Selection on network dynamics drives differential rates of protein domain evolution

BRIAN K. MANNAKEE¹ AND RYAN N. GUTENKUNST *²

¹Division of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona
²Department of Molecular and Cellular Biology, University of Arizona

September 11, 2015

Abstract

The long-held principle that functionally important proteins evolve slowly has recently been challenged by studies in mice and yeast showing that the severity of a protein knockout only weakly predicts that protein’s rate of evolution. However, the relevance of these studies to evolutionary changes within proteins is unknown, because amino acid substitutions, unlike knockouts, often only slightly perturb protein activity. To quantify the phenotypic effect of small biochemical perturbations, we developed an approach to use computational systems biology models to measure the influence of individual reaction rate constants on network dynamics. We show that this dynamical influence is predictive of protein domain evolutionary rate in vertebrates and yeast, even after controlling for expression level and breadth, network topology, and knockout effect. Thus, our results not only demonstrate the importance of protein domain function in determining evolutionary rate, but also the power of systems biology modeling to uncover unanticipated evolutionary forces.

Over evolutionary time, every protein accumulates amino acid changes at its own characteristic rate, which Zuckerkandl and Pauling likened to the ticking of a molecular clock [1]. Remarkably, this evolutionary rate varies by orders of magnitude among proteins. Understanding the determinants of this variation is a fundamental goal in molecular evolution research [2, 3, 4, 5]. Early theoretical work suggested that functional constraints within proteins [1] and the functional importance of each protein to the organism [6, 7] would be key factors in determining evolutionary rates. Yet, empirical studies using knockouts have observed only weak effects. In bacteria [8, 9], yeast [10, 11], and mammals [12] knockout studies conclude that essential proteins evolve only slightly more slowly than non-essential proteins. Moreover, among non-essential genes in yeast, there is little to no correlation between the effect of a protein knockout on growth rate, in a wide range of conditions, and that protein’s evolutionary rate [13, 14, 11], particularly when controlling for expression level [15]. This poor correlation between knockout effects and rates of protein evolution has led some researchers to conclude that function-specific selection plays little role in determining evolutionary rates [4, 5]. This conclusion is, however, contrary to theoretical expectations, the intuition of most molecular biologists, and the reasoning behind much of comparative genomics [16], motivating our search for an alternative measure of protein function.

We reasoned that knockouts do not mimic evolutionarily relevant mutations, which often have small or moderate effects [17]. In particular, most amino-acid changes do not completely destroy a protein’s function, but rather alter its biochemical activity to a greater or lesser extent [18]. The ideal experiment would thus measure the functional effects of many random mutations on many proteins, but such experiments remain challenging [19]. To overcome this experimental limitation, we undertook a computational approach, using biochemically-detailed systems biology models to predict the effects that small perturbations to protein activities will have on the dynamics of the networks in which they function (Fig. 1). We ascribed high and low dynamical influence to protein domains for which amino acid substitutions were predicted to have respectively large or small effects on network dynamics. We hypothesized that network dynamics is a synthetic phenotype that is likely subject to natural

*To whom correspondence should be addressed. Email: rgutenk@email.arizona.edu
A : Illustrative hypothetical signaling network. The dynamical influence of the activator kinase-binding domain (A-KB) is calculated from the influences of the rate constants of the reactions in which it is involved (highlighted in blue): phosphorylation, $k_2$; dephosphorylation, $k_3$; kinase-binding, $k_4$; and kinase-unbinding, $k_5$. B: Illustrative phylogenetic analysis of the kinase-binding domain of the activator protein. C: Partial list of ordinary differential equations that model the dynamics of this network. Here all reactions are assumed to be mass-action, but that is not the case in all models analyzed. D: Dynamics of phosphorylated activator protein levels and sensitivity of those dynamics to changes in rate constant $k_2^+$, following addition of ligand L. Small increases in $k_2^+$ hasten the peak of phosphorylated activator protein and increase its steady-state level. The dynamical influence of rate constant $k_2^+$ is calculated by summing such sensitivities for all molecular species in the network. E: Illustrative plot comparing dynamical influence and evolutionary rate for all domains in the network. A single multi-domain protein can contribute multiple data points.

**Results and Discussion**

**Dynamical influence quantifies the network consequences of small-effect mutations**

A biochemically-detailed systems biology model encapsulates vast amounts of molecular biology knowledge in a form that can be used for *in silico* experimentation [20, 21]. In these models, protein biochemical activities are quantified by reaction rate constants $k$ [22]. To assess the phenotypic effects of small changes in protein activity caused by mutations, we first calculated the dynamical influence of each reaction rate constant (Materials and Methods). To do so, we calculated how a differential perturbation to that constant would change the concentration time course of each molecular species in the network (Fig. 1D), for biologically-relevant stimuli. We then normalized those changes and integrated the squared changes over time. Lastly, we summed over all molecular species in the network. The dynamical influence of a rate constant is thus the total effect that small changes in that rate constant would have on network dynamics.

The dynamical influence of each reaction rate constant quantifies its importance to network dynamics, but there is little data on evolutionary divergence of reaction rate constants to which we can compare. To compare with the abundant genomic data detailing sequence divergence at the domain level, we aggregated the influences of reaction rate constants for all reactions in which a given protein domain is involved. Whenever possible, we analyzed at the domain level, because that is the level at which distinct functions can be assigned to distinct regions of protein sequence [23]. Thus, we defined
the dynamical influence $D$ of a domain to be the geometric mean of the dynamical influences of the reaction rate constants for reactions in which it participates (Fig. 1A).

**Dynamical influence is correlated with protein domain evolutionary rate**

To test whether dynamical influence is informative about protein evolution, we analyzed dynamic protein network models from BioModels [24], a database which not only collects such models but also annotates them with links to other bioinformatic databases [25, 26]. We considered only models with experimental validation that were formulated in terms of molecular species and reactions, were runnable as ordinary differential equations, and contained at least eight distinct UniProt protein annotations. In total, we studied 12 vertebrate [27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38] and 6 yeast [39, 40, 41, 42, 43, 44] signaling and biosynthesis models. We further annotated these models to connect molecular species and reactions with particular protein domains (Dataset S1). For each model, we calculated dynamical influences for each reaction rate constant using the stimulation conditions considered in the model’s original publication (Text S1).

