetries in state variables are very unlikely for large systems.

the well-mixed assumption (19) the system's dynamics is described by [**1**], where $x_i(t)$ is the concentration of species $\mathcal{S}_i$ at time t, the input vector $\mathbf{u}(t)$ represents regulatory signals or external nutrient concentrations, and the vector $\mathbf{y}(t)$ may capture the set of experimentally measurable species concentrations or reaction fluxes. The flux $v_j(\mathbf{x})$ of reaction $\mathcal{R}_j$ follows mass-action kinetics (20, 21)

$$v_j(\mathbf{x})=k_j\prod_{i=1}^{N}x_i^{\alpha_{ji}} \qquad [\mathbf{4}]$$

with rate constants $k_j>0$. The system's dynamics is therefore described by the balance equations

$$\dot{x}_i=f_i(\mathbf{x})=\sum_{j=1}^{R}\Gamma_{ij}\ v_j(\mathbf{x}), \qquad [\mathbf{5}]$$

where $\Gamma_{ij}=\beta_{ij}-\alpha_{ji}$ is the element of the $N\times R$ stoichiometric matrix. The right-hand side of [**5**] represents a sum of all fluxes v_j that produce and consume the species $\mathcal{S}_i$.

Assuming that the outputs $\mathbf{y}(t)$ are the concentrations of a particular set of sensor species, observability aims to identify a minimum set of sensor species from whose concentrations we can determine the concentration of all other species. The advantage of GA in this context is that it bypasses the need to know the system's kinetic constants (which are largely unknown in vivo), and only requires accurate information about the topology of the inference diagram. In the context of metabolism or an arbitrary biochemical reaction system, this is uniquely provided by the full reaction list, which is relatively accurately known for several model organisms (22).

Applying GA to biochemical reaction systems leads to a number of results that elucidate the principles behind biochemical network observability:

a) Species that are not reactants in any reaction, i.e., pure products, will be root SCCs of size 1; hence they are always sensors (e.g., x_6 in Fig. 1*C*).

b) For root SCCs of size larger than 1 (e.g., $\{x_4,x_5\}$ and $\{x_7,x_8,x_9\}$ in Fig. 1*C*), any node could be chosen as a sensor. For example, in Fig. 1*C* we can choose x_7, x_8, or x_9 as the sensor for the root SCC $\{x_7, x_8, x_9\}$. Given that some root SCCs can be quite large, and we need only one node from each root SCC, this considerably reduces the number of sensor nodes (except some pathological cases discussed in *SI Text,* section II, B.

c) The minimal number of sensor nodes that are necessary to observe a biochemical reaction system equals the number of root SCCs of its inference diagram. A minimum set of sensors consists of all pure products and one node from each root SCC with multiple species (e.g., $\{x_5, x_6, x_7\}$ in Fig. 1*C*).

d) As any node from a root SCC can be chosen as a sensor node, there are $\Omega_s=\prod_{i=1}^{N_{\text{root-SCC}}}n_i$ equivalent sensor node combinations, representing the product of all root SCCs' sizes. For example, in Fig. 1*C* we have three root SCCs with sizes $n_i=1$, 2, 3; hence $\Omega_s=1\times 2\times 3=6$. This multiplicity offers significant flexibility in selecting experimentally accessible sensors.

The principles *a*–*d* allow us to formulate the second and main finding: In biochemical networks the minimum set of sensors identified by GA is not only necessary but also sufficient for observability (*SI Text,* section II, B). To demonstrate this we randomly generated 1,000 chemical reaction systems, testing each system's observability using the rank criteria (3). We find that in all connected reaction networks the minimum set of sensors obtained by the GA achieves full observability for the whole system. This sufficiency is rooted in the fact that, when monitoring the GA-predicted minimum sensor set, the probability of developing symmetries in the state variables is close to zero. Indeed, we find that the only systems that fail sufficiency are those with isolated reactions, but such isolated reactions are not only useless from biological perspective, but their chance of occurrence goes to zero exponentially as the number of species (or reactions) increases (*SI Text,* section II, B). Hence, apart from a few pathological cases, there are always algebraic relations between the system's state variables and the successive derivatives of the outputs selected by GA, guaranteeing that the system is observable (5). In the following we discuss a series of application of our main results, demonstrating its impact on the study of biological systems.

