COMPUTATIONAL BIOLOGY

Reduction of Complex Signaling Networks to a Representative Kernel

Jeong-Rae Kim,[1,2] Junil Kim,[1] Yung-Keun Kwon,[1,3] Hwang-Yeol Lee,[1] Pat Heslop-Harrison,[4] Kwang-Hyun Cho[1]*

The network of biomolecular interactions that occurs within cells is large and complex. When such a network is analyzed, it can be helpful to reduce the complexity of the network to a "kernel" that maintains the essential regulatory functions for the output under consideration. We developed an algorithm to identify such a kernel and showed that the resultant kernel preserves the network dynamics. Using an integrated network of all of the human signaling pathways retrieved from the KEGG (Kyoto Encyclopedia of Genes and Genomes) database, we identified this network's kernel and compared the properties of the kernel to those of the original network. We found that the percentage of essential genes to the genes encoding nodes outside of the kernel was about 10%, whereas ~32% of the genes encoding nodes within the kernel were essential. In addition, we found that 95% of the kernel nodes corresponded to Mendelian disease genes and that 93% of synthetic lethal pairs associated with the network were contained in the kernel. Genes corresponding to nodes in the kernel had low evolutionary rates, were ubiquitously expressed in various tissues, and were well conserved between species. Furthermore, kernel genes included many drug targets, suggesting that other kernel nodes may be potential drug targets. Owing to the simplification of the entire network, the efficient modeling of a large-scale signaling network and an understanding of the core structure within a complex framework become possible.

INTRODUCTION

Cellular systems have evolved molecular interaction networks to maintain their complex regulatory functions, which allow cells to perform processes such as differentiation and to respond to the environment. We speculated that such interaction networks were built around certain core structures or "kernels," which would be simpler to analyze without losing essential information. An individual kernel can be defined broadly as a simplified framework of a given complex interaction network that preserves the dynamics and the output of the original network. Identification of such kernels would enable insights into the organization and evolution of biomolecular interaction networks, allow the generation of representative but simplified representations of complete networks that can be modeled, and, eventually, facilitate the exploration of interventions to manipulate cellular systems to perform desired responses (*1*).

When discussing biological networks, we refer to the proteins or genes as "nodes" and the relationships between the proteins or genes as "edges." Although the networks are typically constructed with the names of the encoding genes, the functions are assumed to be performed by the encoded proteins or RNAs. Two general approaches to the study of biological networks are (i) component-wise analysis of individual components in the networks, as in studies of "minimal gene sets," and (ii) computational analysis of simplified networks. Several studies have investigated the minimal gene sets (*2*) required for survival, using computational approaches (*3*–*5*) or experiments with bacterial mutants (*2*, *6*, *7*). One limitation of these component-wise approaches is that they cannot take into account regulatory interactions among the genes. The methods that involve simplifying complex networks generally strive to preserve "static" topological properties, such as the small-world property, scale-freeness, fractality, or modularity (*8*–*21*), and can largely be classified into two categories (*17*), coarse graining and filtering or pruning. Coarse graining refers to the grouping of nodes with respect to various topological properties and replacing each group of nodes with a single node called a coarse-graining unit (CGU), thereby achieving a simpler network representation (*12*, *18*). The filtering or pruning approach deletes nodes classified as less important from scores assigned to the nodes on the basis of the network's topological characteristics. One limitation of these simplified network approaches is that, by primarily focusing on preserving static topological properties of general complex networks, they fail to preserve the dynamical properties of cellular signaling networks. Cellular signaling networks exhibit properties, such as feedback loops, that make preservation of dynamical properties challenging.

The spanning tree network reduction approach of Kim *et al.* (*13*) reduces only the number of edges while preserving all the nodes of the original network. Because the resulting simplified network is a tree, it cannot preserve the dynamics of the original network if the original contains feedback loops (*22*–*25*) or feedforward loops (*26*, *27*). The approach taken by Itzkovitz *et al.* (*12*) replaces network motifs with CGUs, which in principle can preserve the dynamics of a network only if the intrinsic dynamics of each network motif are identically implemented in the CGU of the reduced network. However, it remains unclear how to implement such identical dynamics at each CGU. Song *et al.* (*18*) proposed a reduction scheme that tiles a network with boxes such that the shortest path length of any two nodes in a box is less than a given number called a box size, where the size of the box is $1 + m$, with m the maximum of the shortest paths between two nodes in the box. However, the resulting network does not contain any information on the direction or interaction type (activation or inhibition) of the edges; thus, preservation of dynamic properties is not possible. Using network symmetry, Xiao *et al.* (*19*) proposed a network reduction scheme in which a set of nodes is grouped as one node if the

[1]Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Republic of Korea. [2]Department of Mathematics, University of Seoul, Seoul 130-743, Republic of Korea. [3]School of Computer Science and Information Technology, University of Ulsan, Ulsan 680-749, Republic of Korea. [4]Department of Biology, University of Leicester, Leicester LE1 7RH, UK.
*To whom correspondence should be addressed. E-mail: ckh@kaist.ac.kr

rearrangement of their position within the set does not change the network topology. This approach can be effectively applied to a gene network containing many functionally redundant genes, but it is not effectively applicable to cell signaling networks that usually contain many long cascades.

Here, we describe the "kernel identification algorithm," which is an algorithm that identifies a kernel systematically by considering the relationship between a network's structure and its dynamics. Because of the enormous complexity of biomolecular interaction networks, it is not computationally feasible to find a representative kernel by simultaneously taking into account the dynamics of all possible subnetwork cases. The kernel identification algorithm overcomes this difficulty by recursive sequential replacement of the neighborhood subnetwork of each node with a smaller one that preserved the same dynamics. The neighborhood subnetwork of a node is the network composed of the nodes directly connected to the given node. We show that our algorithm can be applicable to large-scale cell signaling networks to produce smaller, simpler networks that retain the original network's dynamics. Although some coarse-graining methods, such as fractal analysis (*11*, *18*), also perform repetitive substitutions of subnetworks with smaller ones and are as efficient as our method in terms of computational complexity, they generally fail to preserve the dynamical properties of a network.

By applying the kernel identification algorithm, we identified kernels for various signaling networks ranging from bacterium (*Escherichia coli*) and yeast (*Saccharomyces cerevisiae*) to human, and we verified that the identified kernels preserved the input-output dynamics of the original networks. We found that a large proportion of the nodes within the kernels (kernel nodes) corresponded to essential genes, disease-associated genes, genes encoding drug targets, or genes that are part of synthetic lethal gene pairs. Moreover, we found that kernel nodes were encoded by genes conserved in multiple species, suggesting low evolutionary rates, and encoded proteins present in various tissues, suggesting that these kernel-associated genes may serve core cellular functions. The kernel identification algorithm can provide a reduced form of a given network, and this smaller network may provide insight into the design principles of complex biomolecular interaction networks, as well as suggest effective ways to perturb or manipulate the network.

RESULTS

Kernel identification algorithm

We wanted to develop an algorithm that preserved the input and output nodes of a biomolecular interaction network and the input-output dynamics of the original network while reducing the complexity of the network (Fig. 1A). An input node in a network denotes a node without any regulatory inputs (indegree is zero). Likewise, an output node denotes any node that lacks any relationships with downstream nodes (outdegree is zero). For example, in some signaling networks, ligands or receptors (when ligands are not specified) may correspond to input nodes, and transcription factors (when their target genes are not specified) may correspond to output nodes. The remaining nodes in a network are intermediate nodes. We developed an algorithm that minimized the number of intermediate nodes by replacing certain subnetworks within a large network with smaller subnetworks.

To overcome the computational burden that would result from analyzing simultaneously all possible dynamics of biological networks and their subnetworks, the kernel identification network recursively replaces the neighborhood subnetwork of each node with a smaller network, either with fewer nodes or fewer edges or both, with the same dynamics until no further replacement is possible.