![Evolutionary rate is correlated with dynamical influence in signaling and biosynthetic networks.](image)

**Figure 2: Evolutionary rate is correlated with dynamical influence in signaling and biosynthetic networks.** Each point represents a protein domain, plotted given its evolutionary rate dN/dS and dynamical influence. Spearman rank correlations $\rho$ between dynamical influence and evolutionary rate are generally negative, indicative of widespread purifying selection on network dynamics. Expression level is represented by marker size and is weakly correlated with evolutionary rate but not significantly correlated with dynamical influence (Table 1). A: Vertebrate networks. Knockout essentiality is represented by color, and is not significantly correlated with evolutionary rate or dynamical influence (Table 1). B: Yeast networks. Knockout growth rate is represented by color, with red indicating a more severe phenotype. Knockout growth rate is not significantly correlated with evolutionary rate or dynamical influence (Table 1).

Using this novel method, we determined protein domain dynamical influence and evolutionary rate for 18 conserved signaling and metabolic networks (Fig. 2). We quantified the strength of the relationship between dynamical influence and evolutionary rate using Spearman rank correlations ($\rho$),
Correlation coefficient and p-values:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Correlation</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{\omega,D}$</td>
<td>-0.23</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>$\rho_{\omega,B}$</td>
<td>-0.25</td>
<td>0.0008</td>
</tr>
<tr>
<td>$\rho_{\omega,X}$</td>
<td>-0.14</td>
<td>0.0167</td>
</tr>
<tr>
<td>$\rho_{\omega,d}$</td>
<td>-0.21</td>
<td>0.0004</td>
</tr>
<tr>
<td>$\rho_{\omega,C}$</td>
<td>-0.18</td>
<td>0.0016</td>
</tr>
<tr>
<td>$r_{b_{\omega,E}}$</td>
<td>-0.12</td>
<td>0.2061</td>
</tr>
<tr>
<td>$\rho_{\omega,Gr}$</td>
<td>-0.02</td>
<td>0.8531</td>
</tr>
<tr>
<td>$\rho_{D,B}$</td>
<td>+0.17</td>
<td>0.0090</td>
</tr>
<tr>
<td>$\rho_{D,X}$</td>
<td>+0.09</td>
<td>0.0839</td>
</tr>
<tr>
<td>$\rho_{D,d}$</td>
<td>+0.07</td>
<td>0.1805</td>
</tr>
<tr>
<td>$\rho_{D,C}$</td>
<td>+0.06</td>
<td>0.2794</td>
</tr>
<tr>
<td>$r_{b_{D,E}}$</td>
<td>+0.08</td>
<td>0.3378</td>
</tr>
<tr>
<td>$\rho_{D,Gr}$</td>
<td>+0.11</td>
<td>0.2429</td>
</tr>
<tr>
<td>$\rho_{\omega,D</td>
<td>B,X,d,C,E,Gr}$</td>
<td>-0.13</td>
</tr>
</tbody>
</table>

Table 1: Overall correlations between evolutionary rate, dynamical influence, and other explanatory variables. Spearman rank ($\rho$) and rank biserial ($r_{b}$) correlation coefficients for variables evolutionary rate $dN/dS$ ($\omega$), dynamical influence ($D$), expression breadth ($B$, for vertebrates only), expression level ($X$), interaction degree ($d$), interaction betweenness centrality ($C$), knockout essentiality ($E$), and knockout growth rate ($Gr$, for yeast only). Correlations were combined over all analyzed models using the method of Hunter and Schmidt [51, 52], and p-values were calculated via permutation (Materials and Methods) as well as [53, 54]. Dynamical influence is independently predictive of evolutionary rate, as shown by the negative and statistically significant partial correlation after controlling for all other variables.

Overall correlations between evolutionary rate, dynamical influence, and other explanatory variables. Spearman rank ($\rho$) and rank biserial ($r_{b}$) correlation coefficients for variables evolutionary rate $dN/dS$ ($\omega$), dynamical influence ($D$), expression breadth ($B$, for vertebrates only), expression level ($X$), interaction degree ($d$), interaction betweenness centrality ($C$), knockout essentiality ($E$), and knockout growth rate ($Gr$, for yeast only). Correlations were combined over all analyzed models using the method of Hunter and Schmidt [51, 52], and p-values were calculated via permutation (Materials and Methods) as well as [53, 54]. Dynamical influence is independently predictive of evolutionary rate, as shown by the negative and statistically significant partial correlation after controlling for all other variables.

and in 10 of 12 vertebrate networks and 6 of 6 yeast networks, we found a negative correlation. This is consistent with the expectation that most sequences and networks evolve primarily under purifying selection [45], in which natural selection is primarily acting to remove deleterious mutations from the population. Mutations in protein domains with high dynamical influence are predicted to have greater phenotypic effect and thus, in general, be more deleterious. So mutations in those domains are more efficiently removed, and those domains evolve more slowly. Demonstrating the strength of our approach, the two exceptional vertebrate models with a positive correlation, visual signal transduction and interleukin 6 (IL-6) signaling, were recently identified as undergoing network-level adaptation in humans using population genetic data [46]. Positively selected molecular changes in rhodopsin associated with changes in absorption wavelength have been shown to effect dose-response behavior in visual signal transduction [47, 48], suggesting that network-level adaptation may compensate for changes in rhodopsin. As part of the innate immune system, IL-6 and its receptor evolve under strong diversifying selection, so downstream proteins may evolve to maintain signal fidelity. Moreover, viruses are known that directly interfere with proteins downstream of IL-6 [49, 50], potentially driving additional adaptation. Dynamical influence is thus predictive not only about purifying selection but also about adaptive selection.

The strength of the correlation between dynamical influence and protein domain evolutionary rate varies considerably among networks (Tables S1 and S2). To assess the overall strength of the relationship, we combined results across networks as a meta-analysis [51, 52]. This yielded a combined rank correlation of $\rho_{\omega,D} = -0.23$, with a permutation p-value of $p < 10^{-4}$ (Table 1), indicating that the overall pattern is highly unlikely to have arisen by chance.

**Dynamical influence calculation is robust to modeling uncertainties**

We measured dynamical influence using hand-built systems biology models; what effect do uncertainties in these models have on our analysis? To be agnostic about what aspects of network dynamics are critical to fitness, in calculating dynamical influence we summed over the integrated sensitivities of all molecular species in the network. It is, however, often evident that the builders of each model had specific molecular species in which they were most interested. If we restricted our dynamical influence calculation to those species (Text S1), we found very similar correlation with domain evolutionary rate
Figure 3: Effects of model uncertainty on dynamical influence. A: The correlation between dynamical influence and evolutionary rate \((dN/dS)\) is similarly strong in all models if dynamical influence is evaluated using all model molecular species (“full”) or only those deemed most important by the authors of the original study (“key”). B: Dynamical influences are strongly correlated between biologically-plausible parameter sets for a model of EGF/NGF signaling [27]. C: Models with overlapping domains produce positively rank-correlated estimates of dynamical influence.

