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.

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.

Similarly, 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 multilamellar 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).
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 behavior 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
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 proportion of nodes (73.8%) with canalyzing inputs (23). The number of global responses may be a reflection of the high degree 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 appear 100 times, accounting for 8,597 (86%) of the inputs run (SI Fig. 5).
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. 3A, 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 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.

![Figure 3](https://www.pnas.org/cgi/doi/10.1073/pnas.0705088105)
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 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 the 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 not trivial; in particular, comparison of the interactions of the three incorporated pathways with these 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, α1l/lig, αs/lig, α12/13/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 also inputs that 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 network is binary, it is possible to interpret intermediate activity by looking at the different outputs were performed with very similar results. 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 “stressless” 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.

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.