BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 14 2012, pages 1873–1878
doi:10.1093/bioinformatics/bts283

Gene expression

Advanced Access publication May 9, 2012

Genome-scale analysis of interaction dynamics reveals organization of biological networks

Jishnu Das[1,2], Jaaved Mohammed[1,3] and Haiyuan Yu[1,2,*]

[1]Department of Biological Statistics and Computational Biology, [2]Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY 14853, USA and [3]Tri-Institutional Training Program in Computational Biology and Medicine

Associate Editor: Martin Bishop

ABSTRACT

Summary: Analyzing large-scale interaction networks has generated numerous insights in systems biology. However, such studies have primarily been focused on highly co-expressed, stable interactions. Most transient interactions that carry out equally important functions, especially in signal transduction pathways, are yet to be elucidated and are often wrongly discarded as false positives. Here, we revisit a previously described Smith–Waterman-like dynamic programming algorithm and use it to distinguish stable and transient interactions on a genomic scale in human and yeast. We find that in biological networks, transient interactions are key links topologically connecting tightly regulated functional modules formed by stable interactions and are essential to maintaining the integrity of cellular networks. We also perform a systematic analysis of interaction dynamics across different technologies and find that high-throughput yeast two-hybrid is the only available technology for detecting transient interactions on a large scale.

Contact: haiyuan.yu@cornell.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 18, 2011; revised on March 29, 2012; accepted on May 4, 2012

1 INTRODUCTION

The protein–protein interactome of an organism is the network of all biophysically possible interactions of different proteins in that organism (Yu *et al.*, 2008). It is of key importance to accurately map this network as most proteins function by interacting with other proteins (Pawson and Nash, 2000). Moreover, a better understanding of genotype to phenotype relationships in human disease requires modeling of how disease-causing mutations might affect protein interactions and interactome properties (Goh *et al.*, 2007; Wang *et al.*, 2011). Currently, there are two main high-throughput technologies to generate high-quality protein–protein interactomes on a large scale: yeast two-hybrid (Y2H), where a protein interaction reconstitutes a transcription factor which then activates expression of reporter genes (Fields and Song, 1989), and affinity purification followed by mass spectrometry (AP/MS), where proteins bound to tagged baits are co-purified and identified (Rigaut *et al.*, 1999). High-throughput Y2H maps have been generated for yeast, fly, worm and human, while large-scale AP/MS datasets have been generated for yeast, worm and human (Jensen and Bork, 2008; Yu *et al.*, 2008). An alternative approach, adopted by most databases, is to obtain literature-curated (LC) interactions (Cusick *et al.*, 2009).

It has been shown that well-controlled Y2H and AP/MS experiments are both of high quality but of complementary nature—Y2H identifies direct binary interactions whereas AP/MS determines co-complex associations (Jensen and Bork, 2008; Yu *et al.*, 2008). Moreover, gene expression and other functional genomics datasets are routinely integrated with protein–protein interactions to validate their biological relevance—for example, interactions between proteins encoded by co-expressed genes are often considered to be of high quality (Ge *et al.*, 2001; Suthram *et al.*, 2006; von Mering *et al.*, 2002). In these analyses, gene co-expression is normally determined by a high Pearson correlation coefficient (PCC), which really means that the expression levels of the two genes are correlated over most conditions, i.e. they are globally co-expressed (Fig. 1A). Previous studies have shown that interacting proteins within stable complexes also tend to be encoded by globally co-expressed gene pairs (Jansen *et al.*, 2002; Yu *et al.*, 2008). On the other hand, the regulation and coordination of the subcellular machinery is achieved by dynamic transient interactions for example in signal transduction pathways (Jansen *et al.*, 2002). Proteins involved in transient interactions are not globally co-expressed. Rather, they share local blocks of co-expression. Transient interactions and their dynamics have significant biological importance, but most genes in these pathways are often co-expressed only under certain conditions (Fig. 1B). As a result, these are usually discarded as false positives (Ge *et al.*, 2001; Suthram *et al.*, 2006). Here, we take advantage of a novel measurement of expression relationships (Qian *et al.*, 2001) to directly distinguish stable from transient interactions on a genome-wide scale in human and yeast and systematically analyze their topological and biological significance. We also evaluate different technologies in terms of their sensitivity in detecting interaction dynamics on a genomic scale.