Our results are thus not strongly sensitive to which aspects of network function are assumed to be subject to natural selection.

Given a network model, substantial uncertainty can exist about the values of the rate constants \(k\) [22], because they are difficult to measure directly and are thus often fit to experimental data on network behavior [55, 56]. To assess the importance of this rate constant uncertainty to our results, we used an ensemble of 2000 sets of rate constants [57] that were previously identified as consistent with experimental data for one of our models of EGF/NGF signaling [27]. We calculated the dynamical influence of all protein domains in the network using all these sets of rate constants and compared 10,000 randomly chosen pairs of sets of dynamical influences to each other. Dynamical influences calculated using different plausible rate constant sets are highly correlated (Fig. 3B), with a median rank correlation of 0.74, indicating that rate constant uncertainty does not strongly affect our results.

In addition to parameter uncertainty, different modelers may also make different assumptions when studying the same network regarding forms of interaction, which molecular players to include, or which conditions to consider. We assessed the effect of these assumptions using the models in our data which consider overlapping protein domains. The rank correlation between dynamical influences calculated for the same domains using different models varied considerably and was stronger for pairs of models with larger numbers of overlapping domains (Fig. 3C and Table S3). Combining all these correlations in a meta-analysis as before, we found an overall correlation of 0.26. For comparison, the correlation between different groups measuring gene expression in log-phase growth of budding yeast is roughly 0.62 [58], while for degree in protein-protein interaction data, the correlation is 0.11 [59]. Thus model uncertainty plays a strong but not dominant role in our analysis, and it is comparable to variables that have previously been found to be informative about evolutionary rate.

The existence of overlapping protein domains might inflate statistical significance in our meta-analyses across models. To control for this, we ordered the models by correlation between dynamical influence and evolutionary rate (negative to positive) and recalculated all correlations, removing any domains that had appeared in a previous model (Tables S4 and S5). Meta-analysis of these new correlations yielded similar results (Table S6) to the analyses with all domains (Table 1), so overlapping domains do not strongly affect our conclusions.
Dynamical influence predicts evolutionary rates independently from previously known factors

Dynamical influence captures the phenotypic effect of small perturbations to protein domain activity, but how does it correlate with previous factors linked to evolutionary rate? In multicellular organisms, proteins that are expressed in more cell types (i.e., have higher expression breadth) evolve more slowly [60], and this is true in the vertebrate networks we study (Table 1). The significant positive correlation between dynamical influence and expression breadth (Table 1) suggests that protein domains with key roles in these networks exert their effects across multiple tissues, providing a functional explanation for the observed correlation between expression breadth and evolutionary rate.

Expanding from expression breadth, the strongest known correlate with protein evolutionary rate is expression level. Proteins with greater expression evolve more slowly in both yeast [61] and vertebrates [62], which may reflect the costs of protein mis-folding [15, 63, 64, 65] or mis-interaction [66]. In our analysis, we find the expected negative correlation between evolutionary rate and expression level (Table 1), but that correlation is notably weaker than that between evolutionary rate and dynamical influence (Table 1). Indeed, dynamical influence is not significantly correlated with expression level (Table 1), indicating that dynamical influence reveals previously unanticipated evolutionary pressures beyond the strongest previously known correlate.

A significant advantage of our approach is that it captures how molecular inputs are integrated into functional phenotypic outcomes that may be selected upon. One aspect of network biology that has been previously considered in determining protein evolution is topology. Specifically, proteins with more interaction partners (i.e., greater degree) [67] or more central locations within networks (greater betweenness centrality) [68] evolve more slowly. Consistent with this previous work, we find that domain evolutionary rate has a significant negative correlation with both protein degree and betweenness centrality (Table 1). But, intriguingly, dynamical influence of protein domains is not significantly correlated with degree or betweenness centrality (Table 1) of the corresponding proteins. Why is the influence of topology not captured in our dynamics-based analysis of evolutionary rate?

Network topology is a crude measure of function; networks with the same topology can have different dynamics and thus different functions [69]. Thus, our focus on network dynamics rather than topology provides novel insight into protein domain evolution by directly quantifying system output.

These assessments of dynamical influence relative to known contributors to protein evolution clearly indicate that our approach has uncovered previously unappreciated constraints on protein evolution. Is this new insight sufficient to explain the conundrums raised by knockout experiments? In our data, we found that the correlation between evolutionary rate and knockout measures of function was so weak as to be insignificant (Table 1), consistent with prior work [12, 11]. Strikingly, across the eleven vertebrate networks that include both essential and non-essential proteins and the six yeast networks (for which knockout growth rate data is available), we find no statistical correlation between dynamical influence and essentiality or knockout growth rate (Table 1). Thus, the highly significant correlation between dynamical influence and evolutionary rate (Fig. 2, Table 1) provides a new perspective on the influence of protein function on evolutionary rate.

But, evolutionary rates are complex and likely integrate selection on multiple processes [2, 3, 4, 5]. To assess the power of our approach in comparison with alternative integrative analyses, we used partial correlation analysis [12]. Across all our networks, we find that when expression, network topology, and knockout effect are controlled for, the correlation between protein domain evolutionary rate and dynamical influence remains statistically significant (Table 1). Because the predictive power of dynamical influence cannot be explained by other factors, it provides novel and previously inaccessible insight into evolutionary rates.

Conclusions

Dynamical systems biology models offer great promise for developing and testing evolutionary hypotheses [21, 70]. Previous topological and flux-balance analysis of networks has offered some insight into protein evolution [71, 72], but dynamical models contribute substantial biological detail not previously captured by these approaches. We have shown that incorporating that detail can explain the previous lack of correlation between protein function and evolutionary rate. While dynamical models have previously been used to predict the phenotypic effects of mutations [73], here we uniquely compare such
predictions with genomic data sets to reveal previously unexplored links to evolutionary rate. Given the rapid pace of progress in systems biology modeling [74], the anticipated advances in model scope and validation will provide even more robust data sets to uncover previously unanticipated factors influencing evolutionary processes.

Materials and Methods

Dynamical influence

We defined the dynamical influence $\kappa_i$ of reaction rate constant $k_i$ by

$$\kappa_i^2 = \sum_{\text{stimulation conditions } c} \sum_{\text{molecular species } y} \frac{1}{T_c} \int_0^{T_c} \left[ \frac{dy_c(t)}{dk_i} \right] \left( \frac{k_i}{y_{\text{max}}} \right)^2 \text{d}t.$$  

(1)

Here $y_c(t)$ is the time course of molecular species $y$ in condition $c$, evaluated using the rate constant values $k^*$ from the original publication. The derivative $dy_c(t)/dk_i$ of the time course with respect to rate constant $k_i$ measures how sensitive that molecular species is to changes in that rate constant. To make relative comparisons, we normalized these sensitivities by the value $k_i$ of the rate constant and the maximum $y_{\text{max}}$ of molecular species $y$ over all stimulation conditions. To find the total effect of changes in $k_i$, we squared these normalized sensitivities and integrated over the time course of each stimulation condition, and we summed over all molecular species and stimulation conditions.