Small Biochemical Networks. We applied GA to two well-studied biochemical reaction systems, the simplified glycolytic reaction map and a model for ligand binding, confirming that each of these systems is observable through the minimum set of sensor nodes predicted by GA (Fig. 3). The simplified glycolytic reaction map (20) consists of 10 chemical species [glucose (Gluc); ADP; glucose 6-phosphate (G6P); ATP; glucose 1-phosphate (G1P); AMP; fructose 6-phosphate (F6P); fructose 2,6-biphosphate (F26BP); triose phosphate (TP); pyruvate (Pyr)] involved in nine reactions (see Fig. 3*A* and *SI Text,* section III, A for the balance equations). GA predicts that Pyr, which is the pure product of the system, forms the only root SCC of the inference diagram (Fig. 3*B*), hence the system should be observable by measuring Pyr only. We confirm this prediction by calculating the rank of the Jacobian matrix, finding that with sufficient data points on the Pyr concentration one can reconstruct all other metabolite concentrations in the system. The core model of erythropoietin (Epo) and Epo receptor (EpoR) interaction and trafficking (23) consists of six species involved in ligand binding: Epo, EpoR, Epo_EpoR complex, internalized complex Epo_EpoR_i, degraded internalized ligand dEpo_i, and degraded extracellular ligand dEpo_e (see Fig. 3*C* and *SI Text,* section III, B for the balance equations). GA predicts that the minimum sensor set contains two pure products, dEpo_i and dEpo_e (Fig. 3*D*), confirmed via the rank test of the Jacobian matrix.

Exploring the Complete Metabolism. The developed framework is not limited to small pathways, but allows us to study an organism's genome-scale metabolism as well, involving hundreds of metabolites engaged in thousands of reactions. These systems are too large to identify the sensors via brute-force search or to explicitly verify observability via the rank condition [**3**]. However, we can efficiently identify the sensors using GA. We applied GA to the metabolic networks of three well-studied model organisms, *Escherichia coli*, *Saccharomyces cerevisiae*, and *Homo sapiens*, using their complete metabolic reconstruction (22) to identify the sensor nodes. We find that the sensor set necessary for observability of the genome-scale metabolisms represents ~5–10% of the total metabolites (Table 1). This is because the vast majority of metabolites (91% for *E. coli*, 88% for *S. cerevisiae*, and 83% for *H. sapiens*) are in a giant nonroot SCC. Overall, the GA predicts that in principle one can reconstruct the state of the whole metabolism from the concentration of a relatively small fraction of metabolites. We also find that this result is not very sensitive to the assignment of the reaction reversibility during genome-scale metabolic reconstructions (*SI Text,* section III, D).

Target Observability. Notwithstanding the fundamental importance of full observability, aiming to derive the state of each variable in a system, for most applications it is sufficient to infer the state of a certain subset of variables, that we call target variables, like the concentrations of metabolites whose activities are altered by a disease (24). If those target variables