2 RESULTS AND DISCUSSION

2.1 Expression dynamics: global versus local co-expression

For our analysis, we created compendiums of gene expression and high-quality large-scale protein–protein interaction datasets for human and yeast (Supplementary Note SN1). We decided to use time course datasets because four distinct kinds of expression relationships—co-expression, time-shifted, inverted and inverted time-shifted—can be determined using such datasets

*The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

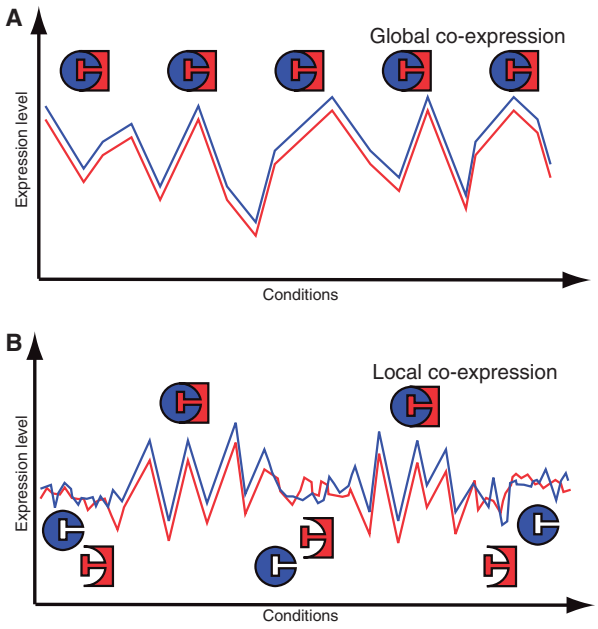

Fig. 1. Cartoon depiction of protein–protein interaction dynamics. (**A**) Gene expression profiles for two proteins that are highly correlated under all conditions indicating a stable or globally co-expressed interaction. (**B**) Two contiguous blocks of significant co-expression indicate this pair of proteins is transiently interacting or locally co-expressed.

(Qian *et al.*, 2001). Details of the biological significance of each of these four categories can be found in Supplementary Note SN2. As the cell is in a different state at each of these time points, we are in fact measuring expression under different intracellular conditions. All datasets are carefully normalized to remove potential noise (Irizarry *et al.*, 2003; Johnson *et al.*, 2007; Luscombe *et al.*, 2003; Yu *et al.*, 2007b). We also compiled high-quality large-scale protein–protein interaction datasets for human and yeast spanning both high-throughput technologies—Y2H and AP/MS. We consolidated high-quality binary interactions in the literature from various databases (Supplementary Note SN3). Although traditionally these LC interactions are considered to be of high quality, recent studies have shown that many of them, especially those supported by only one publication, in fact tend to be false positives (Cusick *et al.*, 2009). To remove unreliable interactions from our analysis, we carefully compiled comprehensive sets of high-quality binary LC interactions supported by multiple publications (named 'LC-multiple') for human and yeast. High-quality LC co-complex associations were obtained from MIPS (Mewes *et al.*, 2011) for yeast and Reactome (D'Eustachio, 2011) for human—two databases generally considered as gold standards for complexes in the corresponding organisms (Jansen *et al.*, 2002; Lage *et al.*, 2007).

