Emergent decision-making in biological signal transduction networks

Tomáš Helikar*, John Konvalina†, Jack Heidel†, and Jim A. Rogers*†‡

*Department of Pathology and Microbiology, University of Nebraska Medical Center, 983135 Nebraska Medical Center, Omaha, NE 68198; and †Department of Mathematics, University of Nebraska, 6001 Dodge Street, Omaha, NE 68182

Edited by Eugene V. Koonin, National Institutes of Health, Bethesda, MD, and accepted by the Editorial Board December 14, 2007 (received for review May 30, 2007)

The complexity of biochemical intracellular signal transduction networks has led to speculation that the high degree of interconnectivity that exists in these networks transforms them into an information processing network. To test this hypothesis directly, a large scale model was created with the logical mechanism of each node described completely to allow simulation and dynamical analysis. Exposing the network to tens of thousands of random combinations of inputs and analyzing the combined dynamics of multiple outputs revealed a robust system capable of clustering widely varying input combinations into equivalence classes of biologically relevant cellular responses. This capability was nontrivial in that the network performed sharp, nonfuzzy classifications even in the face of added noise, a hallmark of real-world decision-making.

information processing | systems biology

Intracellular signal transduction is the process by which chemical signals from outside the cell are passed through the cytoplasm to cellular systems, such as the nucleus or cytoskeleton, where appropriate responses to those signals are generated. Unlike classical biochemical pathways (such as those involved in various metabolic activities) that are generally well understood and characterized by a degree of understandability and efficiency that can be described as elegant, signal transduction pathways are noted for their nonlinear, highly interconnected nature. Stimulation of a given cell surface receptor can induce the activation of a network of tens or even hundreds of cytoplasmic proteins; these networks are not necessarily receptor-specific because different receptors, even those associated with highly differing cellular functions, often activate common sets of proteins (1–3). How differential responses are generated by these networks is not obvious nor is the reason cells evolved such a complicated mechanism for transducing signals. Thus, a full understanding of the mechanism of intracellular signal transduction remains a major challenge in cellular biology.

Similarities in the structure of signal transduction networks to parallel distributed processing networks have led to speculation that signal transduction may involve more than simple passing along of signals. One hypothesis is that signal transduction pathways function as an information-processing system that confers nontrivial decision-making ability (4–8). The number and variety of surface receptors indicates that cells, either as single cells or as part of multicellular organisms, likely encounter a large amount of information from their environments. Thus, surface receptors function as cellular sensory systems that bring in information that must be centralized and integrated and the proper cellular response decided. Decision-making in real-world cellular environments (which are often chaotic, noisy, or contradictory) is unlikely to be relatively trivial (e.g., linear feedback) but, rather, a higher-order, nontrivial decision-making function analogous to neural networks. The ability of individual cells to process information and make nontrivial decisions would have an obvious advantage in terms of adaptation but might also characterize a fundamental difference between living and nonliving systems (9, 10).

Testing the hypothesis of signal transduction networks as nontrivial decision-making systems requires a systems biology approach because it is likely that the decision-making function is an emergent property of the entire system working in concert (11–13). Numerous studies have been performed on the static connectivity maps of signal transduction networks to compare them to other naturally occurring large-scale networks (14). The next major step to extend these results (and a crucial requirement to test explicitly for emergent functions multifamily signal transduction networks) is to study the actual dynamics of a large-scale system (15). To simulate and observe the dynamics of a system, each node's logic (or "instruction set") for activation must be determined based on the activation states of all of its regulatory inputs; i.e., the complete logic of each node in the system must be taken into account. In this study, we have created a large-scale model of signal transduction consisting of three major receptor families; receptor tyrosine kinases (RTKs), G protein-coupled receptors (GPCR), and Integrins. Using logical instruction sets for each node derived from mechanistic data in the biochemical literature, we show that this signal transduction network is able to perform nontrivial pattern recognition, a high-level activity associated with decision-making in machine learning. Nontrivial pattern recognition involves decision-making based on input information that is not necessarily clean or clear-cut; i.e., decision-making in real-world environments. In addition to the ability to classify clearly even relatively indistinct inputs, we show this pattern recognition function is robust in that it is able to perform even under high noise conditions. Together, these results are strong evidence that intracellular signal transduction networks have emergent functions that are characteristic of a nontrivial decision-making system.

Results and Discussion

Because of the highly organized nature of the cytoplasm, the size and shape of the kinetic curves representing the *in vitro* interaction of two signal transduction elements may not represent the true interaction of those elements in the cell. Because of this, continuous, differential equation-based models are difficult to parameterize realistically. This is an important limitation because the dynamics of a continuous model depend highly on the parameter values used. In cases where the function of the system being modeled is known, it is sometimes possible to reverse-engineer the parameters necessary for a continuous model (16–18). In the present study, the emergent functions of the network are only hypothesized, therefore making the reverse-engineering of parameters impossible.

To avoid the problem of parameterizing a quantitative model, a discrete Boolean model of signal transduction was created (19, 20).

Author contributions: J.K., J.H., and J.A.R. designed research; T.H., J.K., and J.A.R. performed research; T.H., J.K., and J.A.R. contributed new reagents/analytic tools; T.H., J.K., and J.A.R. analyzed data; and J.A.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. E.V.K. is a guest editor invited by the Editorial Board.

