SYSTEMS BIOLOGY

A Directed Protein Interaction Network for Investigating Intracellular Signal Transduction

Arunachalam Vinayagam,[1*†] Ulrich Stelzl,[1,2*‡] Raphaele Foulle,[1]
Stephanie Plassmann,[1] Martina Zenkner,[1] Jan Timm,[1] Heike E. Assmus,[3]
Miguel A. Andrade-Navarro,[1] Erich E. Wanker[1‡]

Cellular signal transduction is a complex process involving protein-protein interactions (PPIs) that transmit information. For example, signals from the plasma membrane may be transduced to transcription factors to regulate gene expression. To obtain a global view of cellular signaling and to predict potential signal modulators, we searched for protein interaction partners of more than 450 signaling-related proteins by means of automated yeast two-hybrid interaction mating. The resulting PPI network connected 1126 proteins through 2626 PPIs. After expansion of this interaction map with publicly available PPI data, we generated a directed network resembling the signal transduction flow between proteins with a naïve Bayesian classifier. We exploited information on the shortest PPI paths from membrane receptors to transcription factors to predict input and output relationships between interacting proteins. Integration of directed PPI with time-resolved protein phosphorylation data revealed network structures that dynamically conveyed information from the activated epidermal growth factor and extracellular signal–regulated kinase (EGF/ERK) signaling cascade to directly associated proteins and more distant proteins in the network. From the model network, we predicted 18 previously unknown modulators of EGF/ERK signaling, which we validated in mammalian cell-based assays. This generic experimental and computational approach provides a framework for elucidating causal connections between signaling proteins and facilitates the identification of proteins that modulate the flow of information in signaling networks.

INTRODUCTION

Studies of cellular signal transduction processes indicate that classical signaling pathways are integrated parts of larger molecular interaction networks (*1*–*3*). For instance, receptor tyrosine kinase (RTK) signaling pathways can signal through extracellular signal–regulated kinases (ERKs), and signaling by these receptors affects differentiation, proliferation, survival, and migration. Mutations in RTKs or in downstream proteins, such as Ras and Raf, lead to abnormal RTK signaling through the ERK pathway, which contributes to the development of disease (*4*–*6*). Although the core elements of the RTK to ERK (RTK-ERK) signaling cascade have been well characterized, functional genomic and proteomic studies have shown that signal propagation through this pathway is more complex than previously thought. In a genome-wide RNA interference (RNAi) study, more than 1000 annotated genes, including those encoding the core RTK-ERK pathway components, were identified as influencing ERK phosphorylation (*7*). A proteomics study demonstrated that >600 proteins are dynamically phosphorylated after stimulation of HeLa cells with epidermal growth factor (EGF) (*8*), a ligand for the RTK EGF receptor (EGFR). This proteomics analysis indicated that in addition to the known EGF and ERK pathway components, such as Ras, the mitogen-activated protein kinase kinase kinase (MAPKKK) Raf, the MAPKK MEK (mitogen-activated or extracellular signal–regulated protein kinase kinase), and the MAPK ERK1, a large number of other proteins, including transcription factors, cytoskeletal proteins, and ubiquitin ligases, are phosphorylated in response to EGF. These studies reinforce the fact that to comprehensively understand signal transduction processes in mammalian cells, it is not sufficient to simply investigate the core components of a signaling cascade. Instead, a network view capturing the dynamics of signaling events is necessary to elucidate more completely the molecular alterations of activated cells (*1*, *3*).

Large protein-protein interaction (PPI) networks can be generated by systematic yeast two-hybrid (Y2H) studies (*9*) or protein complex isolation and mass spectrometry approaches (*10*). These networks can advance our understanding of how proteins interact to form large molecular assemblies and cellular machines and how the cell responds to changes in the intracellular or extracellular environment (*2*). However, two problems hamper the use of PPI networks to study signaling networks. PPI networks for signaling proteins are incomplete (*11*), and the links in PPI networks generated from Y2H, complex isolation, or mass spectrometry lack directionality of signal flow and sign (activation or inhibition) (*12*). Such information is necessary to obtain a complete understanding of the dynamics of cellular signaling processes and would strengthen the predictions that can be made from PPI network analyses about key steps in signaling processes and their alterations in disease (*13*–*15*).

Here, we investigated intracellular signaling networks with a combined experimental and computational approach. Using a set of ~450 signaling-related proteins as baits, we identified ~2500 PPIs through repeated Y2H interaction screens. By combining this experimentally derived information with publicly available interaction data, we constructed a more comprehensive PPI network. Because this PPI network lacked information about the directionality of signal flow, we developed a bioinformatic strategy for edge direction prediction and used this to assign the potential direction of signal

[1] AG Neuroproteomics and Computational Biology and Data Mining Group, Max Delbrück Centrum for Molecular Medicine, Robert-Rössle-Strasse 10, D-13125 Berlin-Buch, Germany. [2] Otto-Warburg Laboratory, Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, D-14195 Berlin, Germany. [3] AG Systems Biology and Bioinformatics, Department of Computer Science, University of Rostock, Ulmenstrasse 69, D-18051 Rostock, Germany.
*These authors contributed equally to this work.
†Present address: Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA.
‡To whom correspondence should be addressed. E-mail: stelzl@molgen.mpg.de (U.S.); erich.w@mdc-berlin.de (E.E.W.)

flow along protein interactions from membrane receptors to transcription factors. Whereas previous studies combined heterogeneous information, such as gene expression data, to identify potential signaling pathways in PPI data, here, we predicted directed signaling networks solely from the PPI data. We used the directed PPI network model to analyze the dynamics of protein phosphorylation during EGF signaling, as well as to identify proteins that modulate ERK phosphorylation in mammalian cells.

RESULTS

Generating a Y2H PPI network for signaling proteins

We used a combination of experimental and computational strategies to create a PPI network for cellular signaling. We selected 473 human full-length open reading frames (ORFs) (*16*) on the basis of annotation as members of Kyoto Encyclopedia of Genes and Genomes (KEGG) signal transduction pathways or as direct interaction partners of such proteins (Table 1 and table S1). The ORFs were then used as baits for an interaction mating screen against an array of ~7800 MAT**α** yeast prey strains (*17*). Interactions were identified by spotting yeast colonies onto selective plates followed by β-galactosidase membrane filter assays (*18*). Screens were repeated two to six times to detect weak or transient PPIs, such as interactions between kinases and their substrates that are often missed in a single Y2H screen (*19*). We systematically examined ~10 million potential interactions and generated a Y2H-based PPI data set, linking 1126 human proteins through 2626 unique interactions (Fig. 1A, table S2, and fig. S1). More than half of the PPIs (1457) were identified two or more times in the successive Y2H screens, indicating that the interaction data were reliable.

We established benchmarks for Y2H data quality assessment that included measures for domain-domain interactions, common neighborhood (shared interaction partners), clustering (co-occurrence of interacting proteins in network clusters), biological process similarity, localization similarity, and coexpression (Fig. 1B and table S2). Different measures for benchmarking are needed because PPI data are heterogeneous; for example, different interactions exhibit unique biophysical properties. Therefore, different benchmarks address different subsets of interactions (*20*). For each measure, we calculated relative precision values (the fraction of interactions that matched those already known; see Materials and Methods for details) with PPIs from the Human Protein Reference Database [HPRD (*21*)]. As a positive PPI data set, we selected 10,000 interactions that were either reported in two independent publications or identified with two different experimental techniques. As a negative set, we selected 10,000 random protein pairs for which we did not find any evidence for interaction in the literature. The precision values for the Y2H PPIs identified in this study (between 0.56 and 0.76) are similar to Y2H interactions collected from various small- and large-scale studies in HPRD [13,798 PPIs from 2625 publications (*21*)] (Fig. 1B), indicating that high-quality interaction data were produced by interaction mating. This result is in agreement with a previous study that used empirical measures to show that interactions detected with our screening approach have a precision of ~80% when tested with an independent PPI assay (*19*).

Predicting edge directions from PPI data

Y2H PPI screens, even when repeated several times, detect only about 10 to 20% of all possible biophysical interactions (*19*). To perform a comprehensive bioinformatic analysis, we extended the Y2H-based PPI data with publicly available interaction information. We combined interactions from 10 experimental data sets, each containing more than 80 binary PPIs (table S3), with our Y2H data to create the human PPI network 1 (HPPI1), connecting 9832 proteins through 39,641 interactions.

Although HPPI1 contains information about key components involved in cellular signaling pathways (fig. S2), it does not provide information about the direction of signal flow. Therefore, we developed a naïve Bayesian learning strategy (*22*, *23*) to add directionality to individual PPIs (Fig. 2A and fig. S3). Because a major component of signal flow in cellular signaling cascades is mediated by PPIs, we hypothesized that the shortest paths between receptors and transcription factors could be used to predict the direction of information flow (*24*, *25*). Using human KEGG regulatory pathways, we independently tested this hypothesis and learned that the shortest path connections (SPCs) recall the correct edges from a KEGG PPI network with high frequency at various thresholds (fig. S4).

We computed all 637,099 SPCs between 554 plasma membrane–associated receptors and 1150 transcription factors in the HPPI1 network with the direction of signal flow set as from activated membrane receptors to transcription factors. On average, each interaction was contained in about 20 different SPCs. Because some SPCs between certain families of receptors and transcription factors were strongly overrepresented (fig. S5), we also partitioned the data into subsets by grouping membrane receptors and transcription factors according to 23 and 47 known protein families, respectively. We then constructed a total of 1081 "grouped SPCs," each group representing the union of all SPCs between closely related membrane receptor to transcription factor signaling paths. Using a naïve Bayesian classifier, we derived the directionality of each interaction from the population of shortest paths (SPCs and grouped SPCs) in which they were contained.

Table 1. Bait selection. The 473 baits are individually listed in table S1 with details about their classification. Pathways are as defined by KEGG.