From these datasets, we first calculated the PCC for expression profiles corresponding to interacting protein pairs in the high-quality interaction datasets described above. For a pair of gene expression profiles, PCC reports the global correlation of expression levels across all conditions (Qian *et al.*, 2001). A PCC value close to one indicates the pair of genes is globally co-expressed (Fig. 1A), whereas values close to zero indicate random, uncorrelated expression patterns. We find that the different interaction datasets for both human and yeast are significantly enriched for global co-expression as opposed to random gene pairs (Figs. 2A & B). Since PCC is a linear correlation coefficient and certain co-expression relationships could be non-linear, we also used the maximal information coefficient (MIC) (Reshef *et al.*, 2011) to explore global expression dynamics of the different interaction datasets in human and yeast. MIC belongs to a class of maximal information-based nonparametric exploration (MINE) statistics and has been shown to be very robust in detecting a wide range of associations both linear and not (Reshef *et al.*, 2011). Using MIC, we revalidate the global expression dynamics captured by PCC–all the high-quality interaction datasets in both human and yeast have significantly enriched global co-expression as opposed to random gene pairs (Supplementary Figure SF1). Interacting protein pairs that have PCC greater than a certain cutoff (Supplementary Note SN4) are defined as stable interactions.

However, gene pairs that are only co-expressed under certain conditions could have low and non-significant global PCC/MIC values. These often go undetected in the global nature of the computation, making global correlation an ineffective method for identifying condition-specific characteristics of transient interactions. To define dynamic co-expression relationships, we employed a Smith–Waterman-like dynamic programming algorithm as described previously (Qian *et al.*, 2001). For each pair of genes and their expression profiles, this algorithm calculates local expression-correlation scores (LES) to find subsets of conditions with correlated expression levels (Fig. 1B). Interacting proteins that do not pass the global PCC cutoff but have high LES are defined as transient interactions (see Section 4).

2.2 Interaction dynamics: stable versus transient

Next, to explore interaction dynamics across different technologies, we compared how successful different experimental techniques were in detecting stable and transient interactions. In agreement with previous studies, stable interactions within subcellular complexes show a strong enrichment of proteins encoded by globally co-expressed genes (Figs. 2A & B, Supplementary Figure SF1). On the other hand, although statistically significant, the enrichment of these globally co-expressed pairs is much less for binary interactions from both large-scale Y2H and LC sources. This lack of global co-expression has often been used as an argument to suggest that high-throughput Y2H interactions are of low quality (Ge *et al.*, 2001; Suthram *et al.*, 2006; von Mering *et al.*, 2002). However, a recent study applied orthogonal assays to experimentally confirm that these binary interactions are in fact highly reliable (Yu *et al.*, 2008). Figures 2C and D shows that in both human and yeast, Y2H is the only technology consistently able to identify transient interactions significantly more than random expectation. Surprisingly, binary interactions from the literature are not enriched with transient ones. Given the sociological biases within interactions from the literature (Cusick *et al.*, 2009; Yu *et al.*, 2008), there might be many compounding factors for this result. Stable interactions are easier to recapitulate under different experimental conditions, whereas transient interactions can only be tested under specific conditions. Therefore, transient interactions are more likely to be considered as false positives and not reported in the literature. In addition, in the post-genomic era, many candidate interaction partners are