‡To whom correspondence should be addressed. E-mail: jrogers@unmc.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0705088105/DC1.

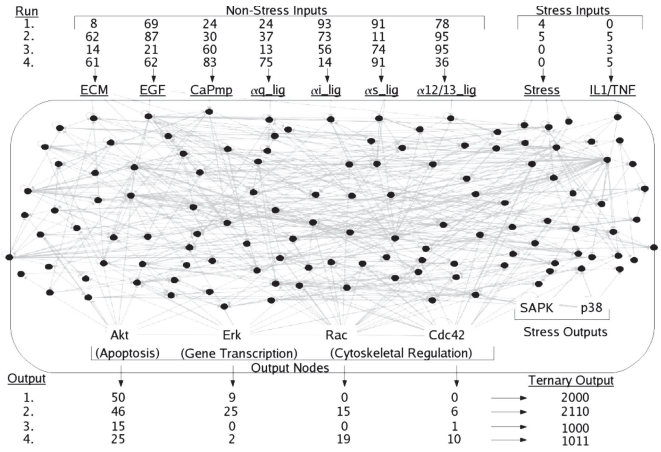

Fig. 1. The Boolean model of signal transduction and method of simulation. The actual connection graph of the 130-node Boolean model is shown inside the cell. The inputs are external to the cell and the outputs are nodes that are part of the network and thus inside the cell. The four nonstress output nodes were selected on the basis of their role in regulating other major cellular functions, as indicated. The stress outputs are the two stress-activated protein kinases SAPK and p38. As a demonstration of how simulations are performed, four random inputs are applied to the network, indicated as runs 1–4. These inputs are stress-limited because the stress inputs are limited to values between 0% and 5% ON, whereas the nonstress inputs are random values between 0% and 100% ON. After the application of each of the inputs, the network is iterated until it reaches a cycle, and the percentage ON of each output is calculated. This results in four corresponding individual outputs, shown at the bottom. The global outputs are the combination of all four individual outputs and are represented by conversion to a ternary string (shown on the bottom right) based on the ranges described in the text.

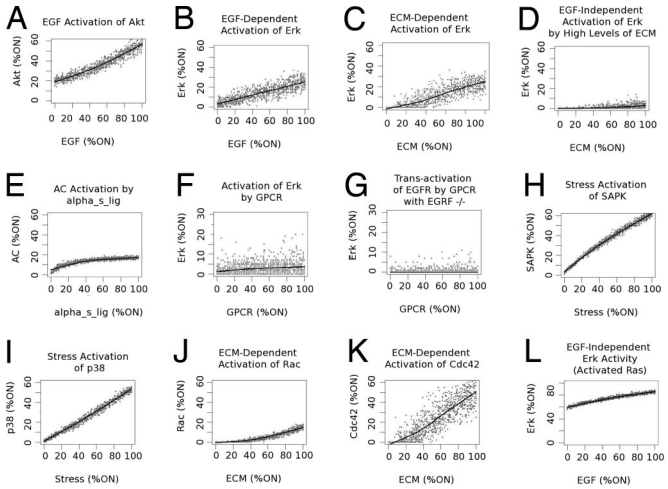

Fig. 2. Qualitative, individual input–output relationships in the Boolean model of signal transduction. (*A*) Positive relationship between EGF and Akt (25). (*B*) Positive relationship between EGF and Erk (34). (*C*) EGF dependence on Integrin stimulation by extracellular matrix (ECM) proteins for Erk stimulation (27). (*D*) Low-level stimulation of Erk by high levels of ECM (36). (*E*) Hormonal stimulators (alpha_s_lig) of G-associated GPCR activation of adenylate cyclase (AC) (37, 38). (*F*) GPCR activation of Erk. (39) (*G*) GPCR stimulation of Erk depends on transactivation of the EGFR (40). (*H* and *I*) Activation of the stress-associated MAPK's SAPK and p38 by stress (33, 34). (*J* and *K*) Activation of Rac and Cdc42 by ECM (28). (*L*) Activating mutations of known protooncogenes such as Ras result in growth factor-independent activation of Erk (41). Note that the references refer to classical, qualitative input–output relationships (not necessarily quantitative dose–response curves), and the dose–response curves presented here are intended to demonstrate how the Boolean model qualitatively reproduces the referenced input–output relationships over a range of inputs.

Because Boolean logic is qualitative in nature (21), there is no need to consider the parameters associated with the individual protein interactions (e.g., initial concentration, pH, etc.). The qualitative logic of cytoplasmic protein interactions is generally straightforward to derive from the biochemical literature, where results are usually expressed in qualitative terms (e.g., protein *x* activates protein *y*, protein *z* deactivates *y*, etc.). Beginning with the classical epidermal growth factor receptor (EGFR) to extracellular signal-regulated kinase (Erk) pathway, all upstream interactions for each member of the pathway were determined by extensive search of the literature, and a logic table (representing an instruction set) was created for each protein node. For each node, the logic table, the literature cited, and an explanation of how the logic was determined can be viewed in an online database, which can be found at http://mathbio.unomaha.edu/Database, and further details on the modeling can be found in *Materials and Methods* as well as supporting information (SI) *Text*. It should be noted that no automated methods were used in the creation of the database, rather, all papers (nearly 800) were read and all pertinent information added to the database by hand. The extent of the connectivity of the network can be seen in Fig. 1.