To determine the rules for subnetwork replacement, we developed and simulated the mathematical models of all two- and three-node networks with ordinary differential equations (see Materials and Methods and Supplementary Model Descriptions), and then clustered the two- and three-node networks according to the similarity in their dynamics (Fig. 1B). We verified that the clustering assignments were similar between linear and Hill-type mathematical models and among the parameter values used (see fig. S1 and Supplementary Model Descriptions). On the basis of the

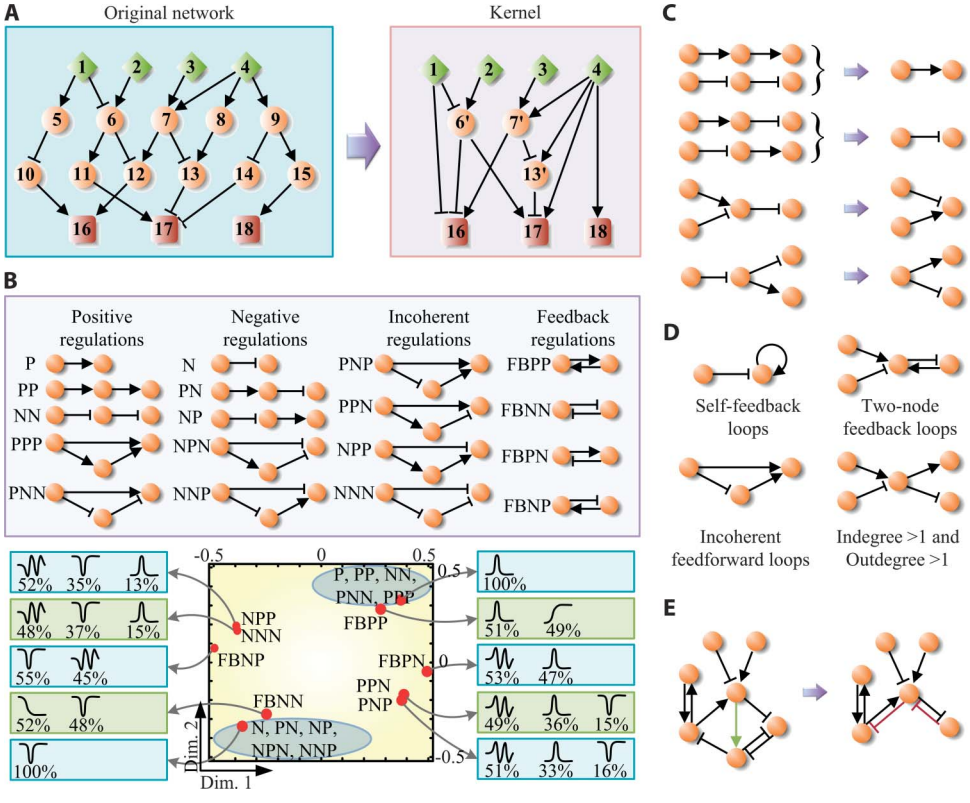
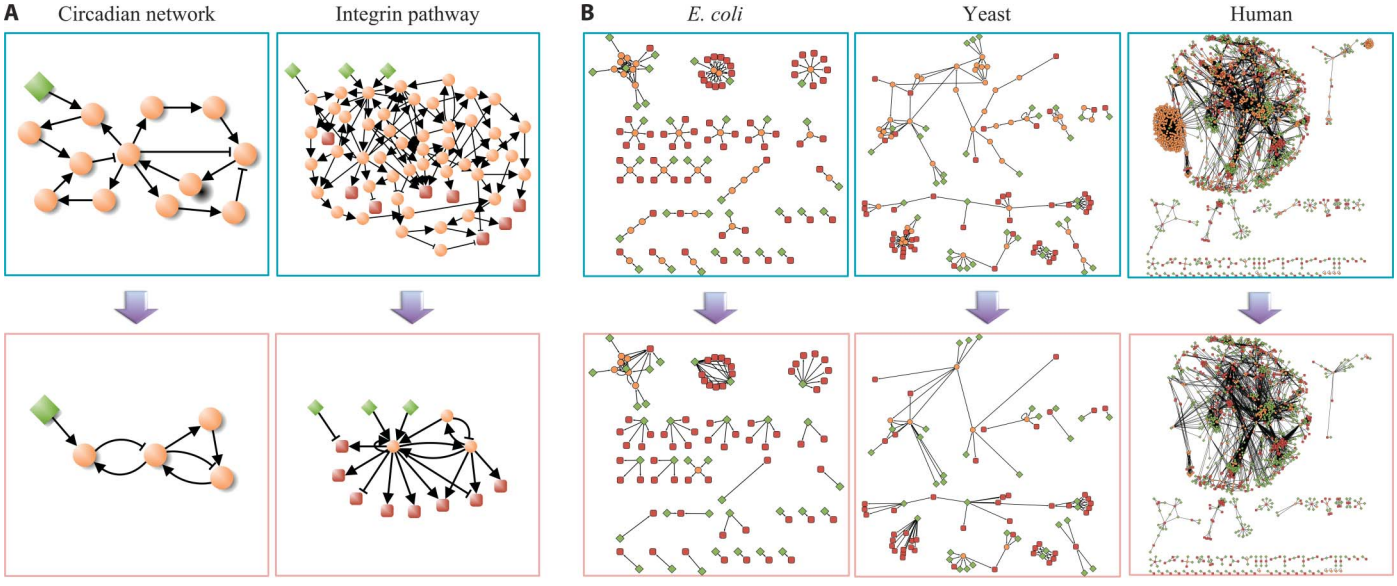


Fig. 1. Scheme of the kernel identification algorithm. (A) Illustration of the kernel identification that preserves the fundamental dynamics of the original network with a node reduction percentage of 73%. (B) Multi-dimensional scaling map for responses of two- and three-node networks (see Materials and Methods for details). If two networks are close on the map, their responses are considered similar. The top box shows the subnetworks. The graph below shows the clustering of the positive and negative regulatory subnetworks (circled in gray). The response curves for each group or subnetwork are shown on either side of the graph. (C) Examples of situations leading to subnetwork replacement. (D) Subnetworks that cannot be reduced. (E) Illustration of subnetwork replacement around an edge.

clustered networks, the algorithm attempts to replace the neighborhood subnetwork of each node with a smaller network (see Fig. 1C for examples of subnetwork replacement and Materials and Methods and fig. S2 for details). The algorithm cannot replace subnetworks either (i) when one node in a three-node subnetwork is also a component node of a self-feedback loop, a two-node feedback loop, or an intermediate node of an incoherent feedforward loop, or (ii) when both the indegree and the outdegree of the node are >1 (Fig. 1D). When a network cannot be reduced any further by the above reduction process, the algorithm reduces the network by replacing the neighborhood subnetwork of a set of edges, taking into account consistency of the types of regulation among the neighboring edges (see Fig. 1E for an example and Materials and Methods and fig. S2 for details). We defined the "node reduction percentage" as [(the number of intermediate nodes removed during reduction)/(the number of intermediate nodes in the original network)] × 100. For the sample network shown in Fig. 1A, the node reduction percentage equals 73% [(8/11) × 100].

We applied the algorithm to the signaling networks of *E. coli*, *S. cerevisiae*, and *H. sapiens*, where we define the signaling network as an integrated network of all the signaling pathways obtained from the KEGG (Kyoto Encyclopedia of Genes and Genomes) database (*28*) for each species, and identified the kernels of those networks (data S3 to S5). We refer to these three networks as the *E. coli*, yeast, and human signaling networks. Through Boolean simulations (*29*, *30*), we verified that the kernels preserved the dynamical properties of the input-output response profiles of the original networks (see Supplementary Model Descriptions and table S1). Because it is not feasible to construct and simulate large-scale networks, such as the human signaling network, by ordinary differential equations, we used Boolean models to verify the preservation of network dynamics.



C

	E. coli	Yeast	Human
Number of the total nodes	**129**	**129**	**1953**
Number of the input nodes	**35**	**36**	**669**
Number of the intermediate nodes	**30**	**43**	**867**
Number of the output nodes	**64**	**50**	**417**
Number of the reduced nodes	**23**	**35**	**699**
Node reduction percentage (%)*	**77**	**81**	**81**

***(Number of the reduced nodes/Number of the intermediated nodes) x 100**

Fig. 2. Kernel structures of biological regulatory networks. (**A**) Kernels of the circadian network and integrin pathway with node reduction percentages of 67% and 94%, respectively. (**B**) Kernels of the networks of *E. coli*, yeast, and human with node reduction rates of 77%, 81%, and 81%, respectively. (**C**) Distributions of input, intermediate, and output nodes, and node reduction percentages of the *E. coli*, yeast, and human networks.

Structural characteristics of networks and kernels

Application of our algorithm to relatively small-sized networks, the circadian regulation network (*31*, *32*) in mammals (data S1) and a generalized integrin signaling pathway (unknown regulations were assumed to be activations) representing data from multiple species (*33*) (data S2), resulted in a node reduction percentage of 67% and 94%, respectively (Fig. 2A). The circadian kernel consisted of two negative feedback loops and one positive feedback loop, a structure consistent with the known core of circadian regulation (*34*). The considerable reduction that we achieved for the integrin network resulted from the following characteristics of the integrin pathway: It had only 8 negative edges (inhibitory regulations) out of 101 edges, and hence most feedforward loops in the pathway were of coherent type, and the pathway consisted of many long signaling cascades (the network diameter of the pathway, the maximum of the shortest path lengths between node pairs, was 14).

A signaling network with a high node reduction percentage contains numerous redundant nodes in terms that are not required to preserve input-output dynamics; thus, the node reduction rate can be considered as a measure of redundancy in signaling networks. To explore the amount of redundancy in signaling networks of three species, *E. coli*, *S. cerevisiae*, and *H. sapiens* (Fig. 2B), we compared the node reduction percentages for the kernels of the *E. coli*, yeast, and human signaling networks. The node reduction percentage for each network was ~80% (Fig. 2C), suggesting that these three signaling networks have a similar proportion of redundant intermediate signaling proteins. The amount of reduction in the number of nodes and edges increased as the proportion of intermediate nodes in the original network increased (Fig. 2C and fig. S3). For example, the human network with 1953 total nodes had the largest proportion of intermediate nodes (44%) and exhibited the greatest reduction in nodes and edges when the kernel was compared to the original network (fig. S3).