We defined the dynamical influence $D_d$ of protein domain $d$ to be the geometric mean of the influences $\kappa$ of the $N_d$ reaction rate constants for reactions in which that domain participates:

$$D_d = \left( \prod_{r=1}^{N_d} \kappa_r \right)^{1/N_d}.$$  

(2)

We took a geometric mean because rate constant sensitivities range over orders of magnitude [57].

We downloaded systems biology models in Systems Biology Markup Language (SBML) format [75] from the Feb. 8, 2012 release of BioModels [24]. We calculated dynamical influence for all protein-related biological parameters in each model, using SloppyCell [76] and simulating under the conditions considered in each model’s original paper (Text S1). These parameters represent a variety of biological phenomena, such as binding and catalytic constants and rates of diffusion and production. We considered only those parameters representing rates of biochemical reactions that depend on protein structure, because we expected constraint on those reactions to have the strongest effect on evolutionary rates. Given the dynamical influences $\kappa$ for each reaction constant, we reviewed the literature to determine the protein domain or domains at which the reaction occurs, and we assigned those influences to that domain or domains (Dataset S1).

Evolutionary rates

We obtained molecular sequences for each protein from Homologene [77] and the Saccharomyces Genome Database [78] (Fig. S1). We quantified protein or domain evolutionary rates using the ratio $dN/dS$ of the rates of nonsynonymous (dN) and synonymous (dS) DNA substitutions, calculated using PAML [79]. Further details and methods for other correlates are in Text S1.

References


**Acknowledgments**

This work was supported by the National Science Foundation, via Graduate Research Fellowship grant DGE-1143953 to BKM. BKM was also supported by an Achievement Rewards for College Scientists scholarship. We thank Tricia Serio for helpful comments on the manuscript.
Supporting Information

Brian K. Mannakee\textsuperscript{1} and Ryan N. Gutenkunst *\textsuperscript{2}

\textsuperscript{1}Division of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona
\textsuperscript{2}Department of Molecular and Cellular Biology, University of Arizona

September 11, 2015

1 Evolutionary rates

UniProt protein ID’s were acquired from the BioModels annotation in the SBML file for each model and converted to NCBI Protein IDs for vertebrates or open reading frame (ORF) numbers for yeast. Some models specified more than one UniProt ID for a single protein, in cases where there is more than one transcript identified and both appear to perform the same function (for example, MEK1 and MEK2). Where more than one UniProt ID was specified, we reviewed the model publication and the protein network literature to select a single transcript. In the case of metabolic flux models that track metabolites rather than proteins, we used the names of the enzymes involved in the reactions to find the appropriate protein identifier.

Vertebrate homologous protein alignments were downloaded from the NCBI Homologene database \cite{1}, and for each protein in the alignment, nucleotide sequence was downloaded from NCBI Entrez \cite{1}. These nucleotide sequences were then used as a template to back-translate the Homologene protein alignments to nucleotide alignments. Yeast gene information for the 7 species in the tree in Fig. S1B was downloaded from the Saccharomyces Genome Database \cite{2} on Nov. 19, 2012. These gene sequences were translated to protein amino acid sequences using Biopython \cite{3}, aligned using ClustalW \cite{4}, and then back-translated to aligned nucleotide sequence using the gene sequence as a template.

Protein domain annotation was done manually using literature review, based on information for the human protein in vertebrate models or the \textit{S. cerevisiae} protein in yeast. Evolutionary rates were calculated using codeML from PAML Version 4.4b \cite{5}, with one dN/dS ratio per tree (model 0), the F3x4 codon substitution model, and a rooted tree, as in \cite{6}. The Mgene=3 setting of codeml was used to estimate a single dN/dS ratio per annotated protein domain. We required a minimum of 4 homologs to include a gene in the analysis, and for each gene any species with more than one homologue was excluded. Because instability is a concern when estimating multiple dN/dS ratios for a single protein sequence, we iterated each codeml run until we acquired three models for which the log-likelihood was within 0.01 of the lowest log-likelihood obtained and then used the model with the lowest log-likelihood.

2 Gene expression and specificity

Vertebrate gene expression and tissue specificity data was compiled from the mouse GNF1M dataset \cite{7}, downloaded from http://bioGPS.org/downloads. The data consist of microarray probes for a number of tissue types, with each probe’s name including the corresponding gene name, which we mapped to Ensembl gene IDs using Ensembl BioMart \cite{8}. We restricted our analysis to normal adult tissues as in Fig. S2 of \cite{6}. To calculate the expression level corresponding to each microarray probe, we took the arithmetic average over replicates of the same tissue and then took the geometric average.

*To whom correspondence should be addressed. Email: rgutenk@email.arizona.edu
over tissues. To calculate the expression level of each gene, we then took the arithmetic average of the probe expression levels corresponding to that gene.

Yeast expression data [9] was obtained from http://younglab.wi.mit.edu/pub/data/orf_transcriptome.txt and used without modification.

3 Gene essentiality and dispensibility

We downloaded mouse knockout phenotype data from the Mouse Genome Informatics database [10] at http://www.informatics.jax.org/phenotypes.shtml on July 11, 2011. We assembled phenotype information for homozygous knockouts and coded a gene as essential if it resulted in one of the following phenotypes: abnormal reproductive system physiology, prenatal lethality, perinatal lethality, postnatal lethality, premature death, abnormal reproductive system morphology, lethality at weaning, preweaning lethality, partial lethality, and all sub-phenotypes of these phenotypes. If homozygous knockout of a gene did not cause one or more of these phenotypes we coded it as non-essential. To validate our parsing of this data, we compared against the results of [11].

Data for yeast knockout growth rate on YPD media were obtained from the file Regression_Tc1_hom.txt downloaded from the Stanford YDPM database http://www-deletion.stanford.edu/YDPM/YDPM_index.html on March 13, 2013.

4 Network degree and centrality

We downloaded protein-protein interaction data for both humans and yeast from the Interologous Interaction Database [12] on April 20, 2012. These data take the form of a list of interactions between two proteins, and the dataset from which the interaction was curated. Because we were interested in experimentally verified interactions we restricted our analysis to the HPRD, BIND, IntAct, and INNATEDB datasets for humans and the Krogan_Core, Yu_GoldStd, YeastHigh, YeastLow, and BIND datasets for yeast. We used the python package NetworkX [13] to load these lists of interactions and compute each protein's degree and its betweenness centrality, which is the fraction of all of the shortest paths between protein pairs in the network that pass through that protein.