To test the model's ability to replicate known qualitative behaviors of the actual biological network, tests were first conducted to find the optimal input settings to do controlled experiments. This is directly analogous to optimization experiments in actual laboratory studies (e.g., determining the optimal medium and plating conditions of a cell before performing a growth factor titration). Thus, a sample of 10,000 random inputs was applied to the network and the behavior of individual outputs was correlated with selected inputs as shown in *SI Text*. Based on these results, optimized conditions were determined and controlled, qualitative input–output experiments were performed by using those conditions with the input of interest varying from 0% to 100%. The results of those controlled experiments can be seen in Fig. 2 and show that many classical, input–output functional relationships in the literature are reproduced by the model. These include the classical relationships of each family of receptors and known interdependencies between those families. These results indicate that Boolean logic can be used to describe each node of a large-scale intracellular signal transduction network qualitatively and the resulting model replicates many of the major known activities of the original system.

With a functioning, large-scale Boolean model of intracellular signal transduction in hand, the next step was to test the hypothesis of emergent information-processing functions in the system. This was accomplished by applying a sample of 10,000 random, stress-limited input combinations and categorizing the activity of the individual output nodes by using three different ranges; 0 (0–9% ON), 1 (10–29% ON), and 2 (30–100% ON), as shown in Fig. 1. Based on these categories, the combined response of all four outputs (i.e., the global response) to a given input combination can be expressed as a ternary string of length four, with each bit representing an individual output node. These ranges were chosen because they reduce the global output space to a more manageable size ($3^4 = 81$ states) and because ranges of this size are at the limits of resolution of actual laboratory data commonly used (e.g., blotting). Results presented do not depend on these ranges because experiments were performed with different ranges (from three to six) with very similar results (see *SI Text*). These runs were performed at 2% noise (a baseline noise level described more fully below).

The results of this analysis with 10,000 runs is shown in Table 1. The most striking aspect of the results shown in Table 1 is the relatively small number of global outputs. There are $3^4 = 81$ possible global outputs of the system, but after 10,000 different inputs, only 38 outputs (50% of the total global output space) are observed, many at low frequency. There are only 15 outputs that

Table 1. Outputs of the network and their average associated inputs

Global output (ternary)	Count	Average input								Average output			
		ECM	EGF	ExtPump	αq_lig	αi_lig	αs_lig	α12_13_lig	IL1_TNF	Stress	Akt	Erk	Cdc42	Rac
1000	2,346	26	26	54	49	48	50	49	2	2	19	3	2	1
2000	1,488	24	67	39	51	49	50	52	1	2	42	4	2	1
1011	952	79	21	55	49	51	50	52	2	2	19	4	19	16
2100	851	31	80	43	51	44	49	50	3	2	49	14	3	1
2110	833	58	78	45	50	53	50	48	2	2	47	15	17	5
2111	626	85	65	28	50	52	50	52	2	2	42	15	21	13
1010	463	50	34	70	46	57	50	44	2	2	21	4	15	5
2010	429	54	70	53	49	54	52	49	1	2	42	5	15	4
2121	361	89	71	55	49	57	47	42	2	2	42	17	37	14
1021	278	87	29	76	48	59	49	45	2	1	19	5	36	20
2011	232	80	46	29	51	55	50	55	1	2	35	6	19	14
1001	149	72	17	40	49	27	55	53	2	2	18	3	5	13
2120	136	69	85	74	48	56	50	38	2	2	52	18	37	6
0000	133	29	12	70	47	9	48	47	2	2	6	2	0	2
1111	112	82	34	48	53	48	54	51	4	2	24	11	21	16
1121	87	89	41	71	51	56	42	41	3	1	23	11	39	19
2021	78	87	61	63	46	63	57	49	1	1	38	6	36	14
1110	71	55	46	64	49	49	48	46	4	2	24	11	17	6
1100	58	35	50	57	51	39	52	46	4	2	25	10	4	2
2020	36	72	79	90	55	58	43	42	0	2	46	5	38	6
0011	30	80	13	90	53	16	52	47	1	2	7	2	18	18
1022	28	98	9	70	52	66	49	46	2	2	16	4	39	31
2200	26	26	97	77	47	26	52	45	4	2	68	37	1	0
2221	25	90	82	47	53	65	27	30	4	1	51	32	46	13
0001	25	79	14	73	54	2	54	45	2	3	6	2	4	14
2001	21	76	52	18	54	36	65	61	1	2	36	6	7	11
2101	20	82	73	22	53	28	40	51	3	2	45	13	6	11
2210	18	62	95	72	64	43	53	37	4	2	62	33	18	4
2220	17	74	95	75	50	53	52	41	4	2	61	33	43	6
1020	14	58	44	91	39	62	45	24	1	2	20	4	34	7
1012	13	98	5	49	46	62	37	51	1	2	18	3	26	30
2211	12	86	84	31	55	36	43	58	4	3	53	31	24	12
0010	10	48	8	90	62	17	47	52	2	2	7	2	12	6
1101	10	78	38	41	42	7	44	58	4	2	22	10	6	12
1120	4	61	56	85	55	50	34	5	4	1	26	14	40	6
0021	3	81	17	97	41	30	68	48	2	0	8	3	34	20
1122	3	96	6	47	53	85	74	21	3	1	21	13	48	31
0022	2	92	3	98	28	15	35	31	2	0	9	2	35	31