Bait class	Baits screened	Baits with interactions	Number of interactions	Average link per bait	Literature overlap	Previously unknown interactions (%)
Signal transduction pathways	257	139	1376	9.9	91	93.4
MAPK	77	43	394	9.2	29	92.6
Insulin	44	19	193	10.2	10	94.8
Apoptosis	13	5	44	8.8	12	72.7
Jak_STAT	10	5	40	8.0	1	97.5
Wnt	8	5	107	21.4	1	99.1
Disease pathway	27	18	89	4.9	11	87.6
Metabolic pathway	33	24	379	15.8	37	90.2
Other cellular process	118	82	801	9.8	58	92.8
Unknown biological process	38	23	180	7.8	15	91.7

The learning algorithm predicted the causal relationships between interacting proteins (edge direction) from eight features on the basis of the shortest PPI path connections and topological network properties (table S4). We trained the classifier by using 828 interactions with a defined direction assignment from KEGG pathways (*26*).

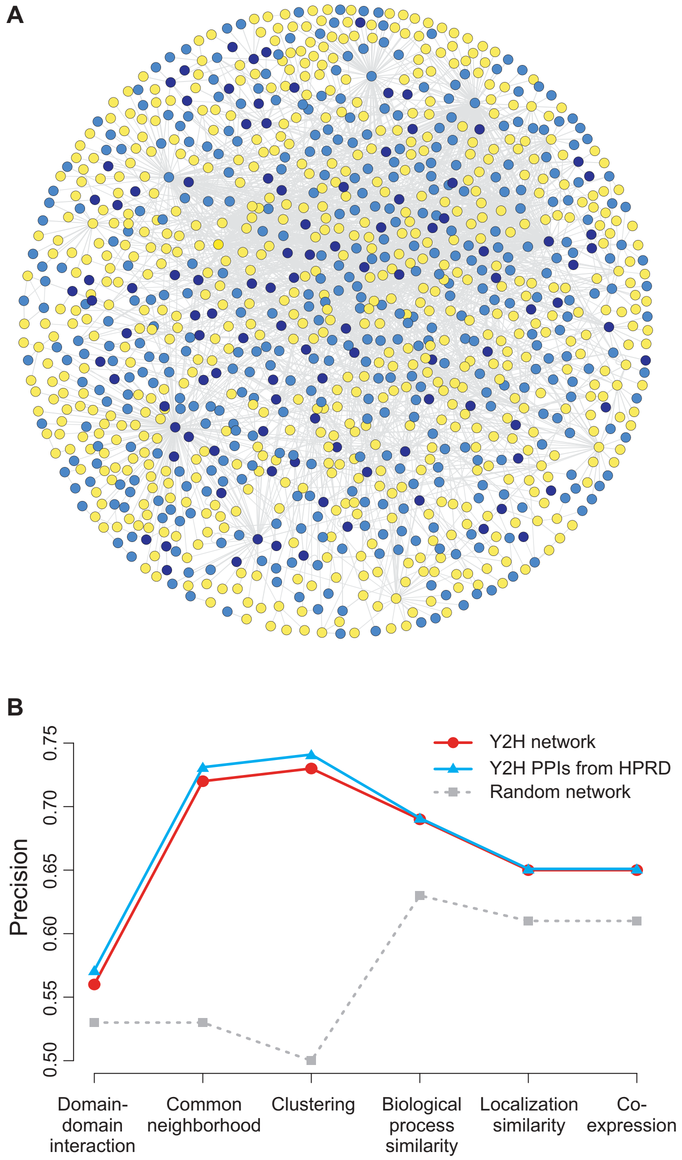

Fig. 1. A Y2H PPI network for signaling proteins. (A) Network view of the PPI data. The network contains 1126 human proteins linked by 2626 unique interactions; 94% of these interactions have not been identified previously. MAPK pathway members are shown as dark blue nodes, annotated members of other signaling pathways are light blue, and proteins that had not been annotated as members of signaling pathways are yellow nodes [drawn with Cytoscape (*68*)]. (B) Confidence profiles for PPI data sets. Six different measures were used to estimate the relative precision of the Y2H data in comparison to literature-curated Y2H PPIs from HPRD. To obtain relative precision values, we benchmarked measures against a set of 10,000 PPIs in HPRD with either two papers reporting an interaction or two experimental methods detecting a PPI.

We evaluated the predictive performance of the classifier on the basis of its receiver operating characteristic (ROC) curve, which shows sensitivity as a function of the false-positive rate (1 − specificity) (Fig. 2B). With an area under the ROC curve (AUC) of 0.73, the result of a 10-fold cross-validation indicates that the classifier exhibited good performance. We obtained similar performances with other classifying algorithms (table S5), showing that the naïve Bayesian classifier was as effective as alternative methods in predicting edge direction. Thus, our studies indicate that potential causal relationships between signaling proteins can be predicted with high precision and recall (Fig. 2C), suggesting that the naïve Bayesian classifier approach reveals biologically meaningful edge directions.

We applied the trained classifier for predicting edge directions of interactions to the HPPI1 data set, using a threshold that results in 70% precision and 69% recall for direction assignment (Fig. 2C, dotted line). We generated a filtered network, termed HPPI2, which contained 32,706 directed PPIs connecting 6339 human proteins, representing potential input and output relationships between interacting proteins that follow the potential information flow from activated membrane receptors to transcription factors (table S6). Because the proteins could not be connected to known membrane receptors and transcription factors, 6935 interactions (17.5%) in the HPPI1 data set remained without direction assignment. Comparison of the properties of the HPPI1 and HPPI2 networks showed that both exhibited known topological features of biological networks, and these characteristics did not differ substantially (fig. S6).

Computational validation of edge directions

To validate the predictive power of the Bayesian classifier, we tested whether the classifier showed similar performance with an independent data set containing 475 directed PPIs from 51 annotated signal transduction pathways from the Database of Cell Signaling (*27*). AUC analysis showed that the results obtained with independent direction information from the Database of Cell Signaling (*27*) and the KEGG (*26*) data were similar (Fig. 2B), confirming that this computational approach revealed meaningful edge directions.

We examined whether the HPPI2 network with directed PPIs contained network motifs characteristic of biological information-processing networks (*28*). We applied the Mfinder algorithm to find recurring interaction patterns (*29*). Particular three-node and four-node motifs, such as the feed-forward loop or the biparallel motif, were observed more frequently (*Z* scores >10) in the HPPI2 network than in randomized networks (fig. S7). Such motifs were previously also found in other signal transduction and directed information-processing networks (*30*), indicating that the HPPI2 network with inferred interaction directions shows similar design principles.

We performed a full triad significance profile analysis (*31*) by calculating the statistical significance of all 13 possible triad motifs in HPPI2 in comparison to randomized networks with the same size and connectivity properties (*29*). We found that three of the triad motifs were overrepresented and that five were underrepresented in the HPPI2 network (Fig. 2D). This pattern of both enriched and depleted local structures corresponds to triad profiles that structurally define a distinct group of information-processing networks, including signal transduction networks (*31*), thus providing independent validation of our computational network modeling strategy.

Analysis of EGF-induced phosphorylation dynamics

The HPPI2 network defines 1183 input and output relationships for 733 proteins that are directly linked to the 28 core components of the

EGF/ERK signaling pathway (Fig. 3A). To analyze the potential chain of events during EGF-stimulated signal propagation, we integrated the PPI data with a time-resolved protein phosphorylation data set from a large-scale, mass spectrometry–based proteomics study (*8*). We mapped 327 of 591 proteins that are dynamically phosphorylated or dephosphorylated upon EGF treatment (EGF-responsive phosphoproteins) in mammalian cells (*8*) onto the HPPI2 network. We determined whether the proteins with altered phosphorylation were output nodes that receive signals from the 28 core proteins of the EGF/ERK pathway (edge directions pointing away from the core pathway) or were input nodes that signal to the core pathway (edge directions pointing to the core pathway proteins). We found that the number of proteins with altered EGF-induced phosphorylation was significantly enriched in the set of output nodes but not in the set of input nodes (Table 2), suggesting that information flows preferentially along the predicted edge directions from the core pathway proteins to their neighbors in the PPI network. We also analyzed the frequency of appearance of proteins with altered phosphorylation (*8*) in first-, second-, or third-degree neighbors of core pathway proteins. In contrast to third-degree output neighbors (Table 2), the frequency of EGF-responsive phosphoproteins among first- and second-degree output neighbors was significantly higher than expected by chance ($P <$ 0.05), supporting the observation that information flows from core pathway proteins through the output edges also to more distant proteins in the network.