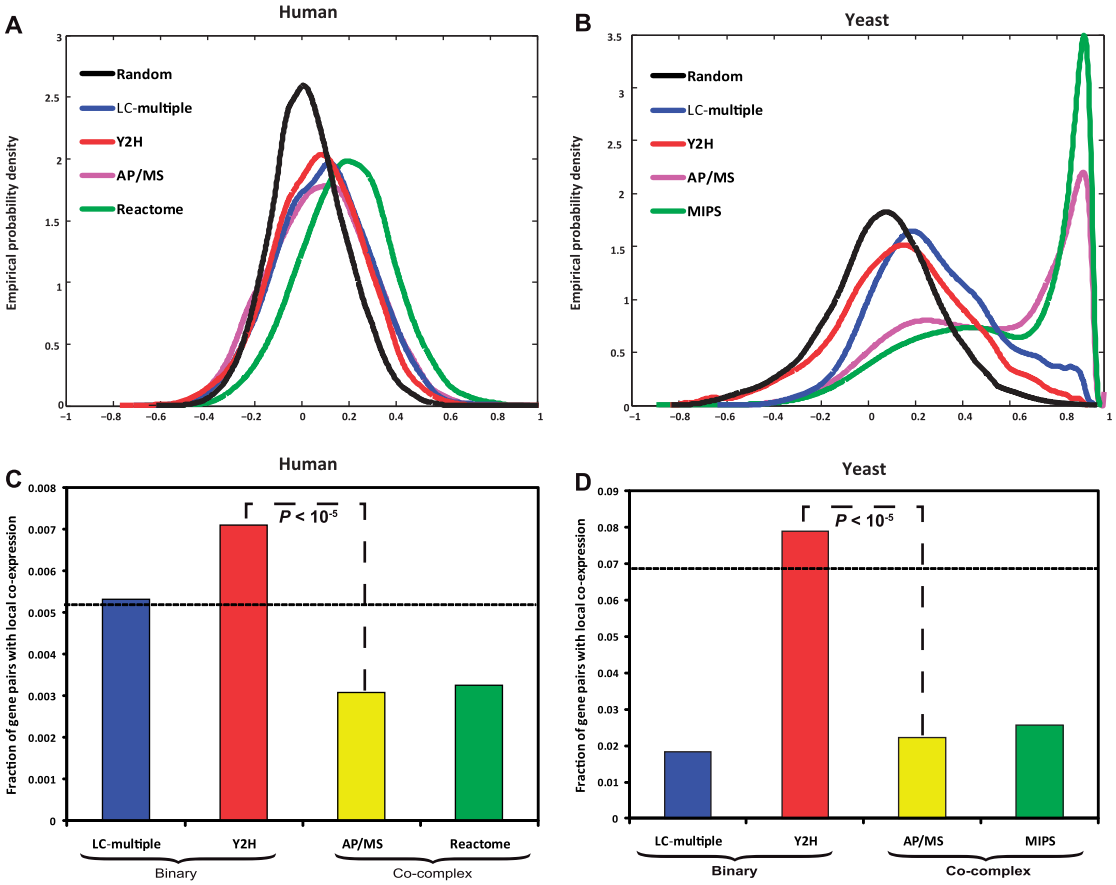

Fig. 2. (**A, B**) Enrichment of PCC of co-expression of interacting proteins (detected by different technologies) as opposed to random gene pairs in human and yeast respectively. (**C, D**) Comparison of transient interactions detected per technology in human and yeast, respectively. The dashed line indicates the overall average detection of transient interactions.

first identified based on gene expression and other genomic features favoring selection of stable interactions over transient ones. This result further highlights the importance of high-throughput Y2H because it is the only technology available to detect transient interactions, confirming that different protein interaction detection technologies capture different modes of biochemical interactions (Jensen and Bork, 2008; Yu *et al.*, 2008).

2.3 Biological significance of transient interactions

To assess the biological significance of transient interactions as defined by our algorithm, we computed functional similarity of protein pairs involved in these interactions. We find that transient interactions are significantly enriched for proteins with similar functions and the fold enrichment is comparable to that of stable interactions in both human and yeast (Supplementary Figure SF2.). These results confirm the validity of our definition of transient interactions. We therefore provide the first method to systematically detect transient interactions on a genomic scale. Although our method might miss certain transient interactions, especially extremely transient ones that are virtually impossible to distinguish from random, our results confirm that those detected by our method are high quality and share significant functional similarity.

A good example of transient interactions identified by Y2H is the interaction between Sfb2 and Sec23. This interaction has been confirmed *in vivo* (Peng *et al.*, 2000). Sec23 is a subunit of the COPII complex required for the budding of transport vesicles from endoplasmic reticulum (Miller *et al.*, 2003). *SFB2* has a 56% sequence identity with *SEC24*, an essential component of COPII involved in cargo selection (Miller *et al.* 2003). Overexpression of *SFB2* can rescue the *sec24* null mutant cells (Kurihara *et al.*, 2000). Furthermore, it has been suggested experimentally that Sfb2 may recognize different export signals from those of Sec24 and may be used under non-normal growth conditions (Miller *et al.*, 2003; Peng *et al.*, 2000). These results agree with the expression dynamics revealed by our new analysis—*SFB2* and *SEC23* are only co-expressed during stress response (Fig. 3A).

