Supplementary Text: Top-Down Network Analysis to Drive Bottom-Up Modeling of Physiological Processes

Christopher L. Poirel[1] Richard R. Rodrigues[2] Katherine C. Chen[3]
John J. Tyson[3] T. M. Murali[1,4,*]

1 The Chen2004 Model

CHEN2004 is a collection of biochemical reactions that describe protein synthesis and degradation, complex formation, regulatory activity, etc. for 27 genes known to be involved in regulating the yeast cell cycle. These reactions are modeled by a set of ordinary differential equations that describe the rate of change of each species (i.e., a protein or protein complex from the model) as a function of the quantities of other species in the model. By solving these ODEs numerically, Chen *et al.* simulated the changing quantities of every species in the model as a wild-type cell progresses through the cell cycle. To refine and test the model, Chen *et al.* then tried to simulate the unique physiological characteristics of 131 mutant strains of budding yeast. In each simulation, changes were made to the "wild-type" parameter set to reflect the genetic makeup of the mutant. For example, if the *CDC20* gene is deleted, then the rate constant for synthesis of Cdc20 protein is set to 0, and the model must reproduce the phenotype of the *cdc20*Δ deletion strain ("inviable, blocked in metaphase"). Of the 131 test strains, CHEN2004 faithfully reproduces the phenotypes of 120 mutants.

2 Assessing Edge Confidence

We assigned a confidence score to each edge in the interactome using a probabilistic approach similar to that of Yeger-Lotem *et al.* [8]. The approach assigns higher confidence to pairs of interacting proteins that participate in the same biological process. Given a pair of proteins u and v, let $I \in \{0, 1\}$ be a binary random variable such that $I = 1$ if u and v truly interact and $I = 0$ otherwise. Let $E = [E_1, \ldots, E_n] \in \{0, 1\}^n$ be a vector of binary random variables where $E_i = 1$ if experiment i supports an interaction between u and v and $E_i = 0$ otherwise. To each edge we compute a score c_{uv} representing the confidence that u and v interact given

[1]Department of Computer Science, Virginia Tech, Blacksburg, VA
[2]Genetics, Bioinformatics, and Computational Biology PhD Program, Virginia Tech, Blacksburg, VA
[3]Department of Biological Sciences, Virginia Tech, Blacksburg, VA
[4]ICTAS Centre for Systems Biology of Engineered Tissues, Virginia Tech, Blacksburg, VA
[*]To whom correspondence should be addressed.

the experimental evidence for this pair as

$$\begin{aligned} c_{uv} &= \Pr(I=1|E) \\ &= \frac{\Pr(E|I=1)\Pr(I=1)}{\Pr(E)} \qquad (1) \\ &= \frac{\Pr(E|I=1)\Pr(I=1)}{\Pr(E,I=0)+\Pr(E,I=1)} \\ &= \frac{\Pr(I=1)\prod_k \Pr(E_k|I=1)}{\Pr(I=0)\prod_k \Pr(E_k|I=0)+\Pr(I=1)\prod_k \Pr(E_k|I=1)}, \qquad (2) \end{aligned}$$

where Equation (1) is an application of Bayes rule, and Equation (2) assumes conditional independence of the experimental evidence types conditioned on I such that $\Pr(E|I) = \prod_k \Pr(E_k|I)$.

Let P and N be disjoint sets of true positive and true negative pairs, respectively. We constructed the set of gold standard positive protein pairs P as all pairs (u, v) such that both u and v were co-annotated by at least one of the Gene Ontology (GO) biological processes listed by Meyers *et al.* [5]. Expert biologists manually curated this list by selecting GO terms specific enough to be verified experimentally and general enough to be tested using high-throughput experiments. We randomly selected protein pairs that were not co-annotated by any of these biological functions as the set of negative protein pairs N, and we chose $|N| = 10 \cdot |P|$ such pairs. We computed the prior probability of an interaction $P(I)$ as

$$\Pr(I=i) = \begin{cases} \frac{|P|}{|P \cup N|}, & \text{if } i=1 \\ \frac{|N|}{|P \cup N|}, & \text{if } i=0. \end{cases}$$

Letting X_k be the set of protein pairs *observed* to interact under experiment k (i.e., the set of edges in the interactome with evidence code k), we computed the probability of an individual experiment E_k conditioned on I as

$$\Pr(E_k=e|I=i) = \begin{cases} \frac{|P \cap X_k|}{|P|}, & \text{if } e=1, i=1 \\ \frac{|N \cap X_k|}{|N|}, & \text{if } e=1, i=0 \\ \frac{|P \setminus X_k|}{|P|}, & \text{if } e=0, i=1 \\ \frac{|N \setminus X_k|}{|N|}, & \text{if } e=0, i=0. \end{cases}$$