αq_lig, αi_lig, αs_lig, and α12_13_lig are abbreviations for generic ligands for the respective Gα subunit of GPCR. Other abbreviations are as in the text. Standard deviations were calculated but are not shown for clarity because the variance can be observed directly in the scatter plots in Fig. 3 and SI Fig. 5.

appear >100 times (19% of the total output space), and they account for 9,389 (94%) of the runs. This result is even more dramatic when the global outputs are categorized by using six different ranges; 0–10%, 11–20%, 21–30%, 31–40%, 41–50%, and >50%. With these six ranges, the size of the output space is $6^4 = 1{,}296$ states, yet there are only 24 outputs (1.9%) that occur >100 times, accounting for 8,597 (86%) of the inputs run (SI Table 5).

The average degree (K) of the current network is 4.4, and the average bias (P) is 69.8%. Although these parameters are predictive of relatively chaotic behavior in autonomous Boolean networks (22), the ordered behavior seen in the relatively small number of global responses may be a reflection of the high proportion of nodes (73.8%) with canalyzing inputs (23). The current network is not autonomous, so interpretation of K, P, and the effects of canalyzing inputs on network behavior may be different from the effects on the random, autonomous networks in which these parameters have been studied (21, 23). However, the fact that the network maps the wide ranging global inputs to a relatively small number of global responses indicates that the current system has a small number of attractors with large basins of attraction, which is consistent with results with autonomous networks with the similar connection parameters (21, 23).

The second prominent feature of the results is the biological significance of both the global and individual outputs observed. From the global output prospective, the output states 0000, 1000, and 2000 are prominently represented because those outputs were associated with 3,967 of the 10,000 random inputs. The outputs 1000 and 2000 represent quiescent states in which the outputs are inactive, with the exception of Akt, a protein that must remain active to suppress apoptosis (24). The state 0000 would be associated with apoptosis because Akt activity is very low. Looking at the individual qualitative input–output relationships within these three prominent global outputs, it can be seen that they are characterized by low levels of extracellular matrix (ECM) and increasing levels of EGF. This is consistent with the responsiveness of Akt activity to EGF signaling found in the literature (25, 26) and consistent with the input–output relationship of Akt and EGF presented in Fig. 2. Despite the fact that EGF increases to high levels in the 2000 output, Erk activity does not increase. As expected from the known dependence of EGF on ECM/Integrin stimulation (27), ECM levels associated with these outputs is low, and Erk activity appears in the