We examined the global topological properties of network density, clustering coefficient, network diameter, and characteristic path length between the original networks and their kernels (Fig. 3). The network densities and average clustering coefficients of the kernels were greater than those of the original networks (Fig. 3, A and B), which means that the nodes of the kernels were more densely connected, and neighborhood nodes of each node were more densely connected to each other. From the comparison of the network diameters and the characteristic path lengths (Fig. 3, C and D), we found that the small-world property of the kernels was stronger; that is, every node in the kernel was on average a smaller number of steps away from any other node in the kernel. For example, in the human network, the kernel nodes were 3.3 steps away from each other, whereas the average number of steps was 6.3 in the original network.

We also analyzed the local properties of the networks, such as the subnetwork structure and the properties of the most highly connected node, the "giant component." We compared the distribution ratios of three-node subnetworks (numbered 1 through 13) between the human network and its kernel (Fig. 3E). The subnetwork structure, which plays the role of a signal splitter (subnetwork ID1), was the most frequently occurring (50%) in the original network, whereas a signal integrator (subnetwork ID3) was dominant (47%) in the kernel. This implies that the human network includes a large number of signal splitters and many signaling pathways are connected by signal integrators. Indeed, the giant component of the original human signaling network included 85% of the total number of nodes (fig. S4).

The local properties of the *E. coli* and yeast networks were different from those of the human network. In these two networks, the signal splitter subnetwork (ID1) was the most frequently occurring three-node subnetwork structure in both the original networks and the kernels (fig. S5), suggesting that these two networks have signal-splitting subnetworks but that most pathways in the networks are isolated. Indeed, most of the pathways in the *E. coli* and yeast networks were short in length (Figs. 2B and 3D), and the networks did not contain extensively connected components (fig. S4). The different structural features related to subnetwork occurrence and node interconnectedness between the human

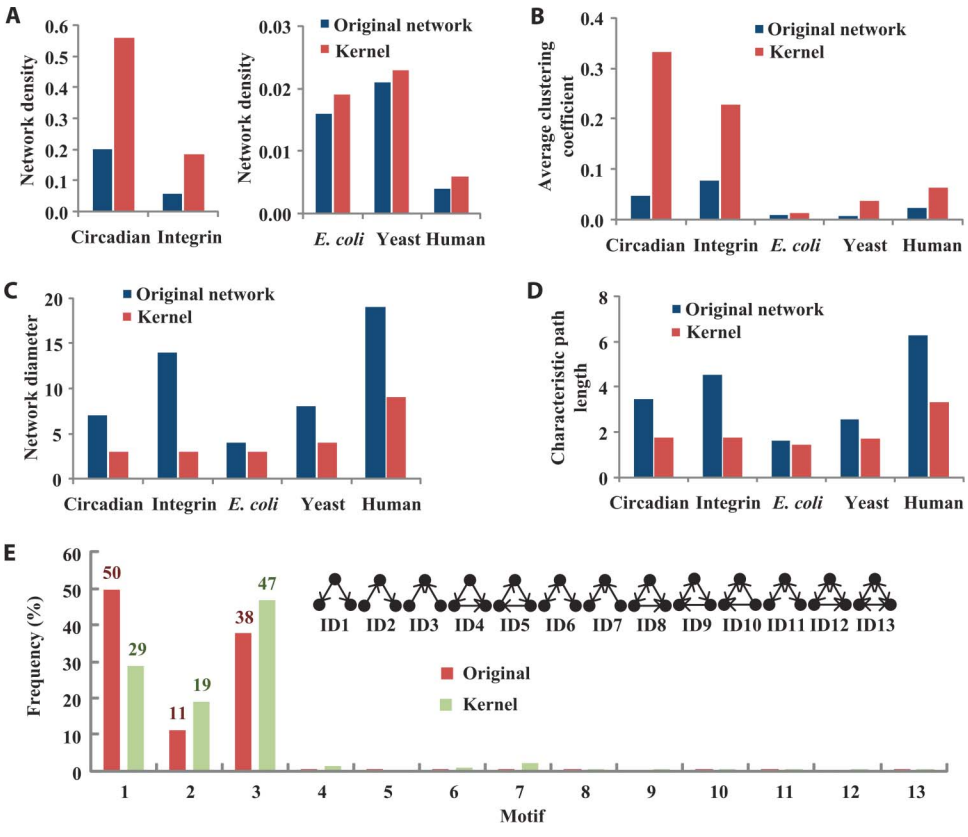


Fig. 3. Contrasting topological properties of original networks and kernels in the networks of *E. coli*, yeast, and human. (**A**) Network density, a measure of the density of edges in a network. (**B**) Average clustering coefficient, a measure of degree to which nodes in a network tend to cluster together. (**C**) Network diameter, the longest path connecting two nodes. (**D**) Characteristic path length, the average number of connections between two nodes. (**E**) Frequency distributions of three-node subnetworks in the human network and its kernel. The shapes of the subnetworks (ID1 through ID13) are shown at the top right.

network and the networks of *E. coli* and yeast may relate to the multifunctionality of kinases, which is reflected in the number of connections that they make. We found that the average indegree and outdegree of human kinases were significantly higher than the average indegree and outdegree of total nodes in both the original and the kernel networks (Fig. 4, A to D). In contrast, indegree and outdegree of the kinases relative to the total nodes in the original and kernel networks of the two single-cell species were not significantly different (Fig. 4). Moreover, this tendency is enforced in the kernels (Fig. 4, B and D, and fig. S6). These results imply that human kinases function to connect multiple signaling pathways.

Enrichment of essential genes, disease genes, and synthetic lethal genes in the kernel

Kernel nodes can be defined as those not deleted during the reduction process, which results in these nodes having similar or increased connectivity in the reduced network relative to the same nodes in the original network. It is possible that the kernel nodes play pivotal roles and that the non–kernel nodes have auxiliary roles in terms of biological processes. We investigated the enrichment of essential genes, disease genes, and synthetic lethal gene pairs in the sets of the kernel nodes and the non–kernel nodes for the human network. If the kernel nodes represented proteins with critical functions, then we would expect that the kernels would be enriched for nodes in each of these classes. We observed that 10% of the non–kernel nodes were essential genes, whereas 32% of the kernel nodes were essential genes (Fig. 5A) [essential genes were defined from Zhang and Lin (*35*); see Materials and Methods], and the difference is statistically significant ($P = 1.38 \times 10^{-21}$). In addition, we observed that most of the essential genes present in the original human network were included in the kernel (fig. S7). Essential genes also tended to be enriched in kernel networks of *E. coli* and yeast (fig. S8). Highly connected proteins in protein-protein interaction networks have a higher probability of being encoded by essential genes (*36*), and the original human network exhibited this property (fig. S9). To determine whether enrichment of essential genes in kernel nodes depended on node degree (the number of inputs and outputs), we considered only the nodes of small degree (<4) for the calculation of the ratio of essential genes (see fig. S10 for the degree distribution in the human network, and note that the number of nodes of degree <4 is about half of the total number of nodes). Even when only nodes with relatively few connections were considered, the kernel nodes were still enriched for essential genes (Fig. 5A). Thus, the kernel-identifying algorithm identified both essential genes represented by those in the kernel network and nonessential genes represented by nodes that were deleted by the network reduction.

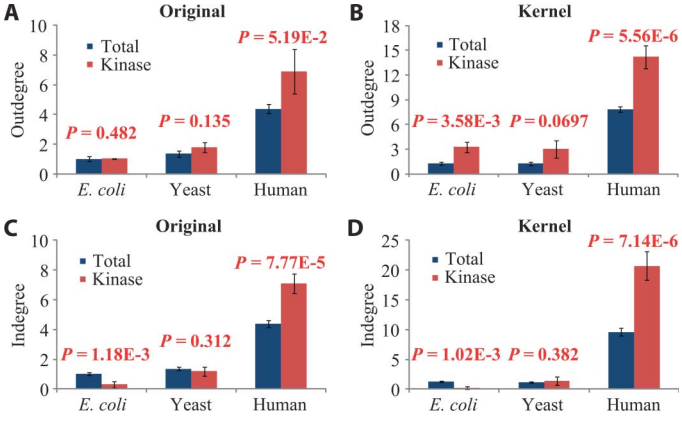


Fig. 4. Human kinases are more connected than those of *E. coli* or yeast. (A) Average outdegrees of the total nodes and the kinases in the original networks. (B) Average outdegrees of the total nodes and the kinases in the kernels. (C) Average indegrees of the total nodes and the kinases in the original networks. (D) Average indegrees of the total nodes and the kinases in the kernels. We extended the original networks by including phosphorylation interactions (edges) for which the regulation (activation or inhibition) was undefined. The error bars represent SE.