Tables S1, S2 and S3 report the confidence scores for individual experimental evidence codes from BioGRID, KID, and YEASTRACT, respectively. Table S3 additionally reports the confidence scores for the Bodenmiller *et al.* experiment and the "Miscellaneous" category, which is the union of all experimental evidences that identified fewer than 25 interactions. The confidence reported for each experiment k was calculated as $\Pr(I=1|E)$ where $E_k = 1$ and $E_j = 0$ for experiment $j \neq k$. Many edges were discovered from multiple experiments, thus when weighting edges in the network, we computed $\Pr(I|E)$ where E is the true vector of experimental evidence codes for the pair of nodes incident on that edge. We computed confidence values close to 1 for many interactions. Such edges may have an unduly large influence on our network-based algorithms, thus we imposed a cap of 0.75 on all edge confidence scores, similar to the approach of Yeger-Lotem *et al.* [8].

BioGRID Experimental Evidence	Confidence
BioGRID Far Western	0.958770
BioGRID FRET	0.947431
BioGRID Co-purification	0.884050
BioGRID Reconstituted Complex	0.852362
BioGRID Co-crystal Structure	0.847589
BioGRID Affinity Capture-Western	0.845484
BioGRID Co-localization	0.829430
BioGRID Co-fractionation	0.817224
BioGRID Protein-peptide	0.637902
BioGRID Two-hybrid	0.576677
BioGRID Affinity Capture-MS	0.554754
BioGRID PCA	0.442114
BioGRID Biochemical Activity	0.364076
BioGRID Affinity Capture-RNA	0.257074
BioGRID Protein-RNA	0.140180

Table S1: BioGRID experimental evidence confidence scores. The confidence reported for experiment k is calculated as $\Pr(I = 1|E)$ where $E_k = 1$ and $E_j = 0$ for experiment $j \neq k$.

KID Experimental Evidence	Confidence
KID LTP Co-localization	0.935624
KID In vivo phosphorylation site mapping using phospho-specific antibodies (Western blot) or by phospho-peptide mapping	0.874626
KID In vivo site-directed mutagenesis in substrate showing same biological consequence as the kinase delete	0.860848
KID Phosphorylation reduced or absent in kinase mutant (Phospho-shifts, Western blot using Phospho-specific antibody)	0.857989
KID Phosphorylation or kinase-dependent change in localization	0.828909
KID In vitro phosphorylation site mapping (Mass Spec, Phospho-specific antibodies by Western, in vitro site-directed mutagenesis)	0.779821
KID Reconstituted complex	0.775033
KID Physical interaction by Two-hybrid or PCA	0.759793
KID HTP in vitro phosphorylation	0.749032
KID LTP in vitro kinase assays	0.735180
KID Co-Immunoprecipitation / Co-purification	0.670492
KID Reduction in phospho-peptide in vivo by mass-spec	0.646097
KID Yeast 2-Hybrid studies or PCA assay	0.641095
KID Co-Immunoprecipitation by Mass Spec	0.413561
KID Localized to same subcellular compartment	0.358568
KID Protein Chip data for in vitro phosphorylated substrates	0.224782
KID HTP In vitro PPI	0.178653

Table S2: KID experimental evidence confidence scores. The confidence reported for experiment k is calculated as $\Pr(I = 1|E)$ where $E_k = 1$ and $E_j = 0$ for experiment $j \neq k$.

YEASTRACT Experimental Evidence	**Confidence**
YEASTRACT Indirect: S1 nuclease protection assays - wild type vs TF mutant	0.871503
YEASTRACT Indirect: Northern blotting - wild type vs TF mutant	0.574786
YEASTRACT Direct: emsa	0.570063
YEASTRACT Direct: DNA footprinting	0.552197
YEASTRACT Indirect: RT-PCR - wild type vs TF mutant	0.536940
YEASTRACT Indirect: lacz - wild type vs TF mutant	0.487956
YEASTRACT Direct: lacz - wild type vs target promoter mutant	0.402631
YEASTRACT Indirect: GFP - wild type vs TF overexpression	0.398648
YEASTRACT Indirect: Proteomics - wild type vs TF mutant	0.332573
YEASTRACT Indirect: Microarrays - wild type vs TF mutant	0.199460
YEASTRACT Direct: ChIP-on-chip	0.194607
YEASTRACT Direct: ChIP	0.182758
YEASTRACT Indirect: Microarrays wild type vs TF mutant	0.154906
Bodenmiller *et al.* [3]	**Confidence**
Bodenmiller phosphorylation	0.236203
"Miscellaneous" Experimental Evidence	**Confidence**
Miscellaneous	0.588910

Table S3: YEASTRACT and miscellaneous experimental evidence confidence scores. The confidence reported for experiment k is calculated as $\Pr(I = 1|E)$ where $E_k = 1$ and $E_j = 0$ for experiment $j \neq k$. We additionally report the confidence scores for the Bodenmiller *et al.* interactions and the Miscellaneous collection of interacting pairs.