We classified the interacting proteins according to their annotation as KEGG (*26*) signaling molecules (for example, kinases and transcription factors) and examined enrichment of EGF-responsive phosphoproteins in these classes. We found that known signaling molecules of other pathways were generally enriched in EGF-responsive phosphoproteins in both the input and the output groups. Kinases with altered phosphorylation were enriched among first- and second-degree output neighbors, and transcription factor phosphoproteins were enriched only in the second-degree output

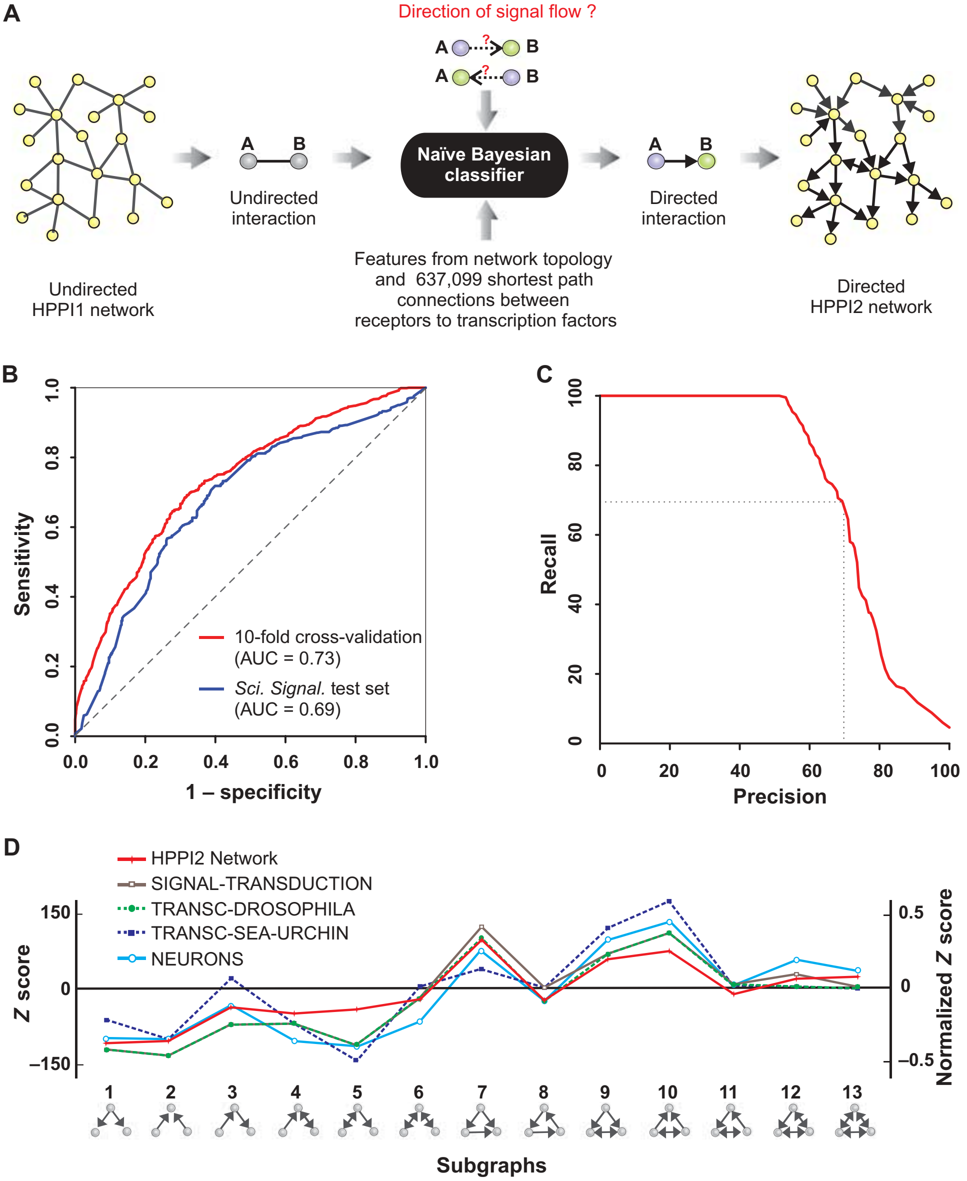

Fig. 2. Predicting the potential directions of signal flow in PPI networks. (**A**) Inferring edge directions from PPI data. For each interaction in the undirected PPI network (HPPI1), a naïve Bayesian classifier was used to predict the edge direction from topological network properties as well as shortest PPI paths connecting membrane receptors and transcription factors. An activated signaling network (HPPI2) was assembled from all interactions that had a direction assigned. (**B**) ROC analysis as an indication of the performance of the naïve Bayesian classifier. The red ROC curve represents results from 10-fold cross-validation. The blue ROC curve is calculated for an independent set of directed signaling interactions from the Database of Cell Signaling (*27*). The discontinuous line represents random performance in the ROC analysis. (**C**) Precision-recall curve as an estimation of the performance of the naïve Bayesian classifier. The optimal cutoff on the score resulted in a 70% precision and 69% recall and was chosen to construct the HPPI2 network (dotted line). (**D**) Three-node subnetwork profiles of the HPPI2, the signal-transduction interactions in mammalian cells from STKE (SIGNAL-TRANSDUCTION), transcription networks that guide development in fruit fly (TRANSC-DROSOPHILA), endomesoderm development in sea urchin (TRANSC-SEA-URCHIN), and synaptic connections between neurons in *Caenorhabditis elegans* (NEURONS). The *Z* scores of each triad from HPPI2 were compared against the normalized *Z* scores of information-processing networks [reference triad significance profiles are taken from (*31*)].

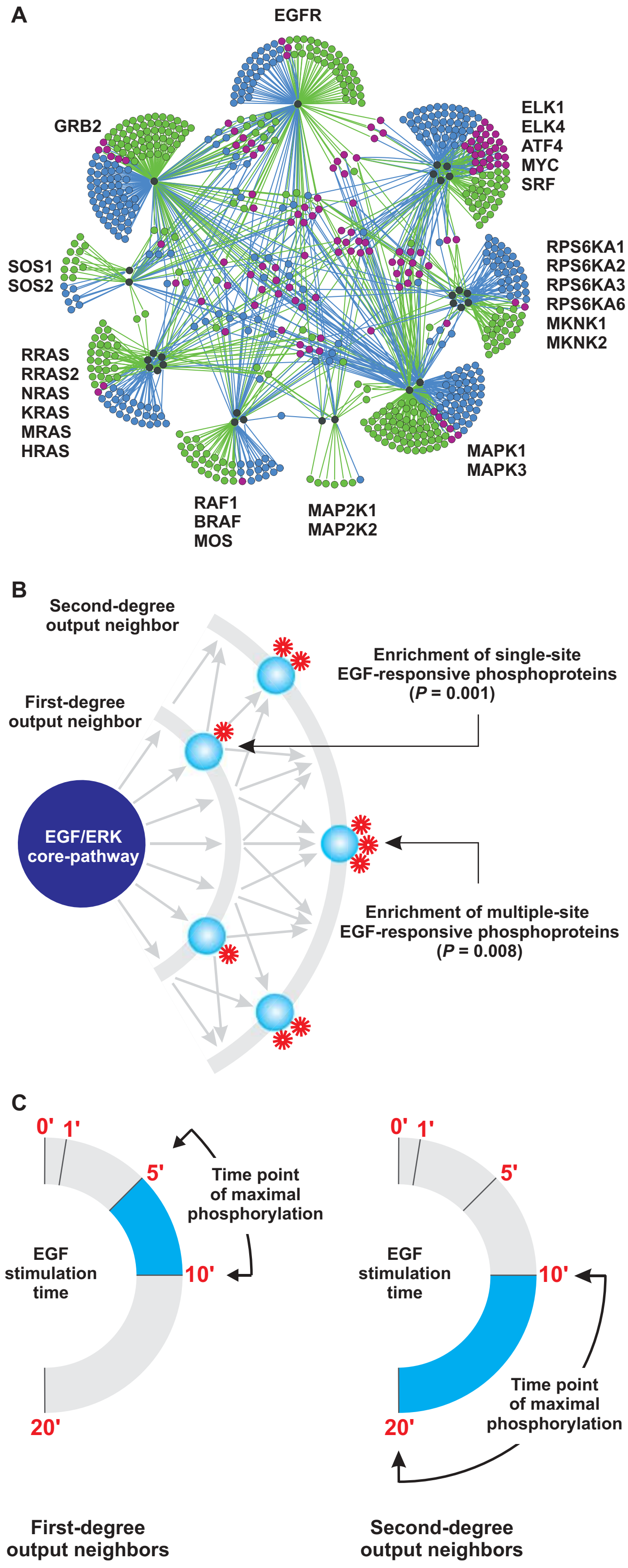
A
EGFR
ELK1
ELK4
ATF4
MYC
SRF
GRB2
SOS1
SOS2
RPS6KA1
RPS6KA2
RPS6KA3
RPS6KA6
MKNK1
MKNK2
RRAS
RRAS2
NRAS
KRAS
MRAS
HRAS
MAPK1
MAPK3
RAF1
BRAF
MOS
MAP2K1
MAP2K2
B
Second-degree output neighbor
First-degree output neighbor
Enrichment of single-site EGF-responsive phosphoproteins (*P* = 0.001)
EGF/ERK core-pathway
Enrichment of multiple-site EGF-responsive phosphoproteins (*P* = 0.008)
C
0'
1'
5'
10'
20'
Time point of maximal phosphorylation
EGF stimulation time
0'
1'
5'
10'
20'
EGF stimulation time
Time point of maximal phosphorylation
First-degree output neighbors
Second-degree output neighbors

Fig. 3. Integration of directed PPIs with dynamic protein phosphorylation data. (A) A network view of core pathway EGF/ERK proteins with 733 direct-interacting partners. Dark gray nodes represent the 28 known EGF/ERK core pathway proteins (labeled with their Entrez Official Gene Symbols) from EGFR downstream in a counterclockwise arrangement. Blue and green nodes correspond to potential input and output nodes, respectively, that are linked to the core pathway. Proteins that are both input and output nodes are shown in purple. (B) Schematic representation of the EGF-responsive phosphoproteins in data sets of first- and second-degree core pathway neighbors. Arrows indicate that the number of output relationships increases with the distance to the core pathway. (C) Schematic representation showing the relationship among the kinetics of EGF-responsive phosphorylation and the distance from the core EGF/ERK core pathway proteins.

nodes (table S7). This suggests that activation of the core pathway influences other pathways and various cellular processes through phosphorylation of a subset of direct (first degree), as well as indirect (second degree), core pathway neighbors.

We used the network model to analyze the number of EGF-responsive phosphorylation sites in direct and indirect neighbors of the core pathway proteins. Proteins phosphorylated at a single site were frequently found among first-degree core pathway neighbors, whereas proteins phosphorylated at more than one site were significantly enriched among second-degree output neighbors (Fig. 3B and table S8). Together with the observation that first- and second-degree core pathway neighbors are connected with a larger number of output links than core pathway proteins are connected with first-degree neighbors, this suggests that core pathway proteins initially activate a relatively small number of directly linked kinases, which then activate a larger number of proteins that are indirectly linked to the core pathway proteins. In this analysis, we can differentiate between first-order and second-order phosphorylation events, supporting the observation that second-site phosphorylation occurs frequently at more distant proteins receiving multiple inputs from the activated core pathway (*32*, *33*).