In the human network, we found a similar enrichment for disease-associated genes, which were defined on the basis of the Online Mendelian Inheritance in Man (OMIM) database (*37*) in the National Center for Biotechnology Information (NCBI). Most kernel nodes (95%) corresponded to Mendelian disease genes (Fig. 5B), and their enrichment in the kernel compared to the non–kernel nodes was statistically significant. As with the essential gene enrichment, we found that the enrichment in disease-associated genes was also not dependent on degree and that even limiting the analysis to nodes with a degree <4 showed a significant enrichment in disease-associated genes in the kernel nodes compared to the non–kernel nodes (Fig. 5B). Many genes are both essential and disease-related. When the classes are taken together, the results suggest that kernel nodes represent biologically important points in the network and often correspond to critical genes.

Synthetic lethality is considered to be closely related to network structure (*38*, *39*). We expected that the kernel nodes would be enriched in synthetic lethal gene pairs. Two genes are called a synthetic lethal gene pair if mutation of either alone is not lethal, but mutation of both leads to death or a significant decrease in the organism's fitness (*40*). We analyzed how many synthetic lethal gene pairs were included in the kernel of the human network. Synthetic lethal pairs of human genes were based on Conde-Pueyo *et al.* (*40*). As expected, most synthetic lethal pairs (93%) occurred between two kernel nodes, and we did not identify any synthetic lethal pairs in the set of non–kernel nodes (Fig. 5C). Our finding that kernels contained not only most essential genes and disease genes but also most synthetic lethal gene pairs suggests that kernel nodes are critical in terms of individual components and, because the synthetic lethality is closely related to network structure (*38*, *39*) and the kernel was obtained in consideration of network structure, in terms of network structure.

Tissue broadness and species broadness of kernel nodes

Because kernels are representative of biological networks, we speculate that kernel nodes may be ubiquitously expressed in various tissues and be conserved among diverse species. We investigated "tissue broadness" and "species broadness" for both kernel nodes and non–kernel nodes of the human network. The tissue broadness (*41*) of a gene is defined as the number of human tissues in which the gene is expressed (*42*), and species broadness as the number of species in which homologs of the gene exist (*43*) (see Materials and Methods). We found that both the tissue broadness (Fig. 5D) and the species broadness (Fig. 5E) of the kernel nodes were significantly larger than those of the non–kernel nodes.

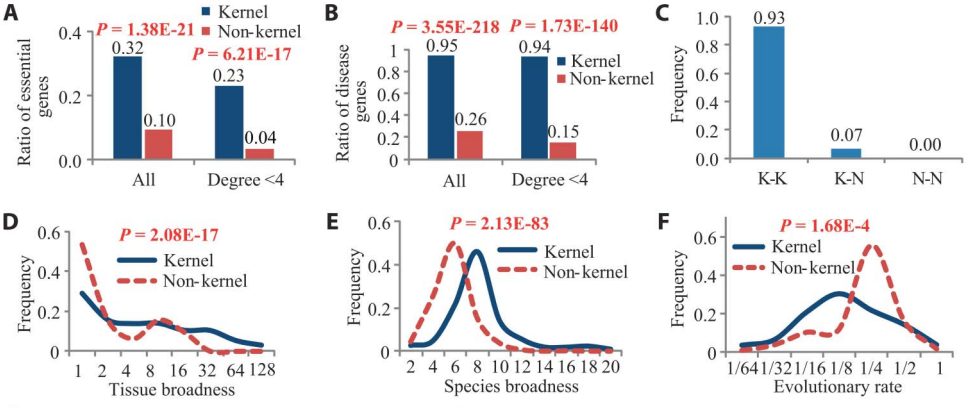
A high value for tissue broadness suggests that the gene is expressed ubiquitously and that the gene plays a common basic cellular function of various types of cells. We found that many kernel nodes were related to metabolic and developmental processes (table S2). Similarly, a high value for species broadness of a gene implies that the gene is evolutionarily conserved; hence, the kernel nodes may represent a conserved core of the network.

Evolutionary rates of kernel nodes and the relationship with function

Large values for tissue and species broadness of the kernel nodes imply that the gene sequences of the kernel nodes might have changed little during evolution. We explored the evolutionary rate (see Materials and Methods) of the kernel nodes with the hypothesis that conserved nodes would have lower evolutionary rates than non–kernel nodes and found that the kernel nodes had significantly lower evolutionary rates than did the non–kernel nodes (Fig. 5F), implying that the gene sequences of the kernel nodes are conserved during evolution.

From a Gene Ontology (GO) analysis, we observed that the functions assigned to the kernel nodes were different from those of the non–kernel nodes. The kernel nodes were mainly related to metabolic processes (48%, table S2) or to developmental processes (49%, table S2), whereas many of the non–kernel nodes were related to sensory perception (62%, table S3). We also observed a relationship between gene functions and evolutionary rates. The genes with relatively higher evolutionary rates were mainly related to immune processes and sensory perception, whereas those with lower evolutionary rates were related to developmental and metabolic processes (Fig. 5G). Because the genes related to metabolic processes play a pivotal role for cell survival (64.4% of the essential genes are related to metabolic processes with the enrichment of $P = 4.9 \times 10^{-29}$), these genes might have been evolutionarily stable. We noticed that the kernel nodes with high evolutionary rates were related to immune processes, whereas the non–kernel nodes with high evolutionary rates were related to sensory perception (Fig. 5G and tables S4 to S9). Because the kernel was determined from the network structure, this type of network reduction process can provide insight into gene functions.

Although genes associated with immune processes and sensory perception had relatively high evolutionary rates (Fig. 5G), these genes were associated with different parts of the network: Immune process–associated genes were enriched in the kernel nodes, whereas sensory perception–associated genes were enriched in the non–kernel nodes. We hypothesized that the genes related to immune processes were represented by nodes within elaborately and tightly regulated network substructures, such as feedback loops, and thus were not eliminated during network reduction. In comparison, we predicted that those genes related to sensory perception would not be represented by nodes in network substructures such as feedback loops. Indeed, the nodes representing the immune response genes were more enriched in feedback loops compared to the nodes representing the sensory perception genes (Fig. 5H). Genes with low evolutionary rate have been negatively selected (*44*) and genes with high evolutionary rate have been positively selected (*45*) during evolution. Hence, these results suggest that the functions of the kernel nodes might have been conserved by negative selection or the deleterious effects of mutations on organisms, whereas those of the non–kernel nodes might have evolved by positive selection or by beneficial effects of mutations.



G

Evolutionary rate (ER) range	Kernel	Non-kernel
ER > 1/4	Immune processes	Sensory perception
1/8 < ER ≤ 1/4	Developmental processes	Sensory perception
1/16 < ER ≤ 1/8	Metabolic processes	No significant GO term
ER ≤ 1/16	Metabolic processes, Signaling pathway	

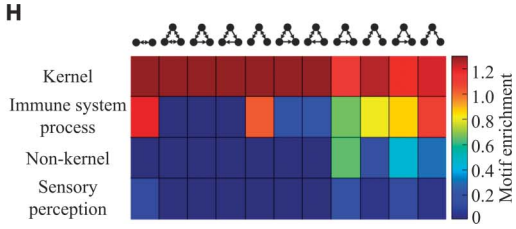


Fig. 5. Biological relevance of the kernel of the human network. (A) Ratios of essential genes of the original human network in the kernel and the non-kernel. Either all the nodes (left) or only the nodes with degree <4 (right) were evaluated and nodes corresponding to essential genes were retained preferentially in the kernel. Node calculations were performed as follows: nodes that represent essential genes in the kernel/total number of nodes in the kernel (blue); nodes that represent essential genes in the non-kernel/total number of nodes in the non-kernel (red); for the low-degree node calculation (the same as described but with nodes of degree <4). (B) Ratios of disease genes in the kernel compared to those not in the kernel were increased fourfold when all the nodes were considered (left) or sixfold when only the nodes with degree <4 (right) were considered. The ratios were calculated as in (A) with nodes that represent disease genes in the numerator. (C) Distribution of synthetic lethal gene pairs. K-K, K-N, and N-N denote the pairs composed of two kernel nodes, a kernel node and a non–kernel node, and two non–kernel nodes, respectively. The frequency represents the proportion of synthetic lethal pair occurring in each group. (D) Distributions of tissue broadness of kernel and non–kernel nodes. (E) Distributions of species broadness for kernel nodes and non–kernel nodes. (F) Distributions of evolutionary rates for the kernel nodes tend to be lower than those of non–kernel nodes. (G) Functional annotations of the four gene classes grouped by the ranges of evolutionary rate for the kernel and non–kernel genes. (H) Enrichment of 11 network motifs of four gene classes in the human network.