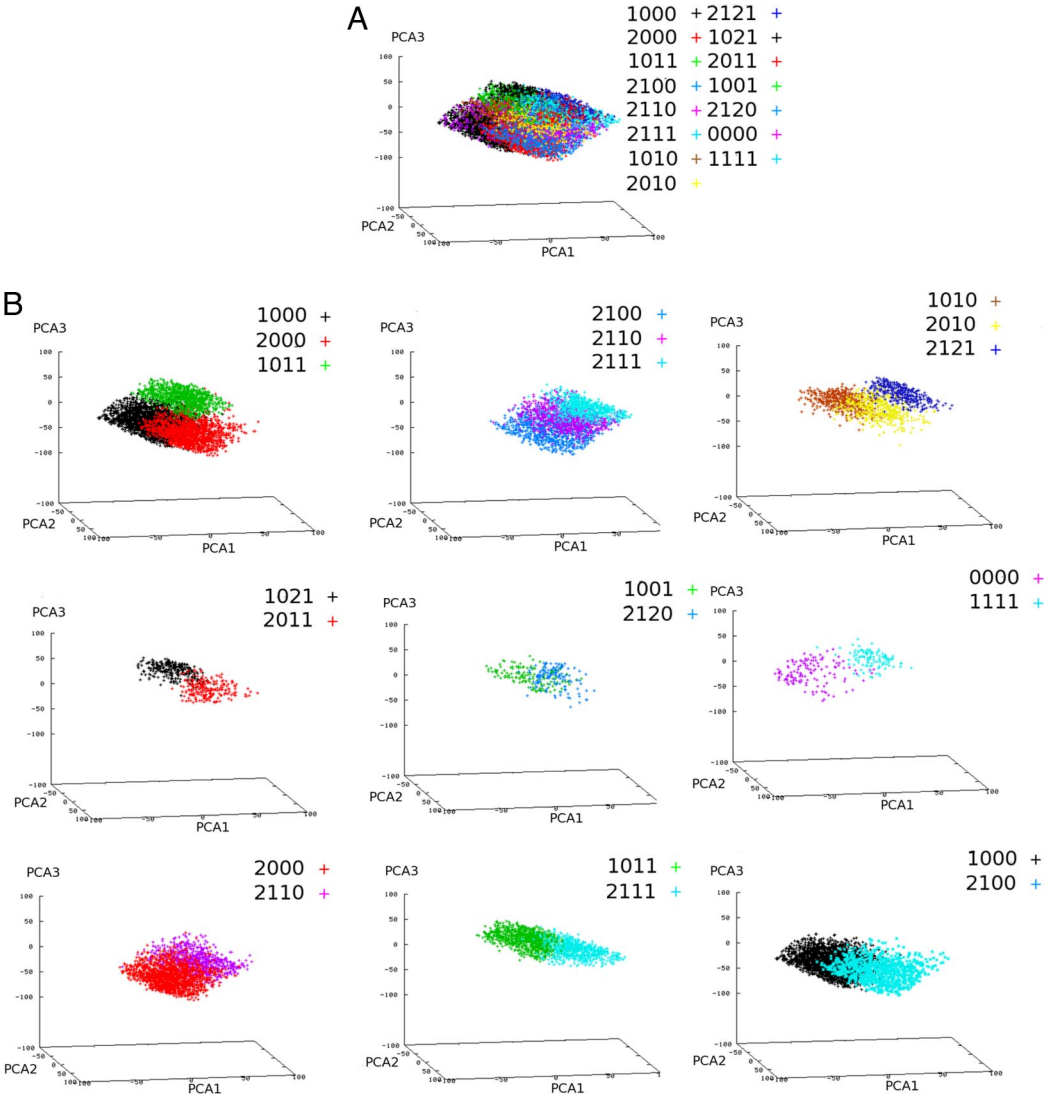

Fig. 3. Scatter plots of all input vectors associated with the first 15 global outputs of Table 1. (*A*) The inputs associated with the 15 most common outputs are plotted in three dimensions by using principle component analysis (PCA, see *Materials and Methods*). All 9,389 inputs plotted together, with each input colored according to which of the 15 outputs it is associated. It appears that all inputs associated with a given output (indicated by the color) are clustered. (*B*) To verify that the model uniquely clusters inputs based on associated outputs, selected colored clusters in *A* are plotted on separate axes so the separation of each cluster is visible. For example, the 2,346 input values associated with the output 1000 (shown as black points) are clustered with little overlap with input values associated with outputs 2000 and 1011, as shown in the first plot. Taken together, these results show that the Boolean signal transduction model divides the input space into distinct equivalence classes that are associated with biologically appropriate global outputs.

global outputs only when input levels of both ECM and EGF are high. Similarly, global outputs with increased Cdc42 and Rac activity correlate strongly with high levels of ECM; both Rac and Cdc42 are classically associated with cytoskeletal regulation in response to ECM (28). As a control, the network was randomly "rewired" 100 times, i.e., the inputs to each node were randomized while preserving the in- and out-degrees as well as the logical table of the individual nodes. The complementary control was also performed 100 times, i.e., the graph of the network was held constant while the logic was randomized. In both controls, the number of outputs diminished to a trivial number of outputs (an average effective number of 1.92 and 1.04, respectively), with no biological significance in the correlation of input and output (see *SI Text*). This indicates that both the graph and the logic are important for the variety and biological significance of the outputs.

The facts that the individual qualitative input–output relationships from Fig. 2 are present in Table 1 and that the global outputs are biologically relevant support the validity of characterizing ranges of output activity. However, the real power of this analysis is the ability to observe how the system clusters combinations of inputs and then maps them to the global outputs. To visualize this mapping at a more detailed level, all 9,389 input vectors associated with the 15 most frequent global outputs were subjected to principal component analysis (PCA) (29). In the resulting plot, shown in Fig. 3*A*, each point (representing an individual input combination) is colored according to the ternary output string with which it was associated. It shows all 9,389 inputs together, and the different colors appear to separate into discrete clusters. To confirm this, the plots in Fig. 3*B* show several combinations of different colors to indicate the degree of overlap of inputs associated with the most common global outputs. The results show that when the random input vectors are plotted in three-dimensional space based on their values, they form a random scatter as expected. But when each vector in that scatter plot is colored based on the global output with which it is associated, all of the input vectors associated with a particular output are not randomly scattered but, rather, clustered in distinct areas with little overlap with inputs associated with other outputs. Thus, this signal transduction system clusters neighborhoods of input combinations into equivalence classes of global outputs; i.e., all input combinations of the same color are considered to be functionally equivalent because they elicit the same global output response. The 100 randomly rewired and random logic control networks were also tested for separation and in rewired networks where there was more than one output to test, the number of outputs that demonstrated clustering of inputs was greatly reduced and separation in the random logic networks was eliminated. The details of PCA, how it was applied to the network, statistical analysis, results with the rewired controls, and further

discussion of the biological relevance of these results can be found in *Materials and Methods* and *SI Text*.

The results presented show that this signal transduction network model is capable of taking a wide array of random hormonal input combinations and classify them into a relatively small number of biologically appropriate, sharply defined equivalence classes of global responses. This function can, by definition, be called pattern recognition, a concept used in machine learning and neural networks (30, 31). The ability to recognize input patterns and classify them is a decision-making function that is a form of information processing. It involves dividing a multidimensional space into associated classes, the boundaries of which must be carefully determined to recognize inputs correctly based on their class association (30). The practical effect of this type of processing is that the very large number of combinations of possible hormonal inputs to which a cell may be exposed (many of which are relatively indistinct) are clustered by the signal transduction network according to the much smaller number of global cellular responses that are possible for a cell to make to each input. Thus, this network is able to make decisions even in the face of less than clear-cut environmental cues that are common in realistic environments.