To investigate the dynamics of EGF-mediated signal propagation, we analyzed time-resolved protein phosphorylation data (*8*) by annotating the proteins according to the highest degree of phosphorylation after 1, 5, 10, or 20 min of EGF stimulation and then determining their frequency of appearance among first-, second-, and third-degree core pathway neighbors. We found that proteins phosphorylated 5 and 10 min after EGF stimulation were significantly enriched among first-degree core pathway output neighbors, whereas proteins phosphorylated after 10 and 20 min were detected significantly more often among second-degree output neighbors (Fig. 3C and table S9). Thus, this correlation confirms that direct neighbors of the activated pathway are phosphorylated faster after EGF stimulation than indirect neighbors.

Notably, such an effect was not observed when the phosphorylation of upstream and downstream core pathway interacting proteins was analyzed in a time-dependent manner. We investigated whether phosphorylation of network proteins correlated with the activation of core pathway members after EGF stimulation. The 28 core pathway proteins were divided into four groups representing upstream to downstream members of the pathway: (i) receptor and adaptor proteins; (ii) Ras, Raf, and MEK; (iii) ERK, Mnk, and Rsk; and (iv) transcription factors. The frequency of phosphorylation of direct and indirect neighbors of proteins belonging to these groups was analyzed in a time-dependent manner with the EGF-responsive phosphoprotein data. We found that neighbors of upstream core pathway proteins were not phosphorylated significantly faster upon EGF stimulation than were neighbors of downstream proteins (fig. S8 and table S10), indicating that transmission of information through the core pathway is not resolvable with the available proteomics data covering time points of 1, 5, 10, and 20 min after EGF stimulation (*5*, *8*).

Predicting potential modulators of phosphorylated ERK from the directed PPI network

Using the input- and output-node specifications in the HPPI2 network, we investigated whether potential modulators of EGF/ERK signaling can be predicted. In a genome-wide RNAi screen, more than 1000 *Drosophila* proteins were identified as modulators of phosphorylated ERK (pERK) (*7*). We mapped 337 from the set of 606 human orthologs of those modulators onto the HPPI2 network and used these as source nodes in a simple flow model (*34*, *35*) to predict potential previously unknown modulators downstream of known pERK modulators in the directed HPPI2 network (Fig. 4A; see Materials and Methods). With this approach, we prioritized 50 proteins (table S11) mostly on the basis of our Y2H PPI data set and subsequently tested their activity in cell-based EGF/ERK signaling assays.

To investigate whether the selected proteins influenced EGF/ERK signaling under conditions where the core pathway is activated with EGF, we transiently transfected human embryonic kidney (HEK) 293 cells with constructs encoding the potential modulator proteins, treated the cells with EGF, and, after 10 min, quantified pERK in cell extracts with a standardized enzyme-linked immunosorbent assay (ELISA) (*36*). We found that 11 of the 50 selected proteins significantly reduced EGF-mediated ERK phosphorylation in HEK293 cells (Fig. 4B and table S11), whereas 6 proteins, similar to constitutively activated MEK1, increased ERK phosphorylation (Fig. 4C). Some of the proteins exhibiting a strong inhibitory effect, such as GADD45A and MAP3K7IP1, had not been previously linked to EGF/ERK signaling [candidate proteins are named according to National Center for Biotechnology Information (NCBI) Entrez Official Gene Symbol].

We assayed whether transient overproduction of the selected target proteins stimulated ERK phosphorylation in the absence of EGF. We used constitutively activated MEK1 to promote ERK phosphorylation in HEK293 cells as a positive control, and we observed an increase in pERK when the proteins PRDX4, XRCC6, NDUFS6, MAP4K2, PEA15, PRKAR1A, or ZAK were overproduced (Fig. 4D). PRKAR1A and ZAK also increased pERK abundance in the presence of EGF, whereas XRCC6, NDUFS6, and PEA15 showed an inhibitory effect in the presence of EGF. In agreement with this observation, it has been reported that PEA15 preferentially interacts with and sequesters pERK in the cytoplasm of mammalian cells (*37*). The effects of a subset of the previously unknown modulators of ERK phosphorylation were confirmed by immunoblotting with antibodies recognizing either total ERK1 and ERK2 or pERK1 and pERK2 (Fig. 4E), supporting the results obtained with ELISA assays.

Selecting 50 proteins at random from either the list of proteins in the directed network, the starting baits, the interacting baits, or the interacting preys would on average yield 2.6 (5.3% of the selected proteins), 4.3 (8.6%), 4.2 (8.4%), or 3.1 (6.2%) of known pERK modulators, respectively. Thus, the detection of 18 of 50 (36%) pERK modifiers with our cell-based assays indicates that network-based strategies can be useful for predicting modulators of cell signaling proteins.

Our cell-based assays revealed that overproduction of MAPK6, an atypical chordate-specific MAPK of largely unknown function (*38*, *39*), significantly reduced ERK phosphorylation after EGF activation (Fig. 4B), suggesting that MAPK6 is linked to EGF/ERK signaling in mammalian cells. Further network analysis revealed 151 interaction partners for MAPK6, a large fraction of which (24 proteins) was also identified as potential modulators of ERK signaling (Fig. 4F). We found that six MAPK6 interaction partners influenced ERK phosphorylation when overproduced, whereas another study showed that eight proteins influence the signaling pathway when their concentrations were reduced with small interfering RNA (siRNA) (*7*). Collectively, our data and the previous data (summarized in table S12) provide evidence that MAPK6 and its interacting partners modulate EGF/ERK signaling in mammalian cells.

DISCUSSION

We present a human PPI data set of ~2600 interactions, which was generated by Y2H matrix screening of 450 signaling-related bait proteins against ~7800 prey proteins. Our interaction data are largely complementary to a Y2H PPI network for MAPKs and associated proteins (*40*), indicating that PPI data for signaling proteins are still incomplete, and additional studies with different genetic as well as biochemical methods are required to obtain a more comprehensive picture of protein associations involved in signal transduction. However, our investigations also indicate that the reliability of the Y2H PPIs identified in this study is high, suggesting that they are a valuable resource for further detailed functional experiments as well as predictive network analyses (*2*, *9*).

For benchmarking, we used six independent features that were previously applied to assign confidence scores to PPIs (*20*, *41*). More than 90% of the interactions scored positive for at least one of the six features, such as common network neighborhood, localization similarity, or coexpression (table S2). The benchmarking of our interaction data set in comparison to a filtered literature-curated data set showed that the Y2H PPIs identified

Table 2. EGF/ERK signal flow in the HPPI2 network. Proteins in the EGF-responsive phosphoprotein data set were mapped to first-, second-, and third-degree neighbors of the EGF/ERK core pathway. The proteins were further grouped into input or output neighbors on the basis of the edge direction relative to the core pathway. For each set of neighbors, the total number of nodes (neighborhood size) and number of mapped proteins were computed. We generated 1000 random sets with the same number of nodes as the respective neighborhood size. Enrichment was measured by comparing the data sets of associated proteins with the 1000 equal-size random sets of simulated interacting proteins ($P < 0.05$). The mean, SD, and *P* values were calculated on the basis of the overlap within each random set.

Data set	Degree of separation	Neighbor size	Input neighbors		*P*	Output neighbors		Random mean ± SD	*P*
			Phospho-proteins	Random mean ± SD		Neighbor size	Phospho-proteins		
EGF-responsive	First	424	26	22 ± 4.3	0.21	436	39	23 ± 4.4	**<0.001**
phosphoproteins	Second	1875	102	96 ± 7.9	0.25	1963	121	102 ± 8.3	**0.016**
(327 proteins)	Third	2254	123	116 ± 8.2	0.221	1622	90	84 ± 7.6	0.274

in this study are of high precision. This is in good agreement with previous work, where we used independent experimental measures to estimate the precision of Y2H PPI data (*19*, *42*).

To address whether PPI networks are useful for investigating information flow in cellular signal transduction pathways (*1*, *3*), we extended the Y2H PPI data set with available PPI information and computationally inferred edge directions following the potential signal flow from membrane receptors to transcription factors. Previous studies have inferred pathways—and in part directionality—from PPI interaction networks by integrating interaction information with different types of data such as expression profiles or functional annotations (*35*, *43*–*49*). Here, we used only network features that were obtained from SPCs between membrane receptors and transcription factors in a naïve Bayesian learning approach to globally predict PPI directionality. To define the directionality between protein pairs, we extracted information from more than 600,000 SPCs. In the learning phase, the directionality of each interaction was derived statistically from a population of shortest paths in which they were contained. For example, even though EGFR interacts directly with ERK under certain conditions (*21*, *50*), we also predict directions for interactions connecting EGFR and ERK through longer paths. Our directed PPI network (HPPI2) contains both