Network kernel and drug targets

Because most kernel nodes in the human signaling network can be mapped to diseases (fig. S7), we speculated that the kernel nodes might be related to drug targets. Hase *et al.* (*46*) showed that drug targets are enriched in the backbone network composed of middle-degree nodes (6 to 38 connections) in a protein-protein interaction network from Rual *et al.* (*47*). We compared the ratios of drug targets in the kernel and non–kernel nodes and found that drug targets were enriched in the kernel (Fig. 6, A and B), which is consistent with the previous work (*46*). Drug targets were identified on the basis of DrugBank (*48*). In addition, we examined the relationship between drug targets and node degree, and we found that nodes that were drug targets had middle degrees (Fig. 6C), which is also consistent with the previous work (see fig. S11 for the degree correlation

between the protein-protein interaction network and the human network). We observed that the neighborhood nodes of drug targets had low degrees (Fig. 6D), and drug targets had low closeness centrality (Fig. 6E), which is a measure of how a given node is close to all other nodes (*49*). These characteristics suggest that the nodes that are drug targets are middle-degree hubs in the kernel but are peripheral nodes such that their perturbation would locally affect the network. On the basis of these results, we suggest that analysis of the topological properties of the kernel may enable discovery of drug targets.

DISCUSSION

Genomic and other experimental techniques have enabled the discovery and study of large-scale biomolecular interaction networks, such as the map of human cancer signaling (*50*). However, the scale of biological networks has typically required the study and modeling of either small subnetworks for performing detailed parameterization or larger sets of nodes with limited opportunity to parameterize the interactions. One means of solving this problem is to condense a biological network into a smaller one that is equivalent in terms of its dynamical and topological aspects. We considered whether interaction networks have evolved from certain core structures and if such reduced networks could represent the dynamics of the source network. To address these questions, we introduced the concept of a kernel of a biological network, which we defined as the minimal essential network that preserves the input-output dynamics of the original network. We created an algorithm by which we systematically identified a kernel by considering the relationship between the network structure and its dynamics (Fig. 1). Because the proposed algorithm reduces signaling networks by considering the locally equivalent subnetworks instead of global equivalents, it is fast and can be applied to large-scale networks containing tens of thousands of nodes and millions of edges. Several studies have simplified complex networks by considering the static properties of the network topology (*8*–*19*). Compared to our approach, these alternatives do not preserve the dynamical properties of the input-output response profiles of the original network (fig. S12). Among the coarse-graining methods, hierarchical modularization methods (*20*, *51*) are most effectively applicable to various networks because they do not require any a priori information on the number or size of modules (that is, groups of clustered nodes) (*51*). However, coarse-graining methods still cannot reduce networks while preserving their dynamics because a module represented as a single node in the reduced network can actually contain many feedback or feedforward loops that entail complex dynamics.

For more detailed modeling of the behavior of individual genes, the parameters for each reduced node can be expanded while maintaining a simplified but fully representative network away from the area under study. With our algorithm, we identified the kernels of several networks ranging from *E. coli* and yeast to human and verified that the identified kernels preserved the fundamental input-output dynamics of the original networks. We found that the kernels comprise nodes representing essential genes, disease genes, drug targets, and synthetic lethal gene pairs (Figs. 5 and 6). Moreover, the kernels contained a high proportion of nodes representing genes with low evolutionary rates and genes that are ubiquitously expressed in various tissues and are present in many species. These results suggest that the kernels might be the backbones of biomolecular interaction networks, and interaction networks might have evolved on the basis of their kernels. We conclude that the analysis of a network kernel can provide new insights into the design principles of complex biomolecular interaction networks, identify potential drug targets, and facilitate modeling and parameterization of the resulting smaller-scale networks. This kernel identification network algorithm should only be applied to networks where neither stochasticity nor a time delay effect is dominant in determining the dynamical properties of input-output response profiles.

Genes corresponding to nodes in the reduced network may be most informative for phylogenetics or evolutionary studies, which may have implications for understanding the domestication of animals. As the number of organisms with well-defined signaling networks increases, it will be possible to investigate the effects of agricultural domestication of animals on both these genes represented by kernel nodes and non–kernel nodes. For example, reduced sensory perception (less flightiness and less need for detection of predators) and increased immune response (living in proximity and unnaturally large, genetically homogeneous groups) are signatures of domestication that are both intensively selected and probable preconditions for introduction to farms. Thus, one could ask if domesticated animals and their wild relatives (such as cows and aurochs) have the kernels representing gene networks that would produce these phenotypes compatible with domestication, whereas animals that have not been domesticated successfully (deer or all sub-Saharan large mammals) would have different kernels. One could also evaluate if the genes represented by the kernels between the domesticated and nondomesticated have these genes (and the kernel nodes) in domesticated animals changed under intense selection in a relatively short evolutionary period. Kernel nodes may also suggest points of

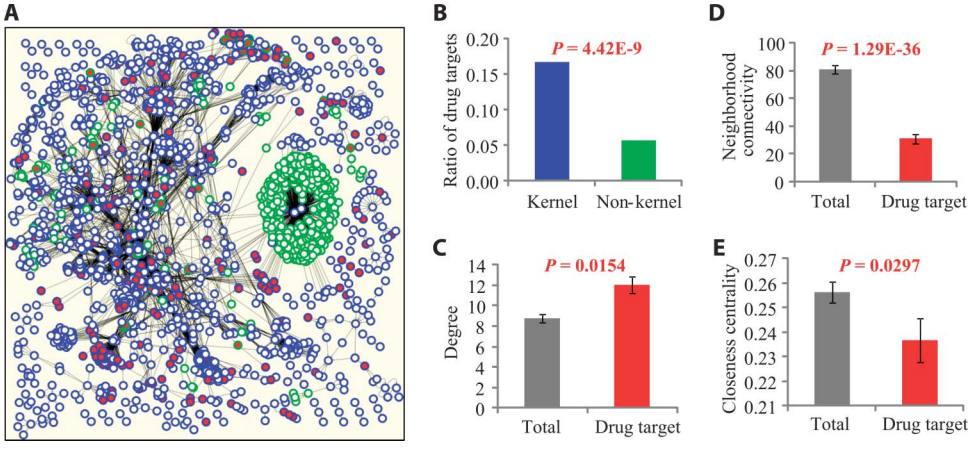


Fig. 6. The kernel is enriched in drug targets. (**A**) Distribution of drug targets in the human network. The blue bordered circles denote kernel nodes; the green bordered circles denote non–kernel nodes. Red (white) fill denotes drug targets, and white fill denotes non–drug targets. (**B**) Ratios of drug targets in the kernel and non-kernel. The ratio was calculated as (nodes that represent drug targets/total number of kernel nodes) or (nodes that represent drug targets/total number of non–kernel nodes). (**C**) Averages of the node degrees of total nodes and drug targets in the original network. (**D**) Averages of the neighborhood connectivities of total nodes and drug targets in the original network. (**E**) Averages of the closeness centralities of total nodes and drug targets in the original network. In (C) to (E), the error bars denote SEs.

genetic modification in agriculture because they lie on the critical path to generation of products and hence may show strong signatures of domestication or be future targets for modification.

The whole kernel can be modeled and then the behavior of a node in the original network can be elucidated from more detailed analyses of the kernel, and the kernel will show the modular organization of the original network as well as the critical input and output edges, which must be included in subnetwork models. We expect that the proposed kernel identification method can also be applied to facilitate the modeling and analysis of middle-scale signaling pathways, such as the epidermal growth factor receptor pathway (*52*–*54*), which already comprises a large number of signaling proteins under various states. Borisov *et al.* proposed a model reduction scheme in which unfeasible protein states are eliminated on the basis of a domain analysis (*55*). Our method reduced the number of signaling proteins to be modeled, as shown in the examples of the circadian network and the integrin pathway (Fig. 2A). The kernel identification method can be used to screen the key signaling components that dominate the dynamics of a given signaling pathway. This is particularly useful in the study of complex signaling networks, because the identification of such key signaling components is a fundamental step for any further analysis (*56*, *57*).

Although we considered a pulsatile stimulus in this study to examine the dynamical properties of the input-output response profiles of a cellular system because many cell signaling inputs can actually be approximated with this form of signal, other types of "biologically relevant" input stimulations, such as long-term constant inputs or oscillatory inputs, can also be approximated by controlling the parameter of the pulse signal (that is, the duration) or the combinations of pulse signals. Future studies will extend this method to determine its effectiveness in reproducing network dynamics in response to the other (even biologically irrelevant) input stimulation patterns.