To determine the robustness of signal transduction decision-making, the above experiments were carried out with different noise levels. For example, if in a given run, an input such as EGF is set to 50% ON, no added noise would mean that the node is a exactly 50% throughout the entire run. Adding 2% noise to a 50% value would mean that the input varied chaotically between 48% and 52% with an average of 50%. Five percent noise for that input would result in the input varying chaotically between 45% and 55%, and so on. In the above simulations, noise was added at what was considered to be a normal background level of 2%. The results of testing of other noise levels of up to 20% can be seen in SI Tables 7–9. It is clear that the pattern recognition ability of the network in terms of global responses is nearly unaffected by even high levels of noise. This surprising level of stability was verified by repeating the individual input–output relationships of Fig. 2 with varying levels of noise. Even at high levels of noise, the input–output relations remained intact (data shown SI Fig. 6), confirming the ability of the system to recognize patterns in even very noisy inputs.

It has long been noticed that complex, interconnected pathways of biochemical signal transduction networks bear a resemblance to parallel, distributed computer networks. This led to conjecture by some that the overwhelming complexity of these networks might not be an accident of evolution but, rather, a key characteristic of a finely tuned information-processing system that is able to make nontrivial decisions (5, 6, 31, 32). To test this hypothesis directly, we have created a large-scale, literature-based, logically complete model of a multifamily signal transduction network. The model was then exposed to tens of thousands of different combinations of environmental stimuli, and the global responses (i.e., combinations of multiple outputs) were observed. The reason for this approach was to look for emergent properties of the system by moving beyond the exploration of the important and now well established dynamics of specific, individual stimulus–response relationships (e.g., bistability) and consider the higher-level relationships between multiple stimuli and the corresponding global responses. This is the essence of the systems approach.

The results clearly show that the network clustered the vast majority of inputs into a small number of biologically appropriate responses. This nonfuzzy partitioning of a space of random, noisy, chaotic inputs into a small number of equivalence classes is a hallmark of a pattern recognition machine and is strong evidence that signal transduction networks are decision-making systems that process information obtained at the membrane rather than simply passing unmodified signals downstream.

Designing systems to perform sophisticated pattern recognition is not a trivial task. Handwriting and face recognition are examples of real-world, sophisticated pattern recognition where noisy inputs must be correctly placed into a nonfuzzy, sharply defined equivalence class (e.g., individual handwriting classified as a particular character or an individual face recognized as an acquaintance); a major goal of artificial intelligence research is to develop machines that are capable of such tasks (30, 31). It should not be surprising that cells would require a similar ability to perform sophisticated pattern recognition. An individual cell is faced with any number of stimuli in the form of chemical ligands binding to their cell-surface receptors. These receptors are varied in type and number and form a sensory system that enables a cell to sense and respond to its environment. Given that any physical environment is chaotic, noisy, and, at times, contradictory, it is clear that cells need the ability to make decisions based on these types of inputs and that their survival depends on that ability.

Finally, the results presented here use literature-based, Boolean modeling of a large-scale biochemical system. All modeling methods have their downsides, and in the case of Boolean models it is that the logic of each node must be expressed in terms of ON/OFF. This seems counterintuitive to many biologists because it is known that many signal transduction components do not have such simple regulation. In reality, this is not a major obstacle to Boolean modeling because proteins that exhibit more complex regulation can be represented by multiple nodes, each representing a separate activation state of the protein of interest (e.g., Raf in our network). The only real downside to Boolean modeling of biochemical systems is that, for any node with a large number of inputs (N), there are 2^N combinations of those inputs that must be accounted for in each logic table. For most input combinations the ON/OFF state of the protein can be derived from the literature in a straightforward way. However, some combinations are not explicitly dealt with in the literature and must be deduced indirectly. For this reason, we do not consider the current network to be perfect. However, this problem is not unsolvable; it only requires laboratory researchers to test qualitatively the input combinations that are unknown. This can be done exactly as the known combinations were determined, thus requiring only an awareness for the need for this information rather than entirely new laboratory methods. Additionally, our development of tools that are able to input and retrieve continuous data to and from the Boolean model means that the only aspect of the model that is actually ON/OFF is the logic tables for each individual node; once the logic is set, the model is used in the same way as continuous models.

Continuous modeling, on the other hand, has the significant problem of parameter estimation. Although there are also ways of dealing with this problem, determination of the large number of parameters of a large-scale network *in vivo* is a much more complicated technical hurdle. All modeling methods also have upsides, and the parameter-free nature of Boolean modeling is a significant advantage, making it complementary to continuous models used for exploring higher-order functions and emergent properties of biological signal transduction networks.

Materials and Methods

The Boolean Model of Signal Transduction. To create a Boolean network, a set of nodes must be identified and a logic table created for each node. The current Boolean model of signal transduction was created by determining the complete logic of the classical EGFR → Erk pathway. More detail on how the logic tables were created for each node (as well as how it is possible to use Boolean modeling for proteins that have more complex activation than simple ON/OFF) can be found in *SI Text*. As guided by the literature, connections to other classical pathways were included in the EGFR → Erk pathway until a relatively autonomous network of 130 nodes was created that included the RTK, GPCR, and Integrin pathways. Given the highly interconnected nature of cytoplasmic protein networks, stopping at even 130 nodes meant that some interactions with proteins outside these three families had to be ignored. However, these were relatively minor compared with the interactions of the three incorporated pathways; these pathways are so intimately connected that they represent a functioning set of nodes that would be impossible to reduce further without ignoring important interactions. Although the model is a nonspecific network in that it

does not represent any one specific cell type, nodes were not included in the network unless they were generally expressed in a wide range of cell types. However, once a node was included in the network, the best information on the logic was used without regard to the cell type.