2.4 Transient interactions key in maintaining network integrity

Traditionally, in network analysis, the focus has been on nodes. Hubs are crucial in maintaining the integrity of biological networks (Albert *et al.*, 2000; Barabasi and Albert, 1999; Jeong *et al.*, 2000). Interaction networks have two broad categories of hubs. Date hubs have low average PCC with their interactors and hold the key in maintaining the integrity of cellular networks, while party hubs have

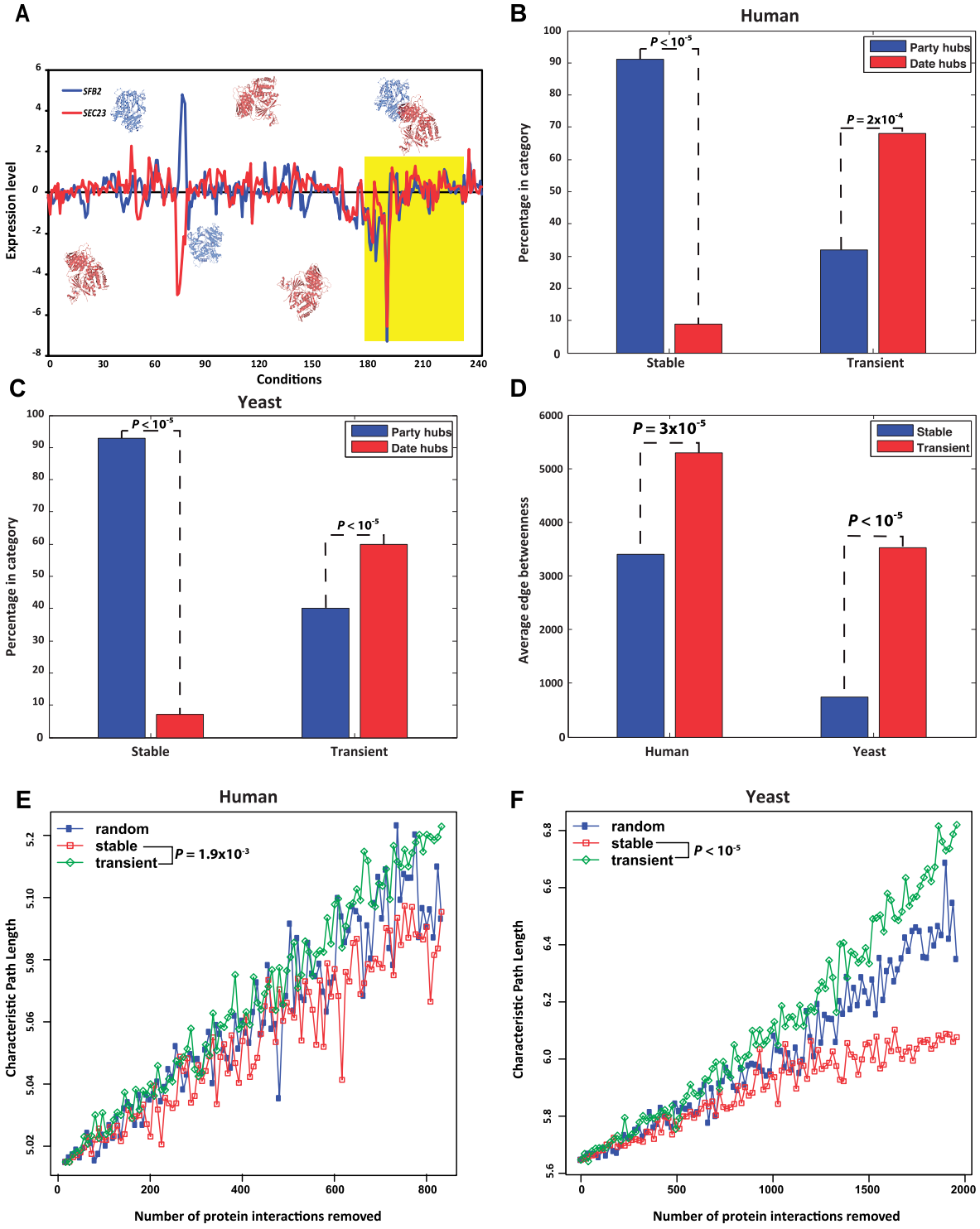

Fig. 3. (**A**) The expression profiles of SFB2 and SEC23 (co-expression only in the final yellow block). (**B, C**) Transient interactions in human are enriched in "date hubs". These have previously been shown to be vital in forming important topological links between stable functional modules. (**D**) Transient interactions in human and yeast have a significantly higher betweenness value–they hold the key in maintaining the integrity of cellular networks. (**E, F**) Characteristic path length as a measure of network connectivity after successive removal of edges of the network. Each data point represents the removal of a fixed percentage of overall nodes of the graph from each interaction type. Random removal occurs on all interactions in the network, which may include other interactions that are still uncategorized as transient or stable. Removal of transient interactions increases path length more sharply than disturbing random or stable interactions.

high average PCC with their interactors and are often contained in tightly organized modules (Han *et al.*, 2004). We find that date hubs have a significant propensity to be involved in transient interactions (Figs. 3B & 3C), suggesting that these play an important role in maintaining the integrity of the networks. To validate this result, we compared the edge 'betweenness' of global and transient interactions (Supplementary Note SN5). Edge betweenness can be used to detect community structure within networks (Girvan and Newman, 2002). Clusters detected by this approach tend to share similar functions (Dunn *et al.*, 2005). We find that transient interactions for both human and yeast have a significantly higher betweenness than stable interactions (Fig. 3D). This implies that transient interactions hold the key in maintaining the integrity of the underlying cellular network. Disrupting these will partition the interactome into disjoint clusters, unable to perform temporally and spatially well-regulated processes.