5 Permutation testing

In our statistical tests, our null model was typically that dynamical influence or evolutionary rate were uncorrelated with other protein domain properties. Because domains share reactions, their influences are not independent, and thus we could not simply permute them to simulate our null model. Instead, we permuted the influences of reaction parameters, which are the most basic unit of our analysis, and then recalculated the influence for each domain based on the new sets of parameter influences. Similarly, many of the correlates with which we compare are defined on the protein level, not the domain level. In these cases, we permuted at the protein level.

6 Models considered

6.1 EGF/NGF signaling [14]

Brown et al. developed a dynamic model of the EGF/NGF signaling network in rat pheochromocytoma (PC12) cells. We downloaded the SBML file BIOMD0000000033.xml from the BioModels database and used it without modification. We simulated this model under two different conditions, 100 ng/ml EGF and 50 mg/ml NGF, as described in [14]. Protein domain, parameter, and reaction assignments are in Dataset S1. In addition to our primary analysis, we calculated dynamical influence over a restricted subset of proteins containing only activated Erk.

6.2 Arachidonic acid signaling [15]

In order to gain insights related to anti-inflammatory drug interaction and design, Yang et al. created and studied a dynamic model of the arachidonic acid signaling network in human polymorphonuclear
leukocytes. We downloaded the SBML file BIOMD0000000106.xml from the BioModels database and used it without modification, integrating the model between 0 and 60 minutes as in [15]. The model contains reactions that create and destroy each tracked molecular species. We excluded these because they are not reactions between modeled proteins. Protein domain, parameter, and reaction assignments are in Dataset S1. In addition to our primary analysis, we calculated dynamical influence over a restricted subset of proteins containing ω-LTB4 and 15-Hete.

6.3 EGF/NGF signaling [16]
Sasagawa et al. developed a model of ERK signaling networks, with parameters derived by fitting model dynamics to in vivo dynamics in PC12 cells, and studied network dynamics under a variety of stimulation conditions. We downloaded SBML file BIOMD0000000049.xml from the BioModels database and used it without modification. For all conditions simulated we integrated the network from 0 to 3600 seconds (60 minutes) as in [16]. The SBML file is coded for constant stimulation by 10 ng/ml EGF, and this is the first of the conditions we simulated. We simulated constant stimulation by 10 ng/ml NGF by using SloppyCell to set EGF concentration to 0 and NGF concentration to 10 ng/ml. Sasagawa et al. also investigated the effect of ramping the concentration of EGF (or NGF) from 0 to 1.5 ng/ml over the course of the simulation. To accomplish this we created assignment rules in SloppyCell which updated the EGF (or NGF) concentration at each time step of the integration, setting it equal to 1.5 * (time/3600) ng/ml. As in the fixed simulation conditions, the network was stimulated by EGF or NGF, but not both. Protein domain, parameter, and reaction assignments are in Dataset S1. In addition to our primary analysis, we calculated dynamical influence over a restricted subset of proteins containing only activated Erk.

6.4 EGF/MAPK cascade [17]
Schoeberl et al. modeled the EGF signaling pathway, comparing simulated time courses with experimental time courses in HeLa cells under several experimental conditions. We downloaded the SBML file BIOMD0000000019.xml from the BioModels database. The model specifies the value of parameter k5 as a piecewise function of another parameter C, and piecewise functions are not supported by SloppyCell, so we removed the piecewise function from the SBML file and used SloppyCell to create two SBML events that replicate it. We simulated the model under three experimental conditions from [17], namely stimulus with 50, 0.5, and 0.25 ng/ml EGF. For all conditions we simulated from 0 to 60 minutes. The model includes receptor internalization reactions which is not modeled mechanistically, and we excluded these reactions from our analysis. Protein domain, parameter, and reaction assignments are in Dataset S1. In addition to our primary analysis, we calculated dynamical influence over a restricted subset of proteins containing only activated Erk.

6.5 Myosin phosphorylation [18]
Maeda et al. developed a computational model of thrombin-dependent Rho-kinase activation and myosin light chain phosphorylation in human umbilical vein endothelial cells. We downloaded the SBML file BIOMD0000000088.xml for this model from the BioModels database. Some parameter names were duplicated, so we modified the model SBML file to assign unique parameter names and used the published parameter values (see Dataset S1). The only simulation condition we considered was the 0.01µM thrombin stimulation published in the SBML file, and we integrated the model from 0 to 3600 seconds as in [18]. We ignored reactions occurring in the extra-cellular compartment, particularly thrombin/thrombin-receptor interactions, and assigned all other reactions to protein domains as outlined in Dataset S1. We did not calculate a sensitivity for the parameter ratio, because it is always multiplied by another parameter \( V_{\text{max}} \) whose sensitivity we did calculate. SloppyCell interprets rate laws in terms of changes in concentration rather than changes in amount as called for by the SBML specification, so we adjusted reaction stoichiometries by a factor of 1/(compartment size) for reactants in compartment c2, the only compartment in all the models we consider with size not equal to 1. In addition to our primary analysis, we calculated dynamical influence over a restricted subset of proteins containing phosphorylated myosin light chain (pMLC) and phosphorylated myosin phosphatase targeting subunit 1 (pMYPT1).
6.6 Extrinsic apoptosis [19]

Albeck et al. developed a model of TRAIL-induced apoptosis and used it to analyze extrinsic apoptosis in HeLa cells. We downloaded the SBML file BIOMD0000000220.xml from the BioModels database and used it without modification. We simulated the model for 10 hours under the 50 ng/ml TRAIL stimulus condition encoded in the SBML file. We excluded extra-cellular reactions involving TRAIL, as well as intra-cellular reactions involving DISC and the TRAIL-DISC complex, because the protein components of the DISC are not specified in the model. We also excluded transport across the mitochondrial membrane and binding of proteins to the inner mitochondrial membrane, because these reactions are not mechanistically specified in the model. Protein domain, parameter, and reaction assignments are in Dataset S1. In addition to our primary analysis, we calculated dynamical influence summing over a restricted subset of proteins containing Caspase 3, Caspase 8, cytosolic Smac, and cytosolic Cytochrome C.