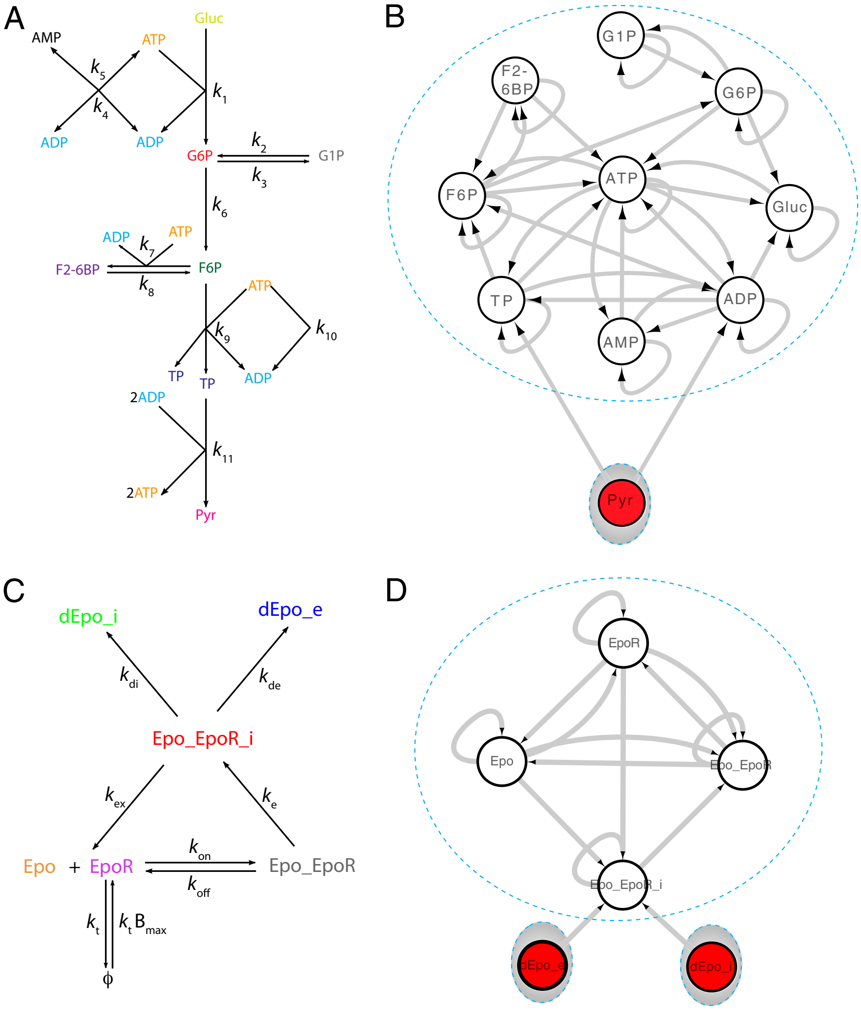

Fig. 3. Biochemical reaction systems and their inference diagrams. (*A*) Simplified glycolytic reaction map (20), where the symbols denote G6P, F6P, TP, F2-6BP, ATP (adenosine 5′-triphosphate), and ADP (adenosine 5′-diphosphate). Source (glucose) and sinks (G1P and Pyr) are also included in this model. Different chemical species are shown in different colors. (*B*) Inference diagram of the reaction system shown in *A* consists of a nonroot SCC of nine species (marked with dashed circle) and a root SCC of one species—the pure product Pyr (shaded in gray), hence indicating that the system can be observed through monitoring the concentration of Pyr only. (*C*) Simple model of ligand binding (23). The symbols denote Epo, EpoR, Epo_EpoR, Epo_EpoR_i, dEpo_i, and dEpo_e, marked with different colors. B_{max} is the maximal amount of receptor at the cell membrane. (*D*) Inference diagram of the reaction system shown in *C* consists of a nonroot SCC of four species (marked with dashed circles) and two root SCCs (shaded in gray). Each root SCC contains one pure product. The system can be observed through monitoring the concentrations of the two pure products, dEpo_i and dEpo_e.

cannot be directly measured, we can invoke target observability, identifying the optimal sensor(s) that can infer the state of the target variables, thus discovering the optimal biomarkers for the respective disease. GA helps us select such optimal sensors as well, following these general principles, representing our third main result:

i) The state of a target node x_t can be observed from a sensor node x_s only if there is a directed path from x_s to x_t in the inference diagram. For example, in Fig. 1*C* x_4 can only be inferred from x_5, whereas x_1 can be inferred from any other nodes.