To further explore topological properties of transient interactions, we examined connectivity in response to progressive edge removal and found that selectively removing transient interactions increased characteristic path length much more sharply than selectively removing stable or random interactions (Figs. 3E & 3F). Biological interactomes are small-world networks and removing a random edge is unlikely to significantly alter connectivity, as most random edges are not essential in maintaining network integrity (Albert *et al.*, 2000). However, selectively disrupting key edges disrupts network structure and increases the characteristic path length significantly. Since removal of transient interactions causes the sharpest increase in path length, these are indeed critical for network integrity.

3 CONCLUSIONS

Here, we utilize a previously described Smith–Waterman-like dynamic programming algorithm to segregate transient interactions from stable complexes on a genomic scale directly from gene expression data. For the first time, we distinguish their biological roles and show that although transient interactions are currently underexplored, they perform key biological functions and are essential to maintaining the integrity of cellular networks. Moreover, we find that Y2H is currently the only technology that is able to determine transient interactions on a large scale. Our findings are likely to generate significant interest in designing experiments to detect transient interactions to further explore their properties.

4 METHODS

4.1 Calculating PCC, MIC and LES

PCC was calculated in a massively parallel Java program (Supplementary Note SN6) using the Parallel Java framework (Kaminsky, 2010). MIC was calculated using a Java implementation provided by Reshef *et al* (2011). Transient interactions for human and yeast were identified with a similar Parallel Java implementation of a Smith–Waterman-like dynamic programming algorithm (Supplementary Note SN7) to calculate LES (Qian *et al.*, 2001). A summarization of the total count and technology-specific count of stable and transient interactions is listed in Supplementary Tables ST1 and ST2.

4.2 Calculating betweenness and functional similarity

Edge betweenness was calculated using the Girvan–Newman algorithm (Girvan and Newman, 2002). Functional similarity was studied using total ancestry measure—a metric that takes the entire biological process tree and calculates the association of each gene with a biological process. For each protein pair query, it computes what fraction of all possible protein pairs that share the same set of Gene Ontology (Ashburner *et al.*, 2000) biological pathway terms as the query pair (Yu *et al.*, 2007a). The calculations are performed using a massively Parallel Java program (Kaminsky, 2010).

The implementations and datasets are available through our supplementary website: http://www.yulab.org/Supp/IntDynamics/.