Inputs to the network are mostly the ligands of surface receptor nodes of which there are seven; epidermal growth factor (EGF), the GPCR stimulators (αq_lig, αi_lig, αs_lig, αi12/13_lig), extracellular matrix (ECM), tumor necrosis factor/interleukin 1 (TNF/IL-1) an idealized hybrid receptor. In addition to those ligands, ''Stress'' is an input representing environmental stress factors such as UV light or reactive oxygen species (33, 34), and there is a nonregulated calcium pump (external calcium pump). Calcium pumps are regulated mostly by calcium and hormonal factors that are not included in the current network (35). The logic for calcium regulation in the network inherently includes the calcium regulation of the pump, but the other regulators cannot be accounted for. Therefore, the calcium pump is considered to be an input and is set at multiple constant levels of activity. Outputs of the network are nodes in the network whose outputs go out to regulators of major cellular functions. These are (*i*) Akt, a major regulator of apoptotic systems (24), (*ii*) the mitogen-activated protein kinase (MAPK) Erk, a major regulator of cell division (34), and (*iii*) Rac and Cdc42, two important regulators of cytoskeletal systems (28). The MAPK's SAPK and p38 are also outputs of the network, but they respond to stress and TNF/IL-1 (33, 34), as documented in *SI Text*. The experiments in the present work involve ''stress-limited'' inputs, meaning that stress and TNF/IL-1 are at low, background levels. Other nodes can be considered to be outputs of the network, and experiments with up to seven different outputs were performed with very similar results.

Methods of Simulation. Although the logic of each node and input to the network is binary, it is possible to interpret intermediate activity by looking at the average ON value of each node when the system has reached a cycle, as all Boolean models must do. Similarly, inputs can be set to a specific average ON by putting the input on an appropriate cycle (for further details on this, see *SI Text*).

The actual simulations are performed by a Boolean simulation program developed by this group called ChemChains. ChemChains is a general Boolean network simulator that is able to incorporate any number of nodes and their logic tables as well as any initial condition or input conditions and iterate the network any desired number of times. The ChemChains program can be freely obtained from J.A.R.

Adding Noise to the Inputs. Experiments are performed at different noise levels by introducing a random noise component to the input that forces the input to vary chaotically within a window around the set input level. The window sizes vary from 2% to 20%, representing background noise to highly noisy inputs. In these studies, all noise levels are tested and 2% noise is considered the standard background levels. Noise was added to the inputs by adding or subtracting from each input set point a percentage of the desired range. The percentage varied randomly within the specified range by using a random number generator within the ChemChains program.

Principal Component Analysis (PCA). PCA was done on all seven inputs and projected onto three dimensions (accounting for 45% of variance of the system) as shown in SI Fig. 5, where a more detailed explanation of PCA and the statistical analysis of the results can be found. To capture more of the variance, various numbers of inputs were tested and it was found that most of the pattern recognition function could be observed by performing PCA on ECM, EGFR, and the external calcium pump inputs and projecting onto three dimensions. This accounts for 100% of the variance and makes the clustering of inputs the most clearly visible. These results are shown in Fig. 3; however, they are not fundamentally different from the original seven-input PCA shown in SI Fig. 5.

ACKNOWLEDGMENTS. We thank J. Maloney for initial help with MAPLE programming, J. Hamilton for creating the initial version of ChemChains, C. Ramey for helpful discussions and creating the PCA program, and S. From for consultation on statistical methods. This work was supported by National Institutes of Health Grant GM067272.