ii) There are important differences in the complexity of the inference process, depending on the size of the subsystem we need to infer for a given sensor choice. The SCC decomposition of the inference diagram helps us formulate the following result: To observe x_t from x_s, we need to reconstruct $\mathcal{N}_s = \sum_{n_i \subset \mathcal{S}_s} n_i$ metabolite concentrations, where $\mathcal{S}_s$ denotes the set of all SCCs that are reachable from x_s, and n_i is the size of the *i*th SCC. This formula can be easily extended to multiple targets.

iii) To identify the optimal sensor node for any target node, we can minimize $\sum_{n_i \subset \mathcal{S}_s} n_i$, representing the minimum amount of information required for the inference process. For example, if x_t is inside an SCC of size larger than 1 (e.g., x_1 in Fig. 1*C*), then the optimal sensor can be any other node in the same SCC (e.g., x_2 or x_3 in Fig. 1*C*). If no node in the same SCC is experimentally accessible, then the optimal sensor node belongs to the smallest SCC that points to x_i (e.g., x_6 in Fig. 1*C*).

Note that this minimization procedure can be implemented for any inference diagrams in polynomial time. Hence, GA can aid the efficient selection of optimal sensors for any targeted node, offering a potentially indispensable tool for biomarker design.

Beyond Metabolism. Although we illustrated GA on biochemical reaction systems, we emphasize that the inference diagram can be constructed for arbitrary nonlinear dynamical systems of form [**1**]; hence, GA can identify the necessary sensor set for arbitrary systems. Moreover, as general nonlinear dynamical systems lack symmetries in their state variables, we expect the GA-predicted sensor set to be sufficient for observability. To show this we have explicitly verified observability for several much-studied dynamical systems, such as Michaelis–Menten kinetics in reaction dynamics (*SI Text,* section III, C), Lotka–Volterra dynamics in ecological systems (*SI Text,* section V, A), and Hindmarsh–Rose model for neuronal systems (*SI Text,* section V, B), in each case finding that the sensors identified by GA are both sufficient and necessary for observability. Given the significant current efforts to elucidate the dynamics of complex systems (25), GA is bound to find applications in a wide range of natural, socioeconomic, or technological system whose dynamics can be cast in the highly general form [**1**], helping identify optimal quantities to monitor their internal state.

Table 1. Genome-scale metabolic networks

Name*	N	L	R	N_{scc}	S_{gscc}	N_{pr}	N_{pp}	N_s
E. coli (iAF1260)	1,668	12,719	3,231	120	1,523	56	60	96
S. cerevisiae (iND750)	1,060	9,080	1,793	112	931	106	78	99
H. sapiens (Recon1)	2,763	21,026	5,283	335	2,290	166	144	293

*For each metabolic network we show the number of nodes (metabolites) N, edges (L), number of elementary reactions (R) and the number of strongly connected components (N_{scc}), and the size of the giant SCC (S_{gscc}) in the inference diagram. The table also lists the number of pure reactants (N_{pr}), pure products (N_{pp}, which are always sensor nodes), and the minimum number of sensor nodes (N_s) predicted by the graphical approach.

Discussion

In many complex systems experimental access is often limited to only a subset of state variables. Hence, we need efficient tools to identify the variables that allow us to infer the state of the whole system. Otherwise, the experiments may waste resources on measuring system variables that are redundant. Our theoretical work helps us identify the necessary sensors for an arbitrary nonlinear dynamical system, serving as the lower bound of the number of system variables we need to monitor. We also show that for many biological systems the necessary sensors are actually sufficient. Hence, our results significantly narrow the candidate variables that one needs to monitor to ensure observability. Given the unprecedented rapid development of biotechnology in the last decade, driving the development of sensitive real-time monitoring tools, our results could have implications from metabolic engineering to synthetic biology and network medicine. For example, studying the role of the GA-identified sensors in cell communication or biomarker design might offer better diagnostic tools, as well as offer rational predictions for potential biomarkers. Moreover, if we consider those unknown system parameters Θ as a special type of state variables with time derivative $\dot{\Theta} = 0$, we can extend the system to contain a larger set of state variables $\{\mathbf{x}(t), \Theta\}$. In this case we can study whether/how those system parameters can be reconstructed or identified from the input–output behavior of the extended system, using the framework of the observability problem developed here. This parameter identifiability problem has its own merit and deserves systematic study. We believe our results could shed light on this challenging problem as well.