Funding: JD is supported by the Tata Graduate Fellowship. JM is supported in part by NIH Training Grant 1T32GM083937, Tri-Institutional Training Program in Computational Biology & Medicine, awarded by the National Institute of General Medical Sciences. HY is supported by US National Institute of General Medical Sciences. This work was funded by US National Institute of General Medical Sciences grant R01 GM097358 to HY.

Conflict of Interest: none declared.

REFERENCES

Albert,R., *et al.* (2000) Error and attack tolerance of complex networks, *Nature*, **406**, 378–382.

Ashburner,M., *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Barabasi,A.L. and Albert,R. (1999) Emergence of scaling in random networks, *Science*, **286**, 509–512.

Cusick,M.E., *et al.* (2009) Literature-curated protein interaction datasets. *Nat. Methods*, **6**, 39–46.

D'Eustachio,P. (2011) Reactome knowledgebase of human biological pathways and processes. *Methods Mol. Biol.*, **694**, 49–61.

Dunn,R., *et al.* (2005) The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics*, **6**, 39.

Fields,S. and Song, O.(1989) A novel genetic system to detect protein-protein interactions. *Nature*, **340**, 245–246.

Ge,H., *et al.* (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat. Genet.*, **29**, 482–486.

Girvan,M. and Newman,M.E. (2002) Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA*, **99**, 7821–7826.

Goh,K.I., *et al.* (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.

Han,J.D., *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**, 88–93.

Irizarry,R.A., *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Jansen,R., *et al.* (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Res.*, **12**, 37–46.

Jensen,L.J. and Bork,P. (2008) Biochemistry. Not comparable, but complementary. *Science*, **322**, 56–57.

Jeong,H., *et al.* (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Johnson,W.E., *et al.* (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.

Kaminsky,A. (2010) Building parallel programs: SMPs, clusters, and Java. Course Technology, Cengage Learning, Boston, MA.

Kurihara,T., *et al.* (2000) Sec24p and Iss1p function interchangeably in transport vesicle formation from the endoplasmic reticulum in Saccharomyces cerevisiae. *Mol. Biol. Cell.*, **11**, 983–998.

Lage,K., *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.

Luscombe,N.M., *et al.* (2003) ExpressYourself: a modular platform for processing and visualizing microarray data. *Nucleic Acids Res.*, **31**, 3477–3482.

Mewes,H.W., *et al.* (2011) MIPS: curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Res.*, **39**, D220–D224.

Miller,E.A., *et al.* (2003) Multiple cargo binding sites on the COPII subunit Sec24p ensure capture of diverse membrane proteins into transport vesicles. *Cell*, **114**, 497–509.

Pawson,T. and Nash,P. (2000) Protein-protein interactions define specificity in signal transduction. *Genes Dev.*, **14**, 1027–1047.

Peng,R., *et al.* (2000) Evidence for overlapping and distinct functions in protein transport of coat protein Sec24p family members. *J. Biol. Chem.*, **275**, 11521–11528.

Qian,J., *et al.* (2001) Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *J. Mol. Biol.*, **314**, 1053–1066.

Reshef,D.N., *et al.* (2011) Detecting novel associations in large data sets. *Science*, **334**, 1518–1524.

Rigaut,G., *et al.* (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.*, **17**, 1030–1032.

Suthram,S., *et al.* (2006) A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics*, **7**, 360.

von Mering,C., *et al.* (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403.

Wang,X., *et al.* (2011) Network-based methods for human disease gene prediction. *Brief Funct Genomics*, 10, 280–293.

Yu,H., *et al.* (2008) High-quality binary protein interaction map of the yeast interactome network. *Science*, **322**, 104–110.

Yu,H., *et al.* (2007a) Total ancestry measure: quantifying the similarity in tree-like classification, with genomic applications. *Bioinformatics*, **23**, 2163–2173.

Yu,H., *et al.* (2007b) Positional artifacts in microarrays: experimental verification and construction of COP, an automated detection tool. *Nucleic Acids Res.*, **35**, e8.