1. Jordan JD, Iyengar R (1998) Modes of interactions between signaling pathways. *Biochem Pharmacol* 55:1347–1352.
2. Fambrough D, McClure K, Kazlauskas A, Lander ES (1999) Diverse signaling pathways activated by growth factor receptors induce broadly overlapping, rather than independent, sets of genes. *Cell* 97:727–741.
3. Jordan JD, Landau EM, Iyengar R (2000) Signaling networks: the origins of cellular multitasking. *Cell* 103:193–200.
4. Bray D (1990) Intracellular signalling as a parallel distributed process. *J Theor Biol* 143:215–231.
5. Bray D (1995) Protein molecules as computational elements in living cells. *Nature* 376:307–312.
6. Schamel WW, Dick TP (1996) Signal transduction: Specificity of growth factors explained by parallel distributed processing. *Med Hypotheses* 47:249–255.
7. Bhalla US, Iyengar R (1999) Emergent properties of networks of biological signaling pathways. *Science* 283:381–387.
8. Fernandez P, Sole RV (2005) in *Power Laws. Scale-Free Networks and Genome Biology*, eds Koonin EV, Wolf YI, Karev GP (Springer, New York), pp 206–224.
9. Hopfield JJ (1994) Physics, computation, and why biology looks so different. *J Theor Biol* 171:53–60.
10. Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. *Nature* 402:C47–52.
11. Kitano H (2002) Systems biology: A brief overview. *Science* 295:1662–1664.
12. Cho K, Wolkenhauer O (2003) Analysis and modelling of signal transduction pathways in systems biology. *Biochem Soc Trans* 31:1503–1509.
13. Van Regenmortel MHV (2004) Reductionism and complexity in molecular biology. Scientists now have the tools to unravel biological and overcome the limitations of reductionism. *EMBO Rep* 5:1016–1020.
14. Albert R (2005) Scale-free networks in cell biology. *J Cell Sci* 118:4947–4957.
15. Zhu H, Huang S, Dhar P (2003) The next step in systems biology: Simulating the temporospatial dynamics of molecular network. *BioEssays* 26:68–72.
16. Rodriguez-Fernandez M, Mendes P, Banga JR (2006) A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *Biosystems* 83:248–265.
17. Chou I, Martens H, Voit EO (2006) Parameter estimation in biochemical systems models with alternating regression. *Theor Biol Med Model* 3:25.
18. Koh G, Teong HFC, Clément M, Hsu D, Thiagarajan PS (2006) A decompositional approach to parameter estimation in pathway modeling: A case study of the Akt and MAPK pathways and their crosstalk. *Bioinformatics* 22:e271–80.
19. Huang S (2001) Genomics, complexity and drug discovery: Insights from Boolean network models of cellular regulation. *Pharmacogenomics* 2:203–222.
20. Shmulevich I, Dougherty ER, Zhang W (2002) From Boolean to probabilistic Boolean networks as models of genetic regulatory networks. *Proc IEEE* 90:1778–1792.
21. Kauffman SA (1993) in *The Origins of Order* (Oxford Univ Press, New York), pp 188–235.
22. Aldana M, Cluzel P (2003) A natural class of robust networks. *Proc Natl Acad Sci USA* 100:8710–8714.
23. Kauffman SA (1993) in *The Origins of Order* (Oxford Univ Press, New York), pp 441–481.
24. Kumar CC, Madison V (2005) AKT crystal structure and AKT-specific inhibitors. *Oncogene* 24:7493–7501.
25. Du K, Tsichlis PN (2005) Regulation of the Akt kinase by interacting proteins. *Oncogene* 24:7401–7409.
26. Kassenbrock CK, Hunter S, Garl P, Johnson GL, Anderson SM (2002) Inhibition of Src family kinases blocks epidermal growth factor (EGF)-induced activation of Akt, phosphorylation of c-Cbl, and ubiquitination of the EGF receptor. *J Biol Chem* 277:24967–24975.
27. Edin ML, Juliano RL (2005) Raf-1 serine 338 phosphorylation plays a key role in adhesion-dependent activation of extracellular signal-regulated kinase by epidermal growth factor. *Mol Cell Biol* 25:4466–4475.
28. Price LS, Leng J, Schwartz MA, Bokoch GM (1998) Activation of Rac and Cdc42 by integrins mediates cell spreading. *Mol Biol Cell* 9:1863–1871.
29. Wall ME, Rechtsteiner A, Rocha LM (2003) in A practical approach to microarray data analysis, eds Berrar DP, Dubitzky W, Granzow M (Kluwer, Boston), pp 91–109.
30. Haykin S (1999) in Neural Networks, A Comprehensive Foundation (Prentice–Hall, Englewood Cliffs, NJ), pp 66–67.
31. Bray D (2003) Molecular networks: The top-down view. *Science* 301:1864–1865.
32. Ma'ayan A, *et al.* (2005) Formation of regulatory patterns during signal propagation in a mammalian cellular network. *Science* 309:1078–1083.
33. Takeda K, Matsuzawa A, Nishitoh H, Ichijo H (2003) Roles of MAPKKK ASK1 in stress-induced cell death. *Cell Struct Funct* 28:23–29.
34. Roux PP, Blenis J (2004) ERK and p38 MAPK-activated protein kinases: A family of protein kinases with diverse biological functions. *Microbiol Mol Biol Rev* 68:320–344.
35. Strehler EE, Treiman M (2004) Calcium pumps of plasma membrane and cell interior. *Curr Mol Med* 4:323–335.
36. Brakebusch C, Bouvard D, Stanchi F, Sakai T, Fässler R (2002) Integrins in invasive growth. *J Clin Invest* 109:999–1006.
37. Milligan G, White JH (2001) Protein–protein interactions at G-protein-coupled receptors. *Trends Pharmacol Sci* 22:513–518.
38. Selbie LA, Hill SJ (1998) G protein-coupled-receptor cross-talk: the fine-tuning of multiple receptor-signalling pathways. *Trends Pharmacol Sci* 19:87–93.
39. Naor Z, Benard O, Seger R (2000) Activation of MAPK cascades by G-protein-coupled receptors: the case of gonadotropin-releasing hormone receptor. *Trends Endocrinol Metab* 11:91–99.
40. Gschwind A, Zwick E, Prenzel N, Leserer M, Ullrich A (2001) Cell communication networks: epidermal growth factor receptor transactivation as the paradigm for interreceptor signal transmission. *Oncogene* 20:1594–1600.
41. Sridhar SS, Hedley D, Siu LL (2005) Raf kinase as a target for anticancer therapeutics. *Mol Cancer Ther* 4:677–685.