Our work also raises a series of fundamental questions worthy of future pursuit. First, for general nonlinear systems GA cannot tell which node in a root SCC should be chosen as a sensor node. Hence in such cases identifying the sensor nodes requires detailed knowledge about the system dynamics. Second, currently the sufficiency of the predicted minimum sensor set can be checked only for rational dynamics (5), raising the need for tools capable of demonstrating the sufficiency for arbitrary complex dynamical systems, like that involved in the Kuramoto model describing synchronization in coupled oscillators (26). Furthermore, noise and measurement uncertainties will likely increase the number of sensors, the degree of which remains to be explored in the context of stochastic control (27, 28). Finally, observability only guarantees that the sensors have access to the necessary information to reconstruct the state of the whole system. To explicitly extract this information we need to construct observers, a well-developed subject in engineering control theory. We demonstrate the construction of such an observer for a linear reaction system in *SI Text,* section IV. However, the systematic adoption of these tools to natural and complex systems could open new avenues in our quest to understand complexity (29).

Materials and Methods

Maximum Matching for Linear Systems. We use the duality between controllability and observability to identify the minimum number of sensors sufficient for observability in a linear system (3, 4, 30). A linear time-invariant system $\dot{\mathbf{x}}(t) = \mathbf{A}\,\mathbf{x}(t) + \mathbf{B}\,\mathbf{u}(t)$ is called a structured system if the state matrix $\mathbf{A}$ and the input matrix $\mathbf{B}$ are structured, i.e., their elements are either fixed zeros or independent free parameters. For a structured system with state matrix $\mathbf{A}$ representing the wiring diagram of its underlying directed weighted network $G(\mathbf{A})$, the minimum number of inputs (or equivalently the minimum number of driver nodes which accept independent signals) required to fully control the system can be calculated by applying the maximum matching algorithm to $G(\mathbf{A})$ (17). For a directed network, an edge subset M is matched if no two edges in M share a common starting node or a common ending node. A node is matched if it is an ending node of an edge in the matching. Otherwise, it is unmatched. A matching of maximum cardinality/size is called a maximum matching. The minimum set of driver nodes which accepts independent signals and enables us to fully control the structured system is given by the set of unmatched nodes with respect to any maximum matchings (17). In case all nodes are matched, any single node can be chosen as the driver node. To assure controllability, each root SCC of $G(\mathbf{A})$ requires an input signal. If there is an unmatched node i inside a root SCC R, then R will be controlled by the same signal connected to node i. If all of the nodes inside a root SCC R are matched, then R can be controlled by any other signal connected to any other unmatched node in the network. Hence the number of actuator nodes,which directly accept signals, can be calculated by counting the unmatched nodes (i.e., the driver nodes) and the root SCCs inside which all of the nodes are matched. Note that by controlling those root SCCs with all nodes matched, we eliminate "inaccessibility" in the system. And, by controlling unmatched nodes, we eliminate all possible "dilations" in the system. A dilation occurs if there is a subset S of the nodes such that its neighborhood set $T(S)$, i.e., the set of all nodes j where a directed edge exists from j to a node in S, has fewer nodes than S itself (6). A structured system is controllable if and only if both inaccessibility and dilations are avoided. By invoking the duality between controllability and observability in linear system, the actuators in system $G(\mathbf{A})$ are just the sensors in its dual (or transposed) system $G(\mathbf{A}^{\mathrm{T}})$, which is obtained by flipping the direction of all edges. By monitoring those sensors, the system $G(\mathbf{A}^{\mathrm{T}})$ is guaranteed to be observable.