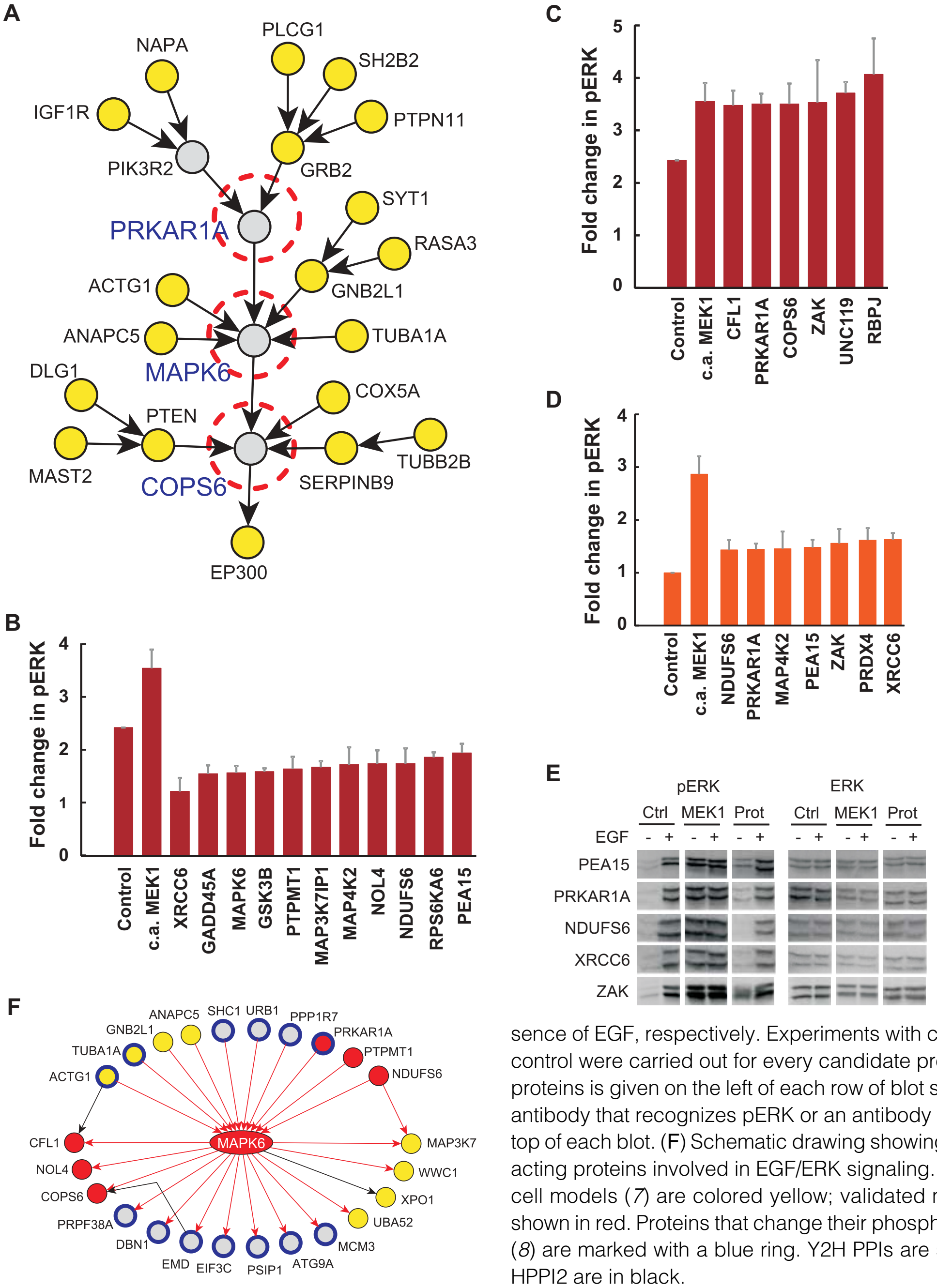

Fig. 4. Identification of EGF/ERK signaling modulators with the directed HPPI2 network. (**A**) Part of the directed network schematically showing a simple signal flow analysis including four proteins (gray) predicted as pERK modulators according to the number of signaling paths that originate from known pERK modulators (yellow), which were defined as orthologs of pERK-regulating proteins identified by RNAi screening in a *Drosophila* cell model (*7*). Proteins outlined with dashed circles represent candidate modulators. (**B**) Proteins that suppressed EGF-stimulated pERK abundance when overproduced by transient transfection in HEK293 cells. In (B) to (D), the selected modulators are labeled with their NCBI Entrez Official Gene Symbol. The indicated proteins were transiently produced in HEK293 cells. Endogenous pERK was compared to control transfections. Fold change of pERK is shown for proteins with at least a 1.4-fold effect in this standardized ELISA assay. SDs from triplicate experiments are shown. The results of all 50 candidate proteins are presented in table S11. (**C**) Proteins that enhanced EGF-stimulated pERK abundance when overproduced by transient transfection in HEK293 cells. (**D**) Proteins that increased pERK abundance in the absence of EGF. (**E**) Western blot analysis showing changes of endogenous ERK phosphorylation in response to transient transfection of PEA15, PRKAR1A, NUDFS6, XRCC6, and ZAK. Columns: Ctrl indicates transfection with empty control plasmid, MEK1 indicates transfection of constitutively active (c.a.) MEK1, and Prot indicates transfection of candidate proteins; +, − indicate the presence or absence of EGF, respectively. Experiments with constitutively active MEK1 and empty plasmid control were carried out for every candidate protein in parallel. The identity of the candidate proteins is given on the left of each row of blot sections. The blots were probed with either an antibody that recognizes pERK or an antibody that recognizes total ERK as indicated as the top of each blot. (**F**) Schematic drawing showing the relationship among MAPK6 and its interacting proteins involved in EGF/ERK signaling. Potential ERK regulators found in *Drosophila* cell models (*7*) are colored yellow; validated modulators predicted from the Y2H data are shown in red. Proteins that change their phosphorylation state in response to EGF stimulation (*8*) are marked with a blue ring. Y2H PPIs are shown as red arrows; other interactions from HPPI2 are in black.

shortcuts, such as the EGFR to ERK relationship, and longer paths, such as interactions from SHC to GRB2 to SOS1 to HRAS, which are in agreement with the known MAPK signal flow (*51*). Thus, our computational strategy does not define a single signaling path with a certain biological function but predicts directed PPIs that might play a role in various cellular pathways and processes. Once the directionality is defined for each interaction, the directed network can be used as a resource to extract flow paths or signaling pathways by applying additional constrains. Although the interactions lack sign (*12*), the directed PPI network supports various modeling strategies and improves the power of predictions made from PPI network analyses (*13*, *14*, *52*).

Changes in protein phosphorylation are crucial for signal transduction and phenotypic alterations in mammalian cells (*53*–*55*). We therefore investigated whether our PPI network with computationally predicted link directions enabled analysis of the dynamics of protein phosphorylation in cellular signaling processes. We integrated the directed PPI information with experimental phosphoproteomics data from EGF-stimulated mammalian cells (*8*) and examined the potential flow of information by monitoring the phosphorylation state of direct and indirect neighbors of EGF/ERK core pathway members. We observed that an EGF-triggered signal spreads from the activated EGF/ERK core signaling pathway through direct and indirect output links to more distant proteins in the network. With this analysis, we could distinguish between first- and second-order events in the signal transduction pathway, as well as resolve protein phosphorylation events in a time-dependent manner (Fig. 3 and Table 2). The results suggested that early first-order phosphorylation events represent specific signals due to direct interactions with pathway members, whereas later, second-order events are less specific due to activation of a multitude of different kinases (*32*, *33*). Therefore, we propose that on the basis of our static, directed PPI network, sequential phosphorylation events can now be predicted in the PPI network. This may enhance our understanding of how the information flow originating from an activated cellular signaling cascade leads to certain cellular phenotypes.

We also identified potential modulators of ERK phosphorylation by combining the directed PPI information with knowledge from known pERK modifiers identified previously in a *Drosophila* cell model (*7*). By following the signal flow from human orthologs of known pERK modulators along the predicted directions through the edges in the PPI network, we inferred previously unknown candidate pERK modulators. We then validated this finding by analyzing the effects of individual overexpression of 50 candidate human proteins on endogenous ERK phosphorylation in cultured cells. These cell-based studies showed that 36% (18 of 50) of the predicted potential modifier proteins altered ERK phosphorylation, and this success rate was substantially higher than that observed with random selections of proteins of equal-size groups (<9%). Although in many cases, overproduction of proteins causes phenotypic effects that are different from those of gene knockdown studies, there is no obvious reason why the overall success rate between unbiased under- and overproduction screens should differ (*56*). In a genome-wide screen in yeast, overproduction of <15% of proteins resulted in a growth defect, and only 3.5% of the proteins produced morphological changes associated with this phenotype (*57*). In another protein overproduction study, only 1.6% of all yeast proteins showed a growth defect upon moderate gene overexpression (*58*). In a screen of ~1000 genes in the developing eye of *Drosophila*, overexpression of less than 6% of the genes produced a phenotype (*59*). These rates are comparable to RNAi knockdown or gene deletion screens (*60*). Our data thus indicate that a network-based prediction of modifier proteins has a higher success rate than can be achieved with unbiased genome-wide protein overproduction or RNAi knockdown screens.

Here, we identified human PPIs that are relevant for signal transduction by systematic Y2H interaction screening, and then, after integrating these data with others, we assigned probable signaling directions to otherwise undirected protein interactions, thus constructing a directed PPI network model for further bioinformatic investigations as well as for hypothesis generation. We suggest that our network strategy is a step toward the generation of more comprehensive, dynamic cellular signaling networks, which will enable better understanding of how receptor-mediated signal flow connects proteins in core signaling pathways with the various other cellular machines and processes.

MATERIALS AND METHODS

Y2H analysis

Y2H interaction screening was performed as described previously (*17*, *18*). In brief, the L40ccU MATa yeast strain was individually transformed with plasmids (pBTM116-D9) encoding bait proteins (table S1) and tested for autoactivation of reporter genes after mating with a MATα strain producing a control Gal4-activation domain fusion protein. Eight clones producing baits that did not self-activate were pooled and mated with the arrayed library of prey L40ccα MATα yeast strains. Our existing prey library (*17*) was extended by subcloning ~2000 additional complementary DNAs (cDNAs) in a Gal4-activation domain Y2H vector (pACT4). In total, prey strains encoding ~7800 human proteins were tested in this screen for interaction. Positive clones were identified by growth on selective plates (-Leu-Trp-Ura-His), lacZ reporter gene activation assays, or both. Subsequently, preys identified in the pooled mating screens were individually tested for interactions with individual baits in a second mating assay with fresh yeast cells. The two-step interaction mating screen was repeated two to six times for different baits (table S2) to identify weak interactions, which are not well sampled in single Y2H screens (*19*). In particular, we repeatedly screened batches of individual baits two, three, or four times. A subset of proteins was screened even more often because the proteins were represented with two different clones in either the bait set or the prey matrix (table S2).