MATERIALS AND METHODS

Classification of two- and three-node networks with respect to dynamics

We constructed mathematical models of two- and three-node networks with ordinary differential equations (see Supplementary Model Descriptions for details) and simulated them 1000 times with random parameter values in the interval [0, 1], where the stimulus was given by a pulse type (fig. S13). Next, we classified 1000 response curves into six types (fig. S14) for each model and represented the numbers of response curves in each of the six classes by a six-dimensional vector (tables S10 and S11). If a stimulus, such as a pulse—most cell signaling inputs can actually be approximated with this form of signal—is given to a cellular system, it can produce one of the following response profiles: a pulsatile response, a monotonic increasing or decreasing response, a sustained oscillatory response, or a damped oscillatory response. We considered these types of response profiles in our examination of the dynamical properties. We defined the dynamical distance $D(X,Y)$ between two models represented by $X = (x_1, x_2, \ldots, x_6)$ and $Y = (y_1, y_2, \ldots, y_6)$ by

$$D(X,Y) = \left| Y/|Y| - X/|X| \right| / \sqrt{2}$$

where $|\ |$ denotes the Euclidean norm

$$|X| = \sqrt{x_1^2 + x_2^2 + \ldots x_n^2} \quad \text{with} \quad X = (x_1, x_2, \ldots x_n)$$

On the basis of this distance, we applied the multidimensional scaling approach (*58*) to classify the two- or three-node networks with respect to their dynamics (Fig. 1B). We used a distance criterion $D < 0.001$ for the determination of the same network dynamics.

Kernel identification algorithm

Our algorithm can be applied to directed networks, such as signaling networks, gene transcription networks, and metabolic networks. For simplicity, we assume that each edge in the networks has only one of two regulation types (activation or inhibition). Hence, a biological network can be represented by a signed graph $G = (V,E)$, where V is a set of nodes and E is a set of edges with signs. Each edge can be represented by $e_{ij} = (v_i, v_j, \sigma_{ij})$, where v_i is a start node, v_j is an end node, and σ_{ij} is a sign (+1, 0, or −1) of the edge (referred to as the "signature"). $\sigma_{ij} = 0$ denotes that two nodes v_i and v_j are not connected by an edge. For each node v_j, we define the set of start nodes of the edges whose end node is v_j, by $VE(v_j) = \{v_i | (v_i, v_j, \sigma_{ij}) \in E\}$. Likewise, we define the set of end nodes of edges whose start node is v_j by $VS(v_j) = \{v_k | (v_j, v_k, \sigma_{jk}) \in E\}$. Let Σ be the set of signed graphs. Our kernel identification algorithm is represented by a map $F: \Sigma \rightarrow \Sigma$ satisfying the following four conditions [for a given $G = (V,E) \in \Sigma$, let $F(G) = (F(V), F(E))$]:

(i) F preserves node-edge relations. That is, for a given edge, $e_{ij} = (v_i, v_j, \sigma_{ij})$, $F(e_{ij}) = (F(v_i), F(v_j), F(\sigma_{ij}))$. Here, $F(\sigma_{ij}) = 0$ denotes an edge deletion.

(ii) F preserves input and output nodes.

(iii) Given v_j satisfying either indegree$(v_j) = 1$ or outdegree$(v_j) = 1$, consider the case that v_j is not involved in a self-feedback loop, an incoherent feedforward loop, or a two-node feedback loop. For any $v_i \in VE(v_j)$ and $v_k \in VS(v_j)$, if $(v_i, v_k, \sigma_{ij}\sigma_{jk}) \in E$, then $F(v_j) = F(v_k)$ and $F(\sigma_{ij}) = F(\sigma_{jk}) = 0$. Otherwise, $F(v_j) = F(v_k)$, $F(\sigma_{jk}) = 0$, and $F(\sigma_{ik}) = \sigma_{ij}\sigma_{jk}$. In this case, a new edge $(F(v_i), F(v_k), F(\sigma_{ik}))$ is contained in $F(E)$.

(iv) Consider the case that there is no node satisfying the conditions for the node-based reduction [step (iii)]. For each edge $e_{ij} = (v_i, v_j, \sigma_{ij}) \in E$, if e_{ij} is not contained in any incoherent feedforward loop, then for any $v_k \in VS(v_j)$ [$v_k \in VE(v_j)$, respectively], $F(v_j) = F(v_i)$, $F(\sigma_{ij}) = F(\sigma_{jk}) = 0$, and $F(\sigma_{ik}) = \sigma_{ij}\sigma_{jk}$ [$F(\sigma_{ki}) = \sigma_{ij}\sigma_{kj}$, respectively].

Our algorithm minimizes the size of $F(G)$ by reordering the nodes in G. We repeated steps (iii) and (iv) for 1000 times reordering of nodes and selected the minimal network for each G.

Essential genes, disease genes, synthetic lethal gene pairs, and drug targets

The essential gene lists for three species, *E. coli*, *S. cerevisiae*, and *H. sapiens*, were obtained from the database DEG (Database of Essential Genes, version 5.4) (*35*). The disease gene list was obtained from OMIM database (*37*) in the NCBI. This list contains 14,388 disease genes, 1536 of which are contained in the original human network. The list of synthetic lethal gene pairs for human was obtained from iHSLN (inferred human SL genes) (*40*). The drug target list was obtained from the DrugBank database (*48*). This list contains 1330 proteins that are drug targets, 275 of which are contained in the original human network.

Tissue broadness

The tissue broadness (*41*) of a gene is defined as the number of tissues in which the gene is the upper outlier, meaning that the mRNA abundance is higher than the sum of the upper quartile and 1.5 times the interquartile range (*59*). We calculated the tissue broadness information using mRNA expression data in 79 human tissues (*42*).

Species broadness

We defined the species broadness of a gene as the number of species in which homologs of the gene exist. The homolog information of 20 species (table S12) was extracted from the HomoloGene database (*43*) in the NCBI.

Evolutionary rate

The evolutionary rates were defined by the ratios of the nonsynonymous substitution rates (dN) and the synonymous substitution rates (dS) for homologous gene pairs in human and mouse and they were obtained from the Human PAML Browser (*60*).

GO analysis

GO (*61*) analysis was performed with the functional annotation tool in DAVID (*62*). We first divided the 1493 human kernel genes into four groups and the 460 deleted genes into three groups on the basis of the ranges of evolutionary rate (see Fig. 5G for the ranges). We then retrieved the GO terms significantly related with each gene group [Benjamini score (*63*) <0.05], using the functional annotation tool applied to the 1953 genes of the human network as a background set. We selected the child GO terms related to parent terms (metabolic process, developmental process, sensory perception, immune process, and signal transduction) in the GO hierarchy.

Motif enrichment analysis

We identified network motifs using MAVisto (*64*) without considering both edge labels and vertex labels. For reliable statistics, 1000 random networks were generated, and the three-node subgraphs with P values <0.05 were considered as network motifs. We defined the enrichment of a motif for each gene class (Fig. 5H) as the ratio of the number of the genes related to the motif contained in the gene class to the expected number of genes related to the motif for the size of the gene class.

Statistical analysis

We performed one-sided two-sample χ^2 tests to evaluate the statistical abundance of the essential genes (Fig. 5A), disease genes (Fig. 5B), and drug targets (Fig. 6B) in the kernels. For the tissue broadness (Fig. 5D), species broadness (Fig. 5E), evolutionary rates (Fig. 5F), degree (Fig. 6C), neighborhood connectivity (Fig. 6D), and closeness centrality (Fig. 6E), the one-sided two-sample t test was applied.

Availability of the software

We have implemented the proposed kernel identification algorithm as software. It is available from http://sbie.kaist.ac.kr/software and as part of the Supplementary Materials.

SUPPLEMENTARY MATERIALS

www.sciencesignaling.org/cgi/content/full/4/175/ra35/DC1

Model Descriptions

Fig. S1. The multidimensional scaling map for classification of responses of the nonlinear (Hill-type) models of two- and three-node networks.

Fig. S2. The flow diagram illustrating the kernel identification algorithm.

Fig. S3. Ratios of kernel to original in terms of nodes and edges for the signaling networks of *E. coli*, yeast, and human.

Fig. S4. Relative size of the giant component, which is the component with the most connections, in the original three networks.

Fig. S5. The frequency distributions of three-node subnetworks in the signaling networks of *E. coli* and yeast compared with the distributions of these subnetworks in their kernels.

Fig. S6. Average indegrees and outdegrees of kinases in the original networks and kernels for the networks of *E. coli*, yeast, and human.

Fig. S7. The frequency of essential genes and disease genes in the kernel nodes and non–kernel nodes.

Fig. S8. The ratio of essential genes contained in the kernel and non–kernel nodes of the networks of *E. coli* and yeast.

Fig. S9. The ratio of essential genes represented in the set of nodes of each degree in the human network.

Fig. S10. Degree distribution and cumulative frequency distribution of degrees in the human network.

Fig. S11. The degree in the human protein-protein interaction network versus the degree in the human signaling network.

Fig. S12. Comparison of the input-output dynamics of the original network and the reduced network after applying five different network-reduction approaches.

Fig. S13. The stimulus pattern used for the simulation of two- and three-node network models.

Fig. S14. Six representative response patterns used for classification of two- and three-node networks.

Table S1. Response coherency between the original signaling network and the corresponding kernel.

Table S2. GO terms related to genes represented by nodes in the kernels (kernel genes).