6.7 EGF/Insulin crosstalk [20]

Borisov et al. developed a model of the Ras/Erk signaling system that incorporates mechanisms of cross-talk between the EGF and Insulin signaling pathways and tested it in HEK293 cells. We downloaded the SBML file BIOMD0000000223.xml from the BioModels database and used it without modification. This SBML model, however, was altered slightly from the originally published model by adding an extra-cellular compartment of size 34. While the model page on the BioModels website says this allows for the use of the original concentrations of EGF we found that this did not create the correct dynamics. Rather than altering the model by changing the size of the extra-cellular compartment we multiplied the desired concentrations of EGF by 34, which produced the correct model dynamics. We simulated the model under four different experimental conditions, 0.01nM or 1 nM EGF with 0 or 100 nM Insulin, as described in [20]. This model contains a reaction in which Akt1 activates mTor via a chain reaction among a number of proteins that are not included in the model. As a result we only applied this reaction’s sensitivity \( \kappa \) to the Akt1 kinase domain, but not to mTor. This is the only reaction in the model in which mTor appears, so mTor was excluded from our analysis of this model. Protein domain, parameter, and reaction assignments are in Dataset S1. In addition to our primary analysis, we calculated dynamical influence summing over a restricted subset of proteins containing active (doubly phosphorylated) Erk and phosphorylated Akt.

6.8 G1 cell cycle progression [21]

Haberichter et al. constructed a dynamical model of mammalian G1 cell cycle progression in order to simulate cell cycle progression dynamics in proliferating cells continuously exposed to growth factors. We downloaded the SBML file BIOMD0000000109.xml from the BioModels database and used it without modification. We simulated the model under the experimental conditions encoded in the SBML file, integrating for 1000 minutes. The model includes reactions that create and destroy proteins, which we excluded from our analysis because they do not involve interaction with any other protein in the model. Protein domain, parameter, and reaction assignments are in Dataset S1. In addition to our primary analysis, we calculated dynamical influence summing over a restricted subset of proteins containing the two activation states, hypo- and hyper-phosphorylated, of retinoblastoma tumor suppressor protein pRb.

6.9 ErbB signaling [22]

Birtwistle et al. built a model of ErbB signaling that describes the response of the signaling network to stimulus of all four ErbB receptors with EGF and HRG (heregulin), comparing model dynamics to dynamics in MCF-7 human breast cancer cells. We downloaded SBML file BIOMD0000000175.xml from the BioModels database and used it without modification. Experimental conditions in the paper include stimulation with 0 nM, 0.5 nM, and 10 nM EGF and HRG in each possible combination of those three stimuli, for a total of 12 experimental conditions, and we simulate each of these 12 conditions for 2000 seconds. As in other models we excluded receptor internalization reactions that are not mechanistically described in the model. Protein domain, parameter, and reaction assignments are in Dataset S1. In addition to our primary analysis, we calculated dynamical influence summing
over a restricted subset of proteins containing active (doubly phosphorylated) Erk and phosphorylated Akt. In particular, we used the normalized active Erk and Akt concentrations described in [22].

6.10  Wnt/Erk crosstalk [23]

Kim et al. created a model of the Erk and Wnt pathways to investigate the effect of a positive feedback loop resulting from crosstalk between Wnt and Erk signaling, and they compared model dynamics with experimental results in HEK293 cells. We downloaded the SBML file BIOMD0000000149.xml from the BioModels database and used it without modification. This model includes a protein X which is postulated to mediate the feedback between the two pathways; it is transcribed as a result Wnt signaling and activates B-raf in the Erk network. We did not include reactions for transcription and degradation of protein X, because they are not mechanistically specified. We did, however, include the reaction in which protein X activates B-raf, which occurs on the Ras-binding domain (RBD) of B-raf, because binding at the RBD is the mode of activation of B-raf, and the reaction is modeled in the same way as Ras activation of B-raf. We excluded creation and destruction of β-catenin as well as degradation of Axin, since they are not mechanistically described in the model. Protein domain, parameter, and reaction assignments are in Dataset S1. In addition to our primary analysis, we calculated dynamical influence summing over a restricted subset of proteins containing active (doubly phosphorylated) Erk and the β-catenin/TCF complex.

6.11  Rod phototransduction [24]

Dell’Orco et al. developed a model of rod phototransduction specifically aimed at describing the mechanism of light adaptation in rod cells, verifying it by reproducing the results of several previous light adaptation response experiments in mouse rod cells. We downloaded the SBML file BIOMD0000000326.xml from the BioModels database and used it with minor modification. We removed two piecewise assignment rules, because SloppyCell does not handle piecewise rules, and replaced them with SBML events. We simulated six flash intensities, replicating those used to create Figure 7 in the publication, by setting the parameter flash0Mag to 1.54, 12.5, 45.8, 184, 800, and 2000. The parameter kP1_rev represents the rate of dissociation of phosphodiesterase from activated G-alpha molecules, and is set to zero in this model, so we excluded it from our analysis. Protein domain, parameter, and reaction assignments are in Dataset S1. In addition to our primary analysis, we calculated dynamical influence summing over a restricted subset of molecular species containing only cyclic GMP.

6.12  IL-6 signaling [25]

Singh et al. developed a model of IL-6 signaling encompassing the Jak/STAT as well as MAP kinase pathways in human hepatocytes. We downloaded the SBML file BIOMD0000000151.xml from the BioModels database and used it without modification. We simulated the experimental conditions encoded in the model, 10 ng/ml IL-6 stimulus and an initial Shp2 concentration of 100 nM. The model includes transcription, translation, and mRNA translocation for the protein SOCS which are not mechanistically detailed, so we excluded them. Protein domain, parameter, and reaction assignments are in Dataset S1. In addition to our primary analysis, we calculated dynamical influence summing over a restricted subset of proteins containing only active STAT3 dimers in the nucleus.

6.13  Trehalose biosynthesis [26]

Smallbone et al. created a model of the trehalose biosynthesis pathway in *Sa. cerevisiae*. We downloaded the SBML file BIOMD0000000380.xml from the BioModels database and used it without modification. This model reaches a steady state, and the model was validated by comparing the steady state concentrations of the metabolites in the model with experimental results in yeast experiencing heat shock. We simulated the model under the heat shock condition used in the publication, and ran the model for 50000 seconds, at which point all metabolites had reached steady state concentrations. The model creates Clb2 in a reaction that is not mechanistically specified, and we excluded this reaction from our analysis. Protein domain, parameter, and reaction assignments are in Dataset S1. In addition to our primary analysis, we calculated dynamical influence summing over only the metabolite of interest, trehalose-6-phosphate.
6.14 Glycolysis [27]

Talser et al. created a model of carbohydrate flux under oxidative stress conditions in \textit{S. cerevisiae}. We downloaded the SBML file `BIOMD0000000247.xml` from the BioModels database and used it with substantial modification. The model file in the BioModels database included unfitted parameter values, but model author Markus Ralser generously provided us with the parameter values they obtained by fitting the model to experimental data, and we reparameterized the SBML file accordingly. We simulated the wild-type experimental conditions encoded in the model for 100 minutes. Protein domain, parameter, and reaction assignments are in Dataset S1. In addition to our primary analysis, we calculated dynamical influence summing over the ratio of NADH to NADPH, the quantity of interest.