Observability Test of Rational Systems. To perform the algebraic observability test of rational dynamic systems, we use Sedoglavic's algorithm with a Maple implementation (5). If a system is algebraically observable, then there are algebraic relations between the state variables and the successive derivatives of the system's inputs and outputs (1, 2). These algebraic relations guarantee that the system is observable and will forbid symmetries. A family of symmetries is equivalent to infinitely many trajectories of the state variables that fit the same specified input–output behavior. If the number of such trajectories is finite, the system is locally observable. If there is a unique trajectory, the system is globally observable. If there are infinitely many trajectories, the system is not observable. Sedoglavic's algorithm tests local algebraic observability for rational systems in polynomial time. The algorithm is mainly based on the generic rank computation of the Jacobian matrix using the techniques of symbolic calculation (5). This algorithm certifies that a system is locally observable and its answer for a nonobservable system is probabilistic with high probability of success. A predicted nonobservable system and its nonobservable variables can be further analyzed to find a family of symmetries, which then can confirm the result.

Generation of Random Chemical Reaction Systems. We generate random chemical reaction systems as follows. We assure each reaction is mass balanced, i.e., it is chemically feasible with respect to mass conservation—the sum of its substrate atoms equals the sum of its product atoms (31). To generate random reaction systems under the constraint of mass balance, we start with several initial chemical compounds composing a few elements, e.g., the six most abundant elements in biological systems: carbon (C), hydrogen (H), nitrogen (N), oxygen (O), phosphorus (P), and sulfur (S). Each compound can be represented by a mass vector. For instance, the mass vector of glucose ($C_6H_{12}O_6$) is given by $m_{C_6H_{12}O_6} = (6,12,0,6,0,0)\cdot(C,H,N,O,P,S)^{\mathrm{T}}$. From the initial compounds, we can generate new compounds through chemical reactions. The stoichiometry coefficients in the reactions are randomly chosen with the constraint that the mass balance is strictly preserved. This can be achieved by tracking the mass vectors of all of the compounds. For example, starting from two compounds $C_3H_6O_3$ and $C_3H_2O_6P_1$, we may have the following reaction $C_3H_6O_3 + C_3H_2O_6P_1 \rightarrow C_3H_5O_6P_1 + C_3H_3O_3$ with a randomly assigned rate constant $k_1 > 0$. The two new compounds $C_3H_5O_6P_1$ and $C_3H_3O_3$ will then be added to the compounds pool and used to generate more compounds. Note that neither the initial nor generated compounds may exist in nature. They are just used to assure that mass balance is exactly preserved. We start from a few randomly generated compounds, and perform the above procedure to create 1,000 chemical reaction systems with up to 221 compounds involved in 121 mass-balanced reactions.

ACKNOWLEDGMENTS. We thank J. Loscalzo, A. Sedoglavic, R. Albert, J. Faeder, J. Baranyi, A. Sharma, G. Basler, P.-J. Kim, Y.-Y. Ahn, C. Song, T. Jia, B. Barzel, and Y. Li for discussions. This work was supported by the Network Science Collaborative Technology Alliance sponsored by the US Army Research Laboratory under Agreement W911NF-09-2-0053; the Defense Advanced Research Projects Agency under Agreement 11645021; the Defense Threat Reduction Agency Award WMD BRBAA07-J-2-0035; and the generous support of Lockheed Martin.