Quality assessment of Y2H data

The frequency of occurrence of six features (domain-domain interactions, common neighborhood, clustering, biological process similarity, cellular localization similarity, and coexpression) in our Y2H PPI data was measured, and precision was calculated for each measure with positive and negative reference PPI sets, each containing 10,000 PPIs. The positive set consists of PPIs from HPRD (*21*) that were either detected by two different assays or reported in two different publications. The negative set contains randomly selected PPI pairs that had not been reported previously. Raw scores were converted into relative precision scores as follows: PSBP = PBP/(PBP + NBP), with PSBP being the precision score corresponding to the raw score, PBP being the number of PPIs in the positive set above the given raw score threshold, and NBP being the number of PPIs in the negative set above the raw score threshold. Relative precision values for each interacting pair and each measure can be found in table S2.

Protein domains were annotated with the InterPro database (version 26.0), and all potential domain-domain interactions in the Y2H data were counted when they matched interacting domain pairs in the DOMINE database (version 1.1) (*61*). Domain-domain interaction information from the DOMINE database includes known and predicted interacting domain-domain pairs. Common neighborhood measure refers to the number of common neighbors (defined as shared direct-interacting partners) in HPPI1 between two interacting proteins. The HPPI1 data set was clustered with the program Cfinder, which uses the clique percolation method (*62*). For each interacting pair, we counted the number of network clusters that contained the two proteins. To determine whether interacting proteins in the network were clustered according to their molecular function (biological process similarity score), we analyzed for every pair of interacting proteins all of their assigned "biological process" Gene Ontology (GO) terms in the (GO Annotation database) to identify the most detailed level of shared GO term, that is, the common parent nodes that have the maximum distance from the root GO nodes. The distance is then used as a raw score of biological process similarity (*41*). We obtained protein localization information from the LOCATE database (human version 6), which is mapped to GO cellular component annotation (*63*). To measure the localization similarity, we used the deepest level of shared GO term (see biological process similarity score). We obtained the expression similarity data from COXPRESdb (human version 6) (*64*). COXPRESdb provides mutual rank and Pearson's correlation coefficient values for all possible gene pairs (data available for 19777 human genes). We used Pearson's correlation coefficient values as the expression similarity score.

Extraction of SPCs and its assessment for the relevance in signaling networks

The SPCs between membrane receptors and transcription factors have been proven useful in the construction of receptor signaling pathways (*24*, *25*). We first systematically analyzed the relevance of SPCs in PPI networks for cellular signal transduction cascades with KEGG pathway data only. We collected all PPIs from human KEGG pathways [PPrel in KEGG (*26*)] and merged them to create one large KEGG PPI network. The KEGG PPI network consists of 7963 interactions between 1744 proteins, 281 of which are membrane receptors and 92 are transcription factors. Using a breadth-first search algorithm (*65*), we then extracted all SPCs between every pair of membrane receptor and transcription factor, resulting in 25,852 SPCs from the KEGG PPI data. A total of 1961 KEGG-SPCs could be compared with known KEGG pathways, because both the receptor and the transcription factor of the KEGG-SPCs were present in the same KEGG pathway. We assessed the fraction of overlapping edges between the KEGG-SPCs and known KEGG pathways in comparison to 100 random SPCs. For random SPCs, edges were scrambled, preserving the number and the degree of the nodes. In comparison to randomized SPCs, the approach recalls the correct edges from the KEGG PPI network with higher frequency at various thresholds (fig. S3). This independent analysis indicates that SPCs correlate well with known signaling pathways and can thus be potentially useful to make edge direction assignments in undirected PPI networks.

HPPI1 combines the Y2H PPI data with other sets of experimentally determined human protein interactions (table S3) and contains 39,641 interactions between 9832 proteins. This includes 576 plasma membrane–localized receptor proteins (collected from Human Plasma Membrane Interactome project) (*66*) and 1166 transcription factors (*67*). Using a breadth-first search algorithm (*65*), we computed all possible shortest paths between a receptor and a transcription factor. We constructed 637,099 SPCs between 554 membrane receptors and 1150 transcription factors (fig. S4). These SPCs connect 6369 proteins through 33,857 interactions. Similar to the characteristic path length of the HPPI1 (4.14), the average length of the SPCs is 4.12. However, we found that SPCs for certain families of membrane receptors (*66*) and transcription factors (*67*) were highly overrepresented (fig. S5). For example, 36,503 SPCs were constructed between the seven-transmembrane receptor family and the zinc finger domain–containing transcription factor family. In contrast, only a single SPC was constructed between the Netrin receptor family and the Glia cell missing transcription factor family. To overcome such biases, we partitioned the data set into subsets grouping membrane receptors and transcription factors according to 23 and 47 known protein families, respectively. A total of 1081 grouped SPCs were constructed as a union of the SPCs connecting receptors and transcription factors of the same families. Each group thus represents closely related membrane receptor to transcription factor signaling paths. Individual interactions contribute to 19 SPCs on average. Within every linear SPC and grouped SPC, the edge directions of the contained interactions were set as from membrane receptor to transcription factor.

Predicting causal interactions with a naïve Bayesian classifier

To predict causal relationships between interacting proteins (edge directions) in HPPI1, we considered each interaction as two different instances, where the interaction between A and B is defined twice, as A→B and B→A. We used the naïve Bayesian classifier available at Weka, version 3.4.11 (*22*, *23*), for classification. As a training set, we selected 828 interactions from the SPCs that overlap with the KEGG pathways (*26*). From these interactions, we created 1656 instances. The instances were then classified as "positive" or "negative" on the basis of their agreement with the directions in the KEGG database. In case of bidirectional interactions, both instances were labeled as positive.

We then computed eight different features from the SPCs, grouped SPCs, and network topological properties (table S4). The features are further discretized with a supervised discretization method that is binning guided by the information in the training data. Using a leave-one-out approach, we tested the classifier's performance with 10-fold validation. We estimated the performance of the classifier with a ROC analysis calculating sensitivity [TP/(TP + FN)] and specificity [TN/(TN + FP)] with different cutoff values on the Bayesian score (Fig. 1B). For the prediction, we used the classifier that was trained with the complete training set including all 1656 instances. The precision [TP/(TP + FP)] and recall [= sensitivity, TP/(TP + FN)], which are measures that do not require a true negative set (TN), were computed at different cutoff values on the Bayesian score. To construct the HPPI2 network (Fig. 1C), we chose 70% precision and 69% recall. It contains 32,706 PPIs with direction assignments connecting 6339 human proteins (table S6).

Network motif analysis

To find the network motifs in HPPI2, we used the Mfinder program (*29*). This software uses a directed network as input and detects network motifs that occur more often in the real network than in random networks with the same size and connectivity properties. We searched for the significantly enriched network motifs with default cutoffs in the HPPI2 compared to 100 random networks. For the triad significance analysis, we used the *Z* score values obtained from the Mfinder program.

Network dynamics

We analyzed the dynamics of signal propagation with a phosphorylated protein data set (*8*) that measured the phosphorylation status of proteins at different time points, such as 0, 1, 5, 10, and 20 min after EGF stimulation. These data contain 883 phosphorylation sites in 591 proteins that exhibited a change in response to EGF treatment over the time period of 20 min. Three hundred and twenty-seven proteins were present in the HPPI2 network and were grouped according to the time point when the changing phosphorylation sites were maximally phosphorylated. Proteins with multiple sites phosphorylated upon EGF treatment could be part of more than one time-responsive data set. Enrichment was measured against 1000 equal-size protein sets picked randomly from the respective group of interacting proteins ($P < 0.05$).

Signal flow analysis to infer ERK modulators

To predict potential pERK modulators, we traced the downstream signaling paths of 337 human orthologs of known *Drosophila* pERK modulators with a signal-flow model. To rank the downstream nodes, the model assumed a signal flow that originates from each of the known pERK regulators and flows through the output edges. The signal coming from the input nodes is equally distributed among all connected output nodes. Thus, the flow model propagates a signal originating from each of the known pERK modulator proteins through the output edges in HPPI2 for three consecutive steps with a breadth-first search algorithm (*65*). The downstream nodes were ranked according to the amount of signal that they received from the various source nodes. Potential signal flow paths were inferred by connecting the nodes with high signal flow. For experimental validation, we selected high-scoring candidate proteins for which full-length cDNAs were available.

ELISA and Western blot–based pERK assays

We established an ELISA-based assay to test the effect of selected proteins on phosphorylation of endogenous ERK in HEK293 cells (*36*). After transient production of FLAG-tagged target proteins in HEK293 cells for 48 hours, cells were serum-starved overnight and then treated with EGF (150 ng/ml) for 10 min. Cells were fixed with 5% paraformaldehyde in microtiter plates and permeabilized for 10 min with 100% MeOH at −20°C. Cells were blocked in 5% bovine serum albumin, immunolabeled with an antibody that recognizes pERK (Cell Signaling #9101, 1:200), and then stained with 4′,6-diamidino-2-phenylindole (DAPI) (1:200) and a secondary Alexa-conjugated antibody (Alexa Fluor 594 goat anti-rabbit antibody, Molecular Probes #A11012, 1:500). The fluorescence (Alexa; excitation, 590 nm; emission, 617 nm) was measured in a fluorescence plate reader and normalized to cell number (DAPI; excitation, 358 nm; emission, 461 nm). Each experiment was performed in triplicates. Data are presented in table S11.