3 Selection of Functional Enrichment Algorithm

A wide variety of functional enrichment methods are available in the literature [2, 4, 6, 7]. These approaches typically perform a term-by-term analysis, reporting the significance of the relationship between each function and a collection of genes being studied. The disadvantage of these approaches is that they typically return long lists of significantly enriched functions, from which the user must determine which are the most relevant. After applying FuncAssociate [2], GSEA [7], and PAGE [4] on our datasets, we found it difficult to distinguish top-ranking functions from one another because they annotated similar collections of genes. However, Model-based Gene Set Analysis (MGSA) [1] simultaneously evaluates all gene sets using a Bayesian approach that integrates overlap between gene sets into the enrichment analysis. MGSA attempts to compute a non-overlapping set of pathways that annotate the study set. MGSA computes a posterior probability for each pathway that reflects how well the pathway overlaps with the study set while not overlapping with other pathways with higher posterior probability. We performed all tests for functional enrichment using MGSA. MGSA allows ranges to be set for two primary parameters α and β. Parameter α controls the fraction of unknown false positive genes, while β controls the fraction of unkown false negatives. We set an upper limit of 0.5 on β. Thus, less than half of the genes from the study set are not annotated by any enriched function. All other parameters were left to their default settings.

4 Extending Chen2004

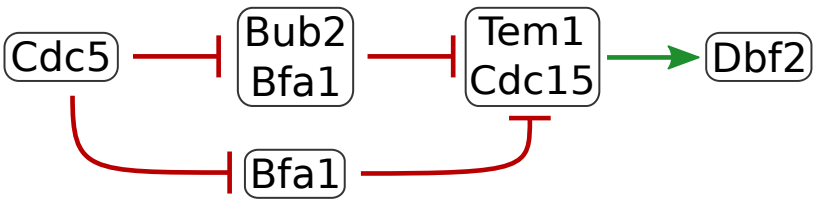

Figure S1: Two regulatory control mechanisms of Dbf2 by Cdc5.

Figure S1 illustrates the two regulatory control mechanisms of Dbf2 by Cdc5 discussed in the main manuscript. We propose the following extensions to Chen2004 that incorporate the second regulatory role of Cdc5 elucidated by Linker (i.e., Cdc5 ⊣ Bfa1 ⊣ Tem1); for consistency, we use the same notation as Chen2004:

$$\frac{d[TEM1_f]}{dt} = \frac{k_{atem} \cdot [LTE1_a] \cdot ([TEM1_T] - [TEM1_f])}{J_{atem} + ([TEM1_T] - [TEM1_f])} - \frac{k_{item} \cdot [BFA1BUB2] \cdot [TEM1_f]}{J_{item} + [TEM1_f]} \tag{3}$$

$$\frac{d[BFA1]}{dt} = \frac{(k_{abfacdc14} \cdot [CDC14] + k_{abfapp2a} \cdot [PP2A]) \cdot ([BFA1_T] - [BFA1])}{J_{abfa} + ([BFA1_T] - [BFA1])} - \frac{k_{ibfacdc5} \cdot [CDC5P] \cdot [BFA1]}{J_{ibfa} + [BFA1]} \tag{4}$$

$$\frac{d[BFA1BUB2]}{dt} = k_{asbfa1bub2} \cdot [BFA1] \cdot [BUB2] - k_{dibfa1bub2} \cdot [BFA1BUB2] \tag{5}$$