Table S3. GO terms related to genes that were excluded from the kernel, but were represented by nodes in the original network (non-kernel).

Table S4. GO terms related to the kernel genes that had evolutionary rates larger than 0.25.

Table S5. GO terms related to the kernel genes that had evolutionary rates between 0.125 and 0.25.

Table S6. GO terms related to the kernel genes that had evolutionary rates between 0.0625 and 0.125.

Table S7. GO terms related to the kernel genes that had evolutionary rates less than 0.0625.

Table S8. GO terms related to non–kernel genes that had evolutionary rates less than 0.25.

Table S9. GO terms related to the non–kernel genes that had evolutionary rates between 0.125 and 0.25.

Table S10. Simulation results for linear models of 18 network structures.

Table S11. Simulation results for Hill-type models of 18 network structures.

Table S12. The list of 20 species examined in the HomoloGene database.

Data S1. The circadian regulatory network data where the first, second, and third columns denote regulator, relation, and target, respectively. [Filename: Circadian_regulatory_network.txt]

Data S2. The integrin signaling pathway data where the first, second, and third columns denote regulator, relation, and target, respectively. [Filename: Integrin_signaling_pathway.txt]

Data S3. The *E. coli* signaling network data where the first, second, and third columns denote regulator, relation, and target, respectively. [Filename: Ecoli_network.txt]

Data S4. The yeast signaling network data where the first, second, and third columns denote regulator, relation, and target, respectively. [Filename: Yeast_network.txt]

Data S5. The human signaling network data where the first, second, and third columns denote regulator, relation, and target, respectively. [Filename: Human_network.txt]

Software. The software of the proposed kernel identification algorithm for MS-DOS–type operating systems. [Filename: kernelfinder.exe]

References

REFERENCES AND NOTES

1. E. V. Koonin, Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.* **1**, 127–136 (2003).
2. J. I. Glass, N. Assad-Garcia, N. Alperovich, S. Yooseph, M. R. Lewis, M. Maruf, C. A. Hutchison III, H. O. Smith, J. C. Venter, Essential genes of a minimal bacterium. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 425–430 (2006).
3. E. V. Koonin, How many genes can make a cell: The minimal-gene-set concept. *Annu. Rev. Genomics Hum. Genet.* **1**, 99–116 (2000).
4. A. R. Mushegian, E. V. Koonin, A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 10268–10273 (1996).
5. C. Pál, B. Papp, M. J. Lercher, P. Csermely, S. G. Oliver, L. D. Hurst, Chance and necessity in the evolution of minimal metabolic networks. *Nature* **440**, 667–670 (2006).
6. C. A. Hutchison III, S. N. Peterson, S. R. Gill, R. T. Cline, O. White, C. M. Fraser, H. O. Smith, J. C. Venter, Global transposon mutagenesis and a minimal mycoplasma genome. *Science* **286**, 2165–2169 (1999).
7. K. Kobayashi, S. D. Ehrlich, A. Albertini, G. Amati, K. K. Andersen, M. Arnaud, K. Asai, S. Ashikaga, S. Aymerich, P. Bessieres, F. Boland, S. C. Brignell, S. Bron, K. Bunai, J. Chapuis, L. C. Christiansen, A. Danchin, M. Débarbouille, E. Dervyn, E. Deuerling, K. Devine, S. K. Devine, O. Dreesen, J. Errington, S. Fillinger, S. J. Foster, Y. Fujita, A. Galizzi, R. Gardan, C. Eschevins, T. Fukushima, K. Haga, C. R. Harwood, M. Hecker, D. Hosoya, M. F. Hullo, H. Kakeshita, D. Karamata, Y. Kasahara, F. Kawamura, K. Koga, P. Koski, R. Kuwana, D. Imamura, M. Ishimaru, S. Ishikawa, I. Ishio, D. Le Coq, A. Masson, C. Mauël, R. Meima, R. P. Mellado, A. Moir, S. Moriya, E. Nagakawa, H. Nanamiya, S. Nakai, P. Nygaard, M. Ogura, T. Ohanan, M. O'Reilly, M. O'Rourke, Z. Pragai, H. M. Pooley, G. Rapoport, J. P. Rawlins, L. A. Rivas, C. Rivolta, A. Sadaie, Y. Sadaie, M. Sarvas, T. Sato, H. H. Saxild, E. Scanlan, W. Schumann, J. F. M. L. Seegers, J. Sekiguchi, A. Sekowska, S. J. Seror, M. Simon, P. Stragier, R. Studer, H. Takamatsu, T. Tanaka, M. Takeuchi, H. B. Thomaides, V. Vagner, J. M. van Dijl, K. Watabe, A. Wipat, H. Yamamoto, M. Yamamoto, Y. Yamamoto, K. Yamane, K. Yata, K. Yoshida, H. Yoshikawa,