For Western blot analysis, cells were lysed after EGF treatment for 20 min on ice in Hepes buffer containing a phosphatase inhibitor cocktail (Sigma). Soluble extracts were prepared, and pERK (Cell Signaling, #9101, 1:1000) and ERK (Cell Signaling, #9102, 1:1000) proteins were detected by SDS–polyacrylamide gel electrophoresis (SDS-PAGE) and immunoblotting.

SUPPLEMENTARY MATERIALS

www.sciencesignaling.org/cgi/content/full/4/189/rs8/DC1

Fig. S1. Topological properties of the Y2H PPI network.
Fig. S2. Gene Ontology annotation of interacting proteins.
Fig. S3. Predicting causal interactions using a naïve Bayesian classifier.
Fig. S4. Extracting canonical signaling pathways as shortest path connections (SPCs) from a KEGG PPI network.
Fig. S5. Distribution of SPCs in groups of SPCs that connect receptor and transcription factor families (grouped SPCs).
Fig. S6. Topological comparison of the undirected HPPI1 and the directed HPPI2.
Fig. S7. Enriched motifs identified with Mfinder.
Fig. S8. Dynamics of EGF/ERK signal flow from receptor to transcription factors.
Table S1. Baits selected for Y2H screening.
Table S2. Y2H interactions in the Y2H PPI network.
Table S3. Human PPI interaction data sets used to construct the HPPI1 network.
Table S4. Features used by a naïve Bayesian classifier to predict edge directions.
Table S5. Comparison of different classifier performance for edge direction prediction.
Table S6. Directed interactions in HPPI2 network.
Table S7. EGF/ERK signal flow mediated by different functional proteins.
Table S8. Enrichment of single- and multisite EGF-responsive phosphoproteins.
Table S9. Dynamics of EGF/ERK signal flow in HPPI2 network.
Table S10. Dynamics of EGF/ERK signal flow from receptor to transcription factors.
Table S11. Proteins tested in an ERK phosphorylation ELISA assay.
Table S12. Characterization of MAPK6-interacting proteins.
References

REFERENCES AND NOTES

1. A. Friedman, N. Perrimon, Genetic screening for signal transduction in the era of network biology. *Cell* **128**, 225–231 (2007).
2. T. Ideker, R. Sharan, Protein networks in disease. *Genome Res.* **18**, 644–652 (2008).
3. C. Jørgensen, R. Linding, Simplistic pathways or complex networks? *Curr. Opin. Genet. Dev.* **20**, 15–22 (2010).
4. E. S. Henson, S. B. Gibson, Surviving cell death through epidermal growth factor (EGF) signal transduction pathways: Implications for cancer therapy. *Cell. Signal.* **18**, 2089–2097 (2006).
5. W. Kolch, Coordinating ERK/MAPK signalling through scaffolds and inhibitors. *Nat. Rev. Mol. Cell Biol.* **6**, 827–837 (2005).
6. P. J. Roberts, C. J. Der, Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer. *Oncogene* **26**, 3291–3310 (2007).
7. A. Friedman, N. Perrimon, A functional RNAi screen for regulators of receptor tyrosine kinase and ERK signalling. *Nature* **444**, 230–234 (2006).
8. J. V. Olsen, B. Blagoev, F. Gnad, B. Macek, C. Kumar, P. Mortensen, M. Mann, Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127**, 635–648 (2006).
9. U. Stelzl, E. E. Wanker, The value of high quality protein–protein interaction networks for systems biology. *Curr. Opin. Chem. Biol.* **10**, 551–558 (2006).
10. M. Gstaiger, R. Aebersold, Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nat. Rev. Genet.* **10**, 617–627 (2009).
11. A. Elefsinioti, M. Ackermann, A. Beyer, Accounting for redundancy when integrating gene interaction databases. *PLoS One* **4**, e7492 (2009).
12. L. J. Lu, A. Sboner, Y. J. Huang, H. X. Lu, T. A. Gianoulis, K. Y. Yip, P. M. Kim, G. T. Montelione, M. B. Gerstein, Comparing classical pathways and modern networks: Towards the development of an edge ontology. *Trends Biochem. Sci.* **32**, 320–331 (2007).
13. S. I. Berger, A. Ma'ayan, R. Iyengar, Systems pharmacology of arrhythmias. *Sci. Signal.* **3**, ra30 (2010).
14. S. Bornholdt, Systems biology. Less is more in modeling large genetic networks. *Science* **310**, 449–451 (2005).
15. 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).
16. J. F. Rual, T. Hirozane-Kishikawa, T. Hao, N. Bertin, S. Li, A. Dricot, N. Li, J. Rosenberg, P. Lamesch, P. O. Vidalain, T. R. Clingingsmith, J. L. Hartley, D. Esposito, D. Cheo, T. Moore, B. Simmons, R. Sequerra, S. Bosak, L. Doucette-Stamm, C. Le Peuch, J. Vandenhaute, M. E. Cusick, J. S. Albala, D. E. Hill, M. Vidal, Human ORFeome version 1.1: A platform for reverse proteomics. *Genome Res.* **14**, 2128–2135 (2004).
17. U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksoz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, E. E. Wanker, A human protein-protein interaction network: A resource for annotating the proteome. *Cell* **122**, 957–968 (2005).
18. H. Goehler, M. Lalowski, U. Stelzl, S. Waelter, M. Stroedicke, U. Worm, A. Droege, K. S. Lindenberg, M. Knoblich, C. Haenig, M. Herbst, J. Suopanki, E. Scherzinger, C. Abraham, B. Bauer, R. Hasenbank, A. Fritzsche, A. H. Ludewig, K. Bussow, S. H. Coleman, C. A. Gutekunst, B. G. Landwehrmeyer, H. Lehrach, E. E. Wanker, A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. *Mol. Cell* **15**, 853–865 (2004).
19. K. Venkatesan, J. F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K. I. Goh, M. A. Yildirim, N. Simonis, K. Heinzmann, F. Gebreab, J. M. Sahalie, S. Cevik, C. Simon, A. S. de Smet, E. Dann, A. Smolyar, A. Vinayagam, H. Yu, D. Szeto, H. Borick, A. Dricot, N. Klitgord, R. R. Murray, C. Lin, M. Lalowski, J. Timm, K. Rau, C. Boone, P. Braun, M. E. Cusick, F. P. Roth, D. E. Hill, J. Tavernier, E. E. Wanker, A. L. Barabási, M. Vidal, An empirical framework for binary interactome mapping. *Nat. Methods* **6**, 83–90 (2009).
20. S. Suthram, T. Shlomi, E. Ruppin, R. Sharan, T. Ideker, A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics* **7**, 360 (2006).
21. T. S. Prasad, K. Kandasamy, A. Pandey, Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. *Methods Mol. Biol.* **577**, 67–79 (2009).
22. P. Domingos, M. Pazzani, On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning* **29**, 103–130 (1997).
23. E. Frank, M. Hall, L. Trigg, G. Holmes, I. H. Witten, Data mining in bioinformatics using Weka. *Bioinformatics* **20**, 2479–2481 (2004).
24. K. D. Bromberg, A. Ma'ayan, S. R. Neves, R. Iyengar, Design logic of a cannabinoid receptor signaling network that triggers neurite outgrowth. *Science* **320**, 903–909 (2008).
25. A. Ma'ayan, S. L. Jenkins, S. Neves, A. Hasseldine, E. Grace, B. Dubin-Thaler, N. J. Eungdamrong, G. Weng, P. T. Ram, J. J. Rice, A. Kershenbaum, G. A. Stolovitzky, R. D. Blitzer, R. Iyengar, Formation of regulatory patterns during signal propagation in a Mammalian cellular network. *Science* **309**, 1078–1083 (2005).