$$[MEN] = \frac{[CDC15] \cdot [TEM1_f]}{[CDC15] + [BFA1]} \tag{6}$$

5 Supplementary Figures

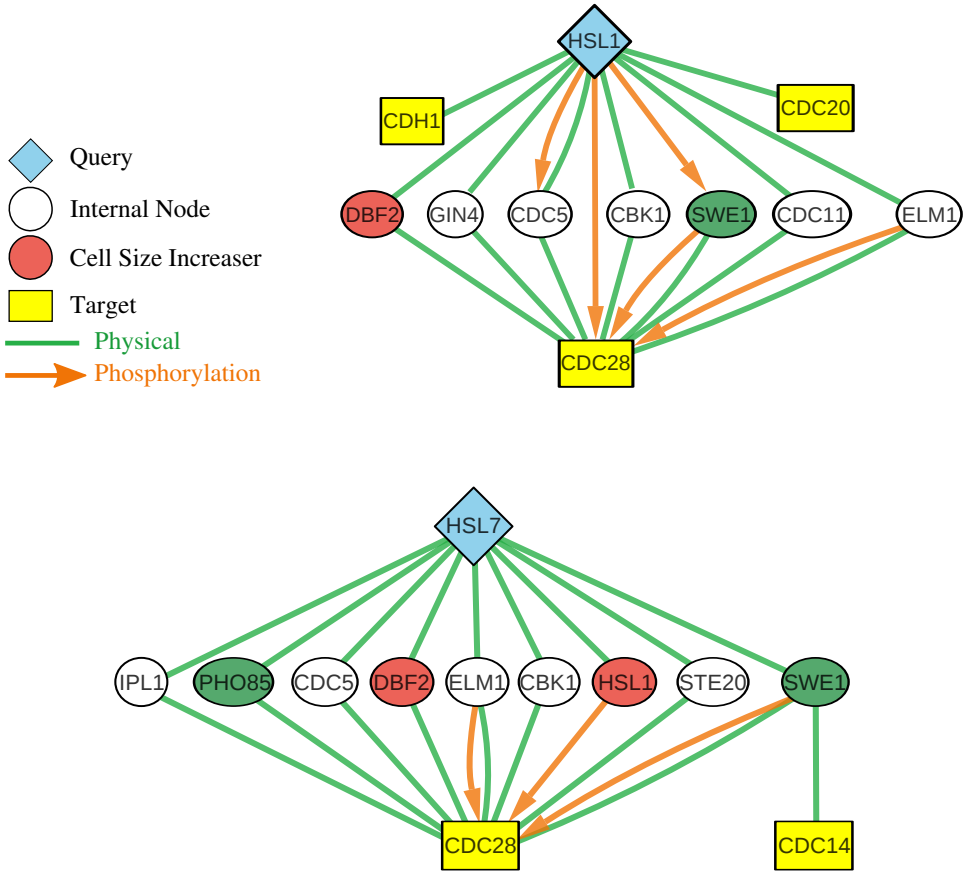

Figure S2: The $k = 10$ shortest paths connecting Hsl1 and Hsl7 to the cell cycle proteins.

References

[1] S. Bauer, J. Gagneur, and P. N. Robinson. GOing Bayesian: model-based gene set analysis of genome-scale data. *Nucleic acids research*, 38(11):3523–3532, June 2010.

[2] G. F. Berriz, O. D. King, B. Bryant, C. Sander, and F. P. Roth. Characterizing gene sets with FuncAssociate. *Bioinformatics*, 19(18):2502–4, 2003.

[3] B. Bodenmiller, S. Wanka, C. Kraft, J. Urban, D. Campbell, P. G. Pedrioli, B. Gerrits, P. Picotti, H. Lam, O. Vitek, M.-Y. Brusniak, B. Roschitzki, C. Zhang, K. M. Shokat, R. Schlapbach, A. Colman-Lerner, G. P. Nolan, A. I. Nesvizhskii, M. Peter, R. Loewith, C. von Mering, and R. Aebersold. Phosphoproteomic analysis reveals interconnected system-wide responses to perturbations of kinases and phosphatases in yeast. *Sci. Signal.*, 3(153):rs4+, Dec. 2010.

[4] S. Y. Kim and D. Volsky. PAGE: Parametric Analysis of Gene Set Enrichment. *BMC Bioinformatics*, 6(1):144+, 2005.

[5] C. Myers, D. Barrett, M. Hibbs, C. Huttenhower, and O. Troyanskaya. Finding function: evaluation methods for functional genomic data. *BMC Genomics*, 7(1):187+, July 2006.

[6] C. L. Poirel, C. C. Owens III, and T. M. Murali. Network-based functional enrichment. *BMC Bioinformatics*, 12(Suppl 13):S14, 2011.

[7] A. Subramanian, P. Tamayo, V. Mootha, S. Mukherjee, B. Ebert, M. Gillette, A. Paulovich, S. Pomeroy, T. Golub, E. Lander, and J. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 2005.

[8] E. Yeger-Lotem, L. Riva, L. J. 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, and E. Fraenkel. Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nature genetics*, 41(3):316–323, March 2009.