1. Diop S, Fliess M (1991) On nonlinear observability, *Proceedings of ECC'91* (Hermès, Paris), Vol 1, pp 152–157.
2. Diop S, Fliess M (1991) Nonlinear observability, identifiability, and persistent trajectories, *Proceedings of the 30th IEEE Conference on Decision and Control* (IEEE Press, New York), Vol 1, pp 714–719.
3. Kalman RE (1963) Mathematical description of linear dynamical systems. *J Soc Ind Appl Math Ser A* 1(2):152–192.
4. Luenberger DG (1979) *Introduction to Dynamic Systems: Theory, Models, & Applications* (Wiley, New York).
5. Sedoglavic A (2002) A probabilistic algorithm to test local algebraic observability in polynomial time. *J Symb Comput* 33(5):735–755.
6. Lin CT (1974) Structural controllability. *IEEE Trans Autom Control* 19(3):201–208.
7. Reinschke KJ (1988) *Multivariable Control: A Graph-Theoretic Approach (Lecture Notes in Control and Information Sciences)* (Springer, Berlin).
8. Murota K (2009) *Matrices and Matroids for Systems Analysis, Algorithms and Combinatorics* (Springer, Berlin).
9. Siddhartha J, van Schuppen JH (2001) *Modelling and control of cell reaction networks* (Centrum Wiskunde & Informatica, Amsterdam), Tech Rep PNA-R0116.
10. Khan UA, Doostmohammadian M (2011) A sensor placement and network design paradigm for future smart grids. *2011 4th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)* (IEEE Press, New York), pp 137–140.
11. Šiljak DD (1978) *Large-scale Dynamic Systems: Stability and Structure* (North-Holland, New York).
12. Khan UA, Moura JMF (2008) Distributing the Kalman filter for large-scale systems. *IEEE Trans Signal Process* 56(10):4919–4935.
13. Doostmohammadian M, Khan UA (2011) Communication strategies to ensure generic networked observability in multi-agent systems. *Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)* (IEEE Press, New York), pp 1865–1868.
14. Cormen TH, Leiserson CE, Rivest RL (1990) *Introduction to Algorithms* (MIT Press, Cambridge, MA).
15. Caldarelli G (2007) *Scale-Free Networks: Complex Webs in Nature and Technology* (Oxford Univ Press, Oxford).
16. Cohen R, Havlin S (2010) *Complex Networks: Structure, Robustness and Function* (Cambridge Univ Press, Cambridge).
17. Liu YY, Slotine JJ, Barabási AL (2011) Controllability of complex networks. *Nature* 473(7346):167–173.
18. Anguelova M April (2004) PhD thesis (Chalmers University of Technology and Göteborg University, Göteborg, Sweden).
19. Iglesias PA, Ingalls BP (2010) *Control Theory and Systems Biology* (MIT Press, Cambridge, MA).
20. Heinrich R, Schuster S (1996) *The Regulation of Cellular Systems* (Springer, Berlin).
21. Palsson BO (2006) *Systems Biology: Properties of Reconstructed Networks* (Cambridge Univ Press, Cambridge, UK).
22. Schellenberger J, Park JO, Conrad TM, Palsson BØ (2010) BiGG: A Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinf* 11:213.
23. Raue A, Becker V, Klingmüller U, Timmer J (2010) Identifiability and observability analysis for experimental design in nonlinear dynamical models. *Chaos* 20(4):045105.
24. Barabási AL, Gulbahce N, Loscalzo J (2011) Network medicine: A network-based approach to human disease. *Nat Rev Genet* 12(1):56–68.
25. Pastor-Satorras R, Vespignani A (2004) *Evolution and Structure of the Internet: A Statistical Physics Approach* (Cambridge Univ Press, Cambridge, UK).
26. Acebrón JA, Bonilla LL, Pérez Vicente CJ, Ritort F, Spigler R (2005) The Kuramoto model: A simple paradigm for synchronization phenomena. *Rev Mod Phys* 77(1): 137–185.
27. Jazwinski AH (1970) *Stochastic Processes and Filtering Theory* (Academic, New York).
28. Åström KJ (2006) *Introduction to Stochastic Control Theory* (Dover, New York).
29. Chaves M, Sontag ED (2002) State-estimators for chemical reaction networks of Feinberg-Horn-Jackson zero deficiency type. *Eur J Control* 8(4):343–359.
30. Chui CK, Chen G (1989) *Linear Systems and Optimal Control* (Springer, New York).
31. Basler G, Nikoloski Z (2011) JMassBalance: Mass-balanced randomization and analysis of metabolic networks. *Bioinformatics* 27(19):2761–2762.
32. Barabási AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286(5439):509–512.