26. M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, M. Hirakawa, From genomics to chemical genomics: New developments in KEGG. *Nucleic Acids Res.* **34**, D354–D357 (2006).
27. *Sci. Signal.* (Database of Cell Signaling, as seen April 23, 2009). http://stke.sciencemag.org/cm/
28. U. Alon, Network motifs: Theory and experimental approaches. *Nat. Rev. Genet.* **8**, 450–461 (2007).
29. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Network motifs: Simple building blocks of complex networks. *Science* **298**, 824–827 (2002).
30. R. Zaidel-Bar, S. Itzkovitz, A. Ma'ayan, R. Iyengar, B. Geiger, Functional atlas of the integrin adhesome. *Nat. Cell Biol.* **9**, 858–867 (2007).
31. R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, U. Alon, Superfamilies of evolved and designed networks. *Science* **303**, 1538–1542 (2004).
32. L. J. Holt, B. B. Tuch, J. Villen, A. D. Johnson, S. P. Gygi, D. O. Morgan, Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science* **325**, 1682–1686 (2009).
33. S. Matsuoka, B. A. Ballif, A. Smogorzewska, E. R. McDonald III, K. E. Hurov, J. Luo, C. E. Bakalarski, Z. Zhao, N. Solimini, Y. Lerenthal, Y. Shiloh, S. P. Gygi, S. J. Elledge, ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science* **316**, 1160–1166 (2007).
34. K. Okada, K. Asai, M. Arita, Flow model of the protein-protein interaction network for finding credible interactions, in *Proceedings of the 5th Asia-Pacific Bioinformatics* Conference, Hong Kong, 15 to 17 January 2007 (Imperial College Press, London, 2007), pp 317–326.
35. E. Yeger-Lotem, L. Riva, L. J. Su, A. D. Gitler, A. G. Cashikar, O. D. King, P. K. Auluck, M. L. Geddie, J. S. Valastyan, D. R. Karger, S. Lindquist, E. Fraenkel, Bridging high-throughput genetic and transcriptional data reveals cellular responses to α-synuclein toxicity. *Nat. Genet.* **41**, 316–323 (2009).
36. A. Friedman, N. Perrimon, High-throughput approaches to dissecting MAPK signaling pathways. *Methods* **40**, 262–271 (2006).
37. E. Formstecher, J. W. Ramos, M. Fauquet, D. A. Calderwood, J. C. Hsieh, B. Canton, X. T. Nguyen, J. V. Barnier, J. Camonis, M. H. Ginsberg, H. Chneiweiss, PEA-15 mediates cytoplasmic sequestration of ERK MAP kinase. *Dev. Cell* **1**, 239–250 (2001).
38. M. A. Bogoyevitch, N. W. Court, Counting on mitogen-activated protein kinases—ERKs 3, 4, 5, 6, 7 and 8. *Cell. Signal.* **16**, 1345–1354 (2004).
39. P. Coulombe, S. Meloche, Atypical mitogen-activated protein kinases: Structure, regulation and functions. *Biochim. Biophys. Acta* **1773**, 1376–1387 (2007).
40. S. Bandyopadhyay, C. Y. Chiang, J. Srivastava, M. Gersten, S. White, R. Bell, C. Kurschner, C. H. Martin, M. Smoot, S. Sahasrabudhe, D. L. Barber, S. K. Chanda, T. Ideker, A human MAP kinase interactome. *Nat. Methods* **7**, 801–805 (2010).
41. J. Yu, R. L. Finley Jr., Combining multiple positive training sets to generate confidence scores for protein–protein interactions. *Bioinformatics* **25**, 105–111 (2009).
42. A. Vinayagam, U. Stelzl, E. E. Wanker, Repeated two-hybrid screening detects transient protein–protein interactions. *Theor. Chem. Acc.* **125**, 613–619 (2010).
43. M. Steffen, A. Petti, J. Aach, P. D'haeseleer, G. Church, Automated modelling of signal transduction networks. *BMC Bioinformatics* **3**, 34 (2002).
44. C. H. Yeang, H. C. Mak, S. McCuine, C. Workman, T. Jaakkola, T. Ideker, Validation and refinement of gene-regulatory pathways on a network of physical interactions. *Genome Biol.* **6**, R62 (2005).
45. G. Bebek, J. Yang, PathFinder: Mining signal transduction pathway segments from protein-protein interaction networks. *BMC Bioinformatics* **8**, 335 (2007).
46. O. Ourfali, T. Shlomi, T. Ideker, E. Ruppin, R. Sharan, SPINE: A framework for signaling-regulatory pathway inference from cause-effect experiments. *Bioinformatics* **23**, i359–i366 (2007).
47. W. Liu, D. Li, J. Wang, H. Xie, Y. Zhu, F. He, Proteome-wide prediction of signal flow direction in protein interaction networks based on interacting domains. *Mol. Cell. Proteomics* **8**, 2063–2070 (2009).
48. N. Yosef, L. Ungar, E. Zalckvar, A. Kimchi, M. Kupiec, E. Ruppin, R. Sharan, Toward accurate reconstruction of functional protein networks. *Mol. Syst. Biol.* **5**, 248 (2009).
49. S. S. Huang, E. Fraenkel, Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Sci. Signal.* **2**, ra40 (2009).
50. A. A. Habib, S. J. Chun, B. G. Neel, T. Vartanian, Increased expression of epidermal growth factor receptor induces sequestration of extracellular signal-related kinases and selective attenuation of specific epidermal growth factor-mediated signal transduction pathways. *Mol. Cancer Res.* **1**, 219–233 (2003).
51. K. S. Ravichandran, Signaling via Shc family adapter proteins. *Oncogene* **20**, 6322–6330 (2001).
52. M. K. Morris, J. Saez-Rodriguez, P. K. Sorger, D. A. Lauffenburger, Logic-based models for the analysis of cell signaling networks. *Biochemistry* **49**, 3216–3224 (2010).
53. P. Beltrao, J. C. Trinidad, D. Fiedler, A. Roguev, W. A. Lim, K. M. Shokat, A. L. Burlingame, N. J. Krogan, Evolution of phosphoregulation: Comparison of phosphorylation patterns across yeast species. *PLoS Biol.* **7**, e1000134 (2009).
54. C. R. Landry, E. D. Levy, S. W. Michnick, Weak functional constraints on phosphoproteomes. *Trends Genet.* **25**, 193–197 (2009).
55. C. S. Tan, C. Jørgensen, R. Linding, Roles of "junk phosphorylation" in modulating biomolecular association of phosphorylated proteins? *Cell Cycle* **9**, 1276–1280 (2010).
56. T. Vavouri, J. I. Semple, R. Garcia-Verdugo, B. Lehner, Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell* **138**, 198–208 (2009).
57. R. Sopko, D. Huang, N. Preston, G. Chua, B. Papp, K. Kafadar, M. Snyder, S. G. Oliver, M. Cyert, T. R. Hughes, C. Boone, B. Andrews, Mapping pathways and phenotypes by systematic gene overexpression. *Mol. Cell* **21**, 319–330 (2006).
58. D. M. Gelperin, M. A. White, M. L. Wilkinson, Y. Kon, L. A. Kung, K. J. Wise, N. Lopez-Hoyo, L. Jiang, S. Piccirillo, H. Yu, M. Gerstein, M. E. Dumont, E. M. Phizicky, M. Snyder, E. J. Grayhack, Biochemical and genetic analysis of the yeast proteome with a movable ORF collection. *Genes Dev.* **19**, 2816–2826 (2005).
59. P. Rørth, A modular misexpression screen in *Drosophila* detecting tissue-specific phenotypes. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 12418–12422 (1996).
60. S. J. Dixon, M. Costanzo, A. Baryshnikova, B. Andrews, C. Boone, Systematic mapping of genetic interaction networks. *Annu. Rev. Genet.* **43**, 601–625 (2009).
61. B. Raghavachari, A. Tasneem, T. M. Przytycka, R. Jothi, DOMINE: A database of protein domain interactions. *Nucleic Acids Res.* **36**, D656–D661 (2008).
62. G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814–818 (2005).
63. J. L. Fink, R. N. Aturaliya, M. J. Davis, F. Zhang, K. Hanson, M. S. Teasdale, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, R. D. Teasdale, LOCATE: A mouse protein subcellular localization database. *Nucleic Acids Res.* **34**, D213–D217 (2006).
64. T. Obayashi, S. Hayashi, M. Shibaoka, M. Saeki, H. Ohta, K. Kinoshita, COXPRESdb: A database of coexpressed gene networks in mammals. *Nucleic Acids Res.* **36**, D77–D82 (2008).
65. T. H. Cormen, C. E. Leiserson, L. Ronald, C. Rivest Stein (MIT Press, 2001).
66. I. Ben-Shlomo, S. Yu Hsu, R. Rauch, H. W. Kowalski, A. J. Hsueh, Signaling receptome: A genomic and evolutionary perspective of plasma membrane receptors involved in signal transduction. *Sci. STKE* **2003**, RE9 (2003).
67. D. N. Messina, J. Glasscock, W. Gish, M. Lovett, An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res.* **14**, 2041–2047 (2004).
68. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

69. **Acknowledgments:** We thank K. Rau, I. Hoyer, and A. Redel for experimental assistance; S. Schnoegl for editorial support and critical reading of the manuscript; and A. Großmann and M. Lalowski for discussions. **Funding:** The work was supported by the Deutsche Forschungsgemeinschaft (SFB618 and SFB740), the Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (NGFN2 KB-P04T03/T05, 01GR0471, 01GR0475; NGFNp NeuroNet-TP1/TP3/TP8, 01GS08170, and 01GS08171), and the Helmholtz Association (MSBN, TP3, and TP5). **Author contributions:** A.V., U.S., and E.E.W. conceived the idea, designed the project, and wrote the manuscript. A.V. performed all computational analyses. U.S. designed the experimental part and coordinated the project. E.E.W. coordinated and supervised the project. U.S., R.F., and S.P. performed the cell culture experiments. U.S., M.Z., and J.T. performed the Y2H screening. H.E.A. and M.A.A.-N. contributed with analysis and discussion. **Competing interests:** The authors declare that they have no competing interests. **Accession numbers:** The PPI data are available in a searchable form from the PPI database MINT (MINT publication id: MINT-8089645).

Submitted 19 November 2010
Accepted 18 August 2011
Final Publication 6 September 2011
10.1126/scisignal.2001699

Citation: A. Vinayagam, U. Stelzl, R. Foulle, S. Plassmann, M. Zenkner, J. Timm, H. E. Assmus, M. A. Andrade-Navarro, E. E. Wanker, A directed protein interaction network for investigating intracellular signal transduction. *Sci. Signal.* **4**, rs8 (2011).

Science Signaling

A Directed Protein Interaction Network for Investigating Intracellular Signal Transduction

Arunachalam Vinayagam, Ulrich Stelzl, Raphaele Foulle, Stephanie Plassmann, Martina Zenkner, Jan Timm, Heike E. Assmus, Miguel A. Andrade-Navarro and Erich E. Wanker

Sci. Signal. **4** (189), rs8.
DOI: 10.1126/scisignal.2001699

Finding More Pieces to the Signaling Puzzle

Even well-studied pathways are likely to be incomplete in terms of our knowledge of all the components and their relationships, and the larger interconnected network that represents the true cellular regulatory landscape remains woefully unknown. Vinayagam *et al*. used an automated yeast two-hybrid interaction mating assay to identify protein-protein interactions (PPIs) among human proteins and then integrated that PPI data set with previously published data to create an undirected human PPI network connecting 9832 proteins through 39,641 interactions. The authors t hen applied a Bayesian learning strategy to assign direction to the interactions among the proteins. The resulting directed n etwork enabled them to evaluate growth factor–induced protein phosphorylation dynamics and to identify previously unknown modulators of the extracellular signal–regulated protein kinase pathway, of which 18 were validated with cell-based assays. This strategy should prove useful in completing the puzzle of the cellular regulatory network.

Science Signaling (ISSN 1937-9145) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. 2017 © The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. The title *Science Signaling* is a registered trademark of AAAS.