U. Zuber, N. Ogasawara, Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 4678–4683 (2003).
8. A. Arenas, J. Duch, A. Fernández, S. Gómez, Size reduction of complex networks preserving modularity. *New J. Phys.* **9**, 176 (2007).
9. M. Girvan, M. E. J. Newman, Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 7821–7826 (2002).
10. J. B. Glattfelder, S. Battiston, Backbone of complex networks of corporations: The flow of control. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **80**, 036104 (2009).
11. K. I. Goh, G. Salvi, B. Kahng, D. Kim, Skeleton and fractal scaling in complex networks. *Phys. Rev. Lett.* **96**, 018701 (2006).
12. S. Itzkovitz, R. Levitt, N. Kashtan, R. Milo, M. Itzkovitz, U. Alon, Coarse-graining and self-dissimilarity of complex networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **71**, 016127 (2005).
13. D. H. Kim, J. D. Noh, H. Jeong, Scale-free trees: The skeletons of complex networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **70**, 046126 (2004).
14. K. Komurov, M. H. Gunes, M. A. White, Fine-scale dissection of functional protein network organization by statistical network analysis. *PLoS One* **4**, e6017 (2009).
15. N. Masuda, Y. Kawamura, H. Kori, Impact of hierarchical modular structure on ranking of individual nodes in directed networks. *New J. Phys.* **11**, 113002 (2009).
16. M. Müller-Linow, C. C. Hilgetag, M. T. Hütt, Organization of excitable dynamics in hierarchical biological networks. *PLoS Comput. Biol.* **4**, e1000190 (2008).
17. M. A. Serrano, M. Boguñá, A. Vespignani, Extracting the multiscale backbone of complex weighted networks. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 6483–6488 (2009).
18. C. Song, S. Havlin, H. A. Makse, Self-similarity of complex networks. *Nature* **433**, 392–395 (2005).
19. Y. Xiao, B. D. MacArthur, H. Wang, M. Xiong, W. Wang, Network quotients: Structural skeletons of complex systems. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **78**, 046102 (2008).
20. I. A. Kovács, R. Palotai, M. S. Szalay, P. Csermely, Community landscapes: An integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics. *PLoS One* **5**, e12528 (2010).
21. F. Chung, The heat kernel as the pagerank of a graph. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 19735–19740 (2007).
22. J. R. Kim, Y. Yoon, K. H. Cho, Coupled feedback loops form dynamic motifs of cellular networks. *Biophys. J.* **94**, 359–365 (2008).
23. S. Y. Shin, O. Rath, S. M. Choo, F. Fee, B. McFerran, W. Kolch, K. H. Cho, Positive- and negative-feedback regulations coordinate the dynamic behavior of the Ras-Raf-MEK-ERK signal transduction pathway. *J. Cell Sci.* **122**, 425–435 (2009).
24. S. Y. Shin, O. Rath, A. Zebisch, S. M. Choo, W. Kolch, K. H. Cho, Functional roles of multiple feedback loops in extracellular signal-regulated kinase and Wnt signaling pathways that regulate epithelial-mesenchymal transition. *Cancer Res.* **70**, 6715–6724 (2010).
25. T. H. Kim, J. Kim, P. Heslop-Harrison, K. H. Cho, Evolutionary design principles and functional characteristics based on kingdom-specific network motifs. *Bioinformatics* **27**, 245–251 (2011).
26. D. Kim, Y. K. Kwon, K. H. Cho, The biphasic behavior of incoherent feed-forward loops in biomolecular regulatory networks. *Bioessays* **30**, 1204–1211 (2008).
27. S. Y. Shin, H. W. Yang, J. R. Kim, W. Do Heo, K. H. Cho, A hidden incoherent switch regulates RCAN1 in the calcineurin-NFAT signaling network. *J. Cell Sci.* **124**, 82–90 (2011).
28. M. Kanehisa, S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
29. Y. K. Kwon, K. H. Cho, Boolean dynamics of biological networks with multiple coupled feedback loops. *Biophys. J.* **92**, 2975–2981 (2007).
30. Y. K. Kwon, K. H. Cho, Quantitative analysis of robustness and fragility in biological networks based on feedback dynamics. *Bioinformatics* **24**, 987–994 (2008).
31. J. C. Leloup, A. Goldbeter, Toward a detailed computational model for the mammalian circadian clock. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 7051–7056 (2003).
32. T. K. Sato, S. Panda, L. J. Miraglia, T. M. Reyes, R. D. Rudic, P. McNamara, K. A. Naik, G. A. Fitzgerald, S. A. Kay, J. B. Hogenesch, A functional genomics strategy reveals Rora as a component of the mammalian circadian clock. *Neuron* **43**, 527–537 (2004).
33. J. T. Parsons, K. H. Martin, J. K. Slack, S. A. Boerner, C. C. Martin, Integrin signaling pathway. *Sci. Signal.* (Connections Map in the Database of Cell Signaling, as seen 23 Aug 2005), http://stke.sciencemag.org/cgi/cm/stkecm;CMP_6880.
34. J. R. Kim, W. S. Bae, Y. Yoon, K. H. Cho, Topological difference of core regulatory networks induces different entrainment characteristics of plant and animal circadian clocks. *Biophys. J.* **93**, L1–L3 (2007).
35. R. Zhang, Y. Lin, DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res.* **37**, D455–D458 (2009).
36. K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, A. L. Barabási, The human disease network. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 8685–8690 (2007).
37. J. Amberger, C. A. Bocchini, A. F. Scott, A. Hamosh, McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.* **37**, D793–D796 (2009).
38. S. L. Ooi, X. Pan, B. D. Peyser, P. Ye, P. B. Meluh, D. S. Yuan, R. A. Irizarry, J. S. Bader, F. A. Spencer, J. D. Boeke, Global synthetic-lethality analysis and yeast functional profiling. *Trends Genet.* **22**, 56–63 (2006).
39. S. L. Wong, L. V. Zhang, A. H. Y. Tong, Z. Li, D. S. Goldberg, O. D. King, G. Lesage, M. Vidal, B. Andrews, H. Bussey, C. Boone, F. P. Roth, Combining biological networks to predict genetic interactions. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 15682–15687 (2004).
40. N. Conde-Pueyo, A. Munteanu, R. V. Solé, C. Rodríguez-Caso, Human synthetic lethal inference as potential anti-cancer target gene detection. *BMC Syst. Biol.* **3**, 116 (2009).
41. X. Gu, Z. Su, Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 2779–2784 (2007).
42. A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, J. B. Hogenesch, A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 6062–6067 (2004).
43. E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, J. Ye, Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **38**, D5–D16 (2010).
44. A. Muto, S. Osawa, The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 166–169 (1987).
45. R. Nielsen, Z. Yang, Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929–936 (1998).
46. T. Hase, H. Tanaka, Y. Suzuki, S. Nakagawa, H. Kitano, Structure of protein interaction networks and their implications on drug design. *PLoS Comput. Biol.* **5**, e1000550 (2009).
47. J. F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, M. Vidal, Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
48. D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **36**, D901–D906 (2008).
49. A. Özgür, T. Vu, G. Erkan, D. R. Radev, Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics* **24**, i277–i285 (2008).
50. Q. Cui, Y. Ma, M. Jaramillo, H. Bari, A. Awan, S. Yang, S. Zhang, L. Liu, M. Lu, M. O'Connor-McCourt, E. O. Purisima, E. Wang, A map of human cancer signaling. *Mol. Syst. Biol.* **3**, 152 (2007).
51. S. Fortunato, Community detection in graphs. *Phys. Rep.* **486**, 75–174 (2010).
52. M. R. Birtwistle, M. Hatakeyama, N. Yumoto, B. A. Ogunnaike, J. B. Hoek, B. N. Kholodenko, Ligand-dependent responses of the ErbB signaling network: Experimental and modeling analyses. *Mol. Syst. Biol.* **3**, 144 (2007).
53. W. W. Chen, B. Schoeberl, P. J. Jasper, M. Niepel, U. B. Nielsen, D. A. Lauffenburger, P. K. Sorger, Input-output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. *Mol. Syst. Biol.* **5**, 239 (2009).
54. R. Samaga, J. Saez-Rodriguez, L. G. Alexopoulos, P. K. Sorger, S. Klamt, The logic of EGFR/ErbB signaling: Theoretical properties and analysis of high-throughput data. *PLoS Comput. Biol.* **5**, e1000438 (2009).
55. N. M. Borisov, A. S. Chistopolsky, J. R. Faeder, B. N. Kholodenko, Domain-oriented reduction of rule-based network models. *IET Syst. Biol.* **2**, 342–351 (2008).
56. B. Schoeberl, E. A. Pace, J. B. Fitzgerald, B. D. Harms, L. Xu, L. Nie, B. Linggi, A. Kalra, V. Paragas, R. Bukhalid, V. Grantcharova, N. Kohli, K. A. West, M. Leszczyniecka, M. J. Feldhaus, A. J. Kudla, U. B. Nielsen, Therapeutically targeting ErbB3: A key node in ligand-induced activation of the ErbB receptor–PI3K axis. *Sci. Signal.* **2**, ra31 (2009).
57. S. Li, S. M. Assmann, R. Albert, Predicting essential components of signal transduction networks: A dynamic model of guard cell abscisic acid signaling. *PLoS Biol.* **4**, e312 (2006).
58. T. F. Cox, M. A. A. Cox, *Multidimensional Scaling* (Chapman and Hall/CRC, Boca Raton, FL, 2000), p. 328.
59. B. R. Rosner, *Fundamentals of Biostatistics* (Duxbury, Pacific Grove, CA, 2000).
60. G. C. Nickel, D. Tefft, M. D. Adams, Human PAML browser: A database of positive selection on human genes using phylogenetic methods. *Nucleic Acids Res.* **36**, D800–D808 (2008).

61. M. A. Harris, J. I. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Collins, K. Eilbeck, S. Lewis, C. Mungall, J. Richter, G. M. Rubin, S. Q. Shu, J. A. Blake, C. J. Bult, A. D. Diehl, M. E. Dolan, H. J. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, G. Binkley, J. M. Cherry, K. R. Christie, M. C. Costanzo, Q. Dong, S. R. Engel, D. G. Fisk, J. E. Hirschman, B. C. Hitz, E. L. Hong, C. Lane, S. Miyasato, R. Nash, A. Sethuraman, M. Skrzypek, C. L. Theesfeld, S. Weng, D. Botstein, K. Dolinski, R. Oughtred, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, N. Mulder, R. Chisholm, P. Fey, P. Gaudet, W. Kibbe, K. Pilcher, C. A. Bastiani, R. Kishore, E. M. Schwarz, P. Sternberg, K. Van Auken, M. Gwinn, L. Hannick, J. Wortman, M. Aslett, M. Berriman, V. Wood, S. Bromberg, C. Foote, H. Jacob, D. Pasko, V. Petri, D. Reilly, K. Seiler, M. Shimoyama, J. Smith, S. Twigger, P. Jaiswal, T. Seigfried, C. Collmer, D. Howe, M. Westerfield, Gene Ontology Consortium, The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.* **34**, D322–D326 (2006).
62. D. W. Huang, B. T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, Y. Guo, R. Stephens, M. W. Baseler, H. C. Lane, R. A. Lempicki, DAVID Bioinformatics Resources: Expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.* **35**, W169–W175 (2007).
63. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
64. F. Schreiber, H. Schwöbbermeyer, MAVisto: A tool for the exploration of network motifs. *Bioinformatics* **21**, 3572–3574 (2005).
65. **Acknowledgments:** We thank S. Baek, D. Kim, D. Lee, H. Byrne, and A. Fletcher for valuable comments on the manuscript. **Funding:** This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea Government, the Ministry of Education, Science and Technology (MEST) (2009-0086964, 2010-0017662, and 2011-0002145). It was also supported by World Class University program through the NRF of Korea funded by the MEST (R32-2008-000-10218-0). **Author contributions:** K.-H.C. designed the research; J.-R.K., J.K., and H.-Y.L. performed simulations; J.-R.K., J.K., H.-Y.L., and K.-H.C. analyzed the data; Y.-K.K., J.-R.K., and K.-H.C. implemented the algorithm; J.-R.K., J.K., P.H.-H., and K.-H.C. wrote the manuscript. **Competing interests:** P.H.-H. is the Director of BioAstral Limited. The other authors declare that they have no competing interests.

Submitted 26 July 2010
Accepted 10 May 2011
Final Publication 31 May 2011
10.1126/scisignal.2001390
Citation: J.-R. Kim, J. Kim, Y.-K. Kwon, H.-Y. Lee, P. Heslop-Harrison, K.-H. Cho, Reduction of complex signaling networks to a representative kernel. *Sci. Signal.* **4**, ra35 (2011).

Reduction of Complex Signaling Networks to a Representative Kernel

Jeong-Rae Kim, Junil Kim, Yung-Keun Kwon, Hwang-Yeol Lee, Pat Heslop-Harrison and Kwang-Hyun Cho (May 31, 2011)

Science Signaling **4** (175), ra35. [doi: 10.1126/scisignal.2001390]