6.15 Cell cycle regulation [28]

Chen et al. developed a model of the cell cycle in \textit{S. cerevisiae} in order to investigate the complex mechanisms of cell cycle control. We downloaded the SBML file `BIOMD0000000056.xml` from the BioModels database and used it with substantial modification. This model was constructed with some reactions combined into assignment rules, making it impossible to use SloppyCell to calculate dynamical influence for individual reaction parameters. We separated these assignment rules into individual reactions and added those reactions to the model file, replacing the assignment rules. In order to verify that the modified model was correct we replicated Figures 3 and 6 from the publication. The model contains reactions causing the degradation of various proteins by SCF, which is not included in the model, and these reactions are not mechanistically described, so we excluded them from our analysis. We simulated the wild-type experimental conditions encoded in the paper for 200 minutes as in Figures 3 and 6 of the publication. Protein domain, parameter, and reaction assignments are in Dataset S1. In addition to our primary analysis, we calculated dynamical influence summing over measures of the timing of cell cycle events: MASS, BUD, ORI, and SPN.

6.16 Mitotic exit [29]

Queralt et al. developed a model of the initiation of mitotic exit in \textit{S. cerevisiae} induced by downregulation of the phosphatase Cdc50. We downloaded the SBML file `BIOMD0000000409.xml` from the BioModels database and used it without modification. This model contains reactions creating and destroying proteins Clb2, Cdc20, securin, separase, Cdc5, and Cdc15 which are not mechanistically described in the model, and we excluded these reactions from our analysis. We simulated the wild type conditions encoded in the model for 50 minutes, as in Figure 7 of the publication. Protein domain, parameter, and reaction assignments are in Dataset S1. In addition to our primary analysis, we calculated dynamical influence summing over only the concentration of active separase.

6.17 Mitotic exit [30]

Vinod et al. created a computational model of mitotic exit in \textit{S. cerevisiae} aimed at investigating the role of separase and Cdc14 endocycles. We downloaded the SBML file `BIOMD0000000370.xml` from the BioModels database and used it with substantial modification. This model was constructed using rate rules rather than reactions in SBML. Because SloppyCell our analysis is focused on reactions, we converted the rate rules to reactions, ensuring that the model remained correct by using it to generate Figure 2 from the publication. The model includes reactions creating, destroying, or (de)activating proteins Clb2, Sic1, Cln1, Cdc20, Cdh1, Swi5, Pds1, Esp1, Cdc5, Polo, and MBF which are not mechanistically detailed, and we excluded these reactions from our analysis. We simulated the wild-type experimental conditions encoded in the model for 120 minutes as in Figure 2 of the publication. Protein domain, parameter, and reaction assignments are in Dataset S1. In addition to our primary analysis, we calculated dynamical influence summing over only the concentration of active separase.

6.18 Pheromone pathway [31]

Kofahl and Klipp modelled the dynamics of the \textit{S. cerevisiae} pheromone pathway. We downloaded the SBML file `BIOMD0000000037.xml` from the BioModels database and used it without modification. The model includes reactions for the destruction of \textit{Ste2} and the export of \textit{Bar} that are not mechanistically detailed, and we excluded them from our analysis. We simulated the wild-type experimental conditions.
encoded in the model for 30 minutes, as in [31]. Protein domain, parameter, and reaction assignments
are in Dataset S1. In addition to our primary analysis, we calculated dynamical influence summing
over only the concentration of complexes M and N, which include Far1 and are required for polarized
growth and cell cycle arrest respectively.

7 Dataset S1

Excel file of data for network parameters and protein domains. Each model corresponds to two
sheets. The first sheet contains the reaction parameters, their dynamical influences, and the reactions
they correspond to. The second sheet contains the protein and domain data, including assignment of
reactions to domains and corresponding references (as PubMed IDs) [32, 33, 34, 35, 36, 37, 38, 39, 40,
41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69,
70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98,
121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141,
142, 143, 29, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162,

References


91.


2011:bar030.

95(5):717–728.

(MGD): premier model organism resource for mammalian genomics and genetics. Nucleic Acids


21(9):2076–82.

using {NetworkX}. (Pasadena, CA USA), pp. 11–15.


[64] Delgado ML et al. (2001) The glyceraldehyde-3-phosphate dehydrogenase polypeptides encoded by the Saccharomyces cerevisiae TDH1, TDH2 and TDH3 genes are also cell wall proteins. *Microbiology* 147(Pt 2):411–7.


Table S1: Correlations in vertebrate models. Spearman rank (ρ) and rank biserial (rb) correlation coefficients for variables evolutionary rate dN/dS (ω), dynamical influence (D), expression level (X), interaction degree (d), interaction betweeness centrality (C), and knock-out essentiality (E). Domains with missing values for any correlate were dropped prior to calculating correlations, and N represents the number of domains used in the analysis.

Table S2: Correlations in yeast models. Spearman rank (ρ) and rank biserial (rb) correlation coefficients for variables evolutionary rate dN/dS (ω), dynamical influence (D), expression level (X), interaction degree (d), interaction betweeness centrality (C), knock-out essentiality (E), and knock-out growth rate (Gr). Domains with missing values for any correlate were dropped prior to calculating correlations, and N represents the number of domains used in the analysis.
Between-model correlations between protein domain dynamical influences. For each pair of models with at least four overlapping domains, shown is the Spearman rank correlation and number of overlapping domains.

Table S3: Between-model correlations between protein domain dynamical influences.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\rho_{D,B}$ (p-val, N)</th>
<th>$\rho_{E}$ (p-val, N)</th>
<th>$\rho_{C}$ (p-val, N)</th>
<th>$\rho_{X}$ (p-val, N)</th>
<th>$\rho_{D}$ (p-val, N)</th>
<th>$\rho_{D,C}$ (p-val, N)</th>
<th>$\rho_{D,E}$ (p-val, N)</th>
<th>$\rho_{D,X}$ (p-val, N)</th>
<th>$\rho_{D,C}$ (p-val, N)</th>
<th>$\rho_{D,E}$ (p-val, N)</th>
<th>$\rho_{D,X}$ (p-val, N)</th>
<th>$\rho_{D,C}$ (p-val, N)</th>
<th>$\rho_{D,E}$ (p-val, N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGF/NGF signaling [14]</td>
<td>(0.0054, 32)</td>
<td>(0.0256, 28)</td>
<td>(0.3363, 32)</td>
<td>(0.4694, 31)</td>
<td>(0.0048, 28)</td>
<td>(0.1389, 28)</td>
<td>(0.8508, 32)</td>
<td>(0.0159, 31)</td>
<td>(0.1213, 27)</td>
<td>(0.7808, 31)</td>
<td>(0.9588, 31)</td>
<td>(0.0537, 31)</td>
<td>(0.4039, 31)</td>
</tr>
<tr>
<td>Arachidonic acid signaling [15]</td>
<td>(0.1140, 11)</td>
<td>(0.7827, 10)</td>
<td>(0.3909, 10)</td>
<td>(0.3794, 10)</td>
<td>(0.6067, 10)</td>
<td>(0.8169, 10)</td>
<td>(0.7336, 10)</td>
<td>(0.9367, 10)</td>
<td>(0.0952, 10)</td>
<td>(0.4502, 10)</td>
<td>(0.4039, 10)</td>
<td>(0.0537, 10)</td>
<td>(0.4039, 10)</td>
</tr>
<tr>
<td>EGF/NGF signaling [16]</td>
<td>(0.7471, 16)</td>
<td>(0.3152, 14)</td>
<td>(0.7841, 16)</td>
<td>(0.2025, 15)</td>
<td>(0.2909, 14)</td>
<td>(0.2404, 16)</td>
<td>(0.1822, 16)</td>
<td>(0.7808, 15)</td>
<td>(0.6155, 14)</td>
<td>(0.0952, 14)</td>
<td>(0.4039, 14)</td>
<td>(0.0537, 14)</td>
<td>(0.4039, 14)</td>
</tr>
<tr>
<td>EGF/MAPK cascade [17]</td>
<td>(0.0005, 4)</td>
<td>(0.3385, 4)</td>
<td>(0.6614, 4)</td>
<td>(0.0000, 4)</td>
<td>(0.3313, 4)</td>
<td>(0.0000, 4)</td>
<td>(0.3313, 4)</td>
<td>(0.0000, 4)</td>
<td>(0.3313, 4)</td>
<td>(0.0000, 4)</td>
<td>(0.3313, 4)</td>
<td>(0.0000, 4)</td>
<td>(0.3313, 4)</td>
</tr>
<tr>
<td>Receptor activation [18]</td>
<td>(0.0005, 0)</td>
<td>(0.0256, 0)</td>
<td>(0.0256, 0)</td>
<td>(0.3363, 0)</td>
<td>(0.4694, 0)</td>
<td>(0.0048, 0)</td>
<td>(0.1389, 0)</td>
<td>(0.8508, 0)</td>
<td>(0.0159, 0)</td>
<td>(0.1213, 0)</td>
<td>(0.7808, 0)</td>
<td>(0.9588, 0)</td>
<td>(0.0537, 0)</td>
</tr>
<tr>
<td>EGF/Insulin crosstalk [20]</td>
<td>(0.0125, 7)</td>
<td>(0.4012, 7)</td>
<td>(0.6842, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
</tr>
<tr>
<td>G1 cell cycle progression [21]</td>
<td>(0.0005, 1)</td>
<td>(0.0256, 1)</td>
<td>(0.0256, 1)</td>
<td>(0.3363, 1)</td>
<td>(0.4694, 1)</td>
<td>(0.0048, 1)</td>
<td>(0.1389, 1)</td>
<td>(0.8508, 1)</td>
<td>(0.0159, 1)</td>
<td>(0.1213, 1)</td>
<td>(0.7808, 1)</td>
<td>(0.9588, 1)</td>
<td>(0.0537, 1)</td>
</tr>
<tr>
<td>ErbB signaling [22]</td>
<td>(0.0125, 7)</td>
<td>(0.4012, 7)</td>
<td>(0.6842, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
</tr>
<tr>
<td>Wnt/Erk crosstalk [23]</td>
<td>(0.0125, 7)</td>
<td>(0.4012, 7)</td>
<td>(0.6842, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
</tr>
<tr>
<td>Rho-kinase activation [24]</td>
<td>(0.0125, 7)</td>
<td>(0.4012, 7)</td>
<td>(0.6842, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
</tr>
<tr>
<td>IL-6 signaling [25]</td>
<td>(0.0125, 7)</td>
<td>(0.4012, 7)</td>
<td>(0.6842, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
<td>(0.0000, 7)</td>
<td>(0.3313, 7)</td>
</tr>
</tbody>
</table>

Table S4: Correlations in yeast models with duplicate domains removed. As in Table S1, but removing any domain that appeared in a previous model (in the order shown).

Table S5: Correlations in yeast models with overlapping domains removed. As in Table S2, but removing any domain that appeared in a previous model. (in the order shown)
<table>
<thead>
<tr>
<th>correlation</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{\omega,D}$</td>
<td>-0.22 0.0010</td>
</tr>
<tr>
<td>$\rho_{\omega,B}$</td>
<td>-0.20 0.0509</td>
</tr>
<tr>
<td>$\rho_{\omega,X}$</td>
<td>-0.05 0.5103</td>
</tr>
<tr>
<td>$\rho_{\omega,d}$</td>
<td>-0.19 0.0142</td>
</tr>
<tr>
<td>$\rho_{\omega,C}$</td>
<td>-0.18 0.0219</td>
</tr>
<tr>
<td>$\rho_{\omega,E}$</td>
<td>-0.18 0.1499</td>
</tr>
<tr>
<td>$\rho_{\omega,G_r}$</td>
<td>+0.01 0.9181</td>
</tr>
<tr>
<td>$\rho_{D,B}$</td>
<td>+0.19 0.0300</td>
</tr>
<tr>
<td>$\rho_{D,X}$</td>
<td>+0.07 0.3344</td>
</tr>
<tr>
<td>$\rho_{D,d}$</td>
<td>+0.07 0.3193</td>
</tr>
<tr>
<td>$\rho_{D,C}$</td>
<td>+0.05 0.4546</td>
</tr>
<tr>
<td>$\rho_{D,E}$</td>
<td>+0.14 0.2057</td>
</tr>
<tr>
<td>$\rho_{D,G_r}$</td>
<td>+0.03 0.8096</td>
</tr>
<tr>
<td>$\rho_{\omega,D,B,X,d,C,E,G_r}$</td>
<td>-0.20 0.0035</td>
</tr>
</tbody>
</table>

Table S6: Overall correlations with overlapping domains removed. Meta-analyses as in Table 1, but based on the correlations without overlapping domains (Tables S4 and S5).

![Phylogenetic trees for species used in this study.](A) Vertebrate (B) Yeast

Figure S1: Phylogenetic trees for species used in this study. Branch lengths represent amino acid divergence.