

Initial Large-scale Exploration of Protein-protein Interactions in the Human Brain

Jake Y. Chen*, Andrey Y. Sivachenko, Russell Bell, Connie Kurschner, Irene Ota, and Sudhir Sahasrabudhe

Myriad Proteomics, Inc., 2150 W. Dauntless Ave, Salt Lake City, UT 84116, USA.

ABSTRACT

Study of protein interaction networks is crucial to post-genomic systems biology. Aided by high-throughput screening technologies, biologists are rapidly accumulating protein-protein interaction data. Using a random yeast two-hybrid (R2H) process, we have performed large-scale yeast two-hybrid searches with approximately fifty thousand random human brain cDNA bait fragments against a human brain cDNA prey fragment library. From these searches, we have identified 13,656 unique protein-protein interaction pairs involving 4,473 distinct known human loci. In this paper, we have performed our initial characterization of the protein interaction network in human brain tissue. We have classified and characterized all identified interactions based on Gene Ontology (GO) annotation of interacting loci. We have also described the “scale-free” topological structure of the network.

INTRODUCTION

Development of high-throughput screening technologies, and the completion of a number of genome sequencing projects, has led to the emergence of post-genomic systems biology. In this new era, biologists are gaining increasing interest in analyzing system-wide biological data, including whole-genome sequence data, cross-genome homolog information, global gene expression profiles, transcriptional regulatory networks, and protein interaction networks. The system-scale analysis of biological data can provide traditional biologists with new perspectives to understand functions of interconnected genes and proteins in their complex cellular and molecular context. We characterize systems biology studies as following a paradigm that consists of iterative cycles of two distinct but related stages. The first stage is a data-driven “bottom-up” knowledge discovery process. At this stage, computational scientists sift through large volumes of primary biological data to reveal high-order “patterns” or “models” that can characterize the underlying data. The second stage is a hypothesis-driven “top-down” knowledge discovery process. At this stage, computational biologists and biologists team up to infer and validate previously unknown details of biological processes or molecular functions as guided by global perspectives of high-order “models”.

At Myriad Proteomics, we are undertaking a system-scale human protein-protein interaction discovery project, which involves high-throughput yeast two-hybrid screening, tandem-affinity purification/mass spectrometry analysis, robotics, software engineering, data-driven bioinformatics, and target-driven drug discovery activities [1]. The scope of our efforts are unprecedented, since all other reported protein interaction projects so far have concentrated either on small non-mammalian organisms [9, 12] or on small regional networks [7]. We believe that collecting and analyzing such data will help us understand protein function and molecular signaling networks leading to drug target discoveries [3]. In this short paper, we describe high-

* To whom correspondence should be addressed. E-mail: jchen@myriad-proteomics.com, Tel: (801) 303-1722.

level characteristics of the human brain protein interaction data and their sub-networks, which have been collected and compiled from random yeast-2-hybrid (R2H) high-throughput process. Specifically, we describe general statistical properties of our data set, present visualization of the structure of the data, and examine the topology of the interaction network.

METHODS

In this study, we have based our analysis on data generated from a “*random yeast two-hybrid*” (R2H) system. For a comprehensive review of related experimental and computational data generation techniques, refer to [14, 17]. Compared with the standard directed yeast two-hybrid system [5], the R2H system makes the following two distinctions. First, we prepare bait cDNA libraries using randomly fragmented cDNAs prior to subcloning. Our aim is to generate cDNA insert sizes averaging 800 to 900 base pairs, which produce small hybrid proteins that facilitate protein-protein interactions. Second, we prepare a mixed yeast bait clone library in the beginning of the process, but then clonally isolate the individual baits and separately perform the mating of one yeast bait clone at a time against a whole yeast prey clone library (which we call a “R2H search”). Therefore, we can perform many R2H searches of anonymous bait clones in parallel, and leave the task of revealing the identity of interacting pairs from positive clones to the subsequent sequencing of both the bait and prey clone cDNA inserts. In all, the R2H system, in conjunction with our extensive use of automation and array-format (96x and 384x formats), enables us to accommodate a throughput of approximately 6,000 R2H searches per week.

We have designed and implemented a Laboratory Information System (LIMS) to manage the R2H data collection process, and a database platform based on data modeling design principles described in [6] to mine and explore our interaction data set. Our major bioinformatics data preparation steps include: collecting primary experimental data through LIMS, performing base-calling, cleaning sequences and clipping vector regions using CrossMatch, assembling sequencing reads from both 5'-end and 3'-end of the vector inserts using CAP3, identifying the assembled sequence insert by performing BLAST against the NCBI REFSEQ database, and annotating sequences using imported databases such as LocusLink and Gene Ontology [2, 11]. We have collected and analyzed interactome data involving more than 50,000 R2H searches against a prey cDNA library from expressed mRNAs in homogenized human brain. From these R2H searches, we have created protein-protein interaction data set for this study, which contains 13,656 uniquely identified binary protein-protein interaction pairs involving 4,473 distinct protein loci.

We have performed data analysis and visualization by using a combination of software tools, including Oracle9i, S-Plus Analytic Server 2, and Spotfire DecisionSite Browser 7.1.

RESULTS

We collected the statistics of a subset of our interaction data that comprises approximately 50,000 R2H searches. We determined three major characteristics of the data. First, we calculated **search positive rate**, which we define as the number of positive clones either “observed” or “observed-and-picked” per search. We found that ~33% of searches generated between 1 and 100 positive clones, ~2% generated more than 100 positive clones, and the remaining ~65% generated no positive clones (*null* searches). Thus, there were 17,371 searches (~90% of them contained unique baits) that gave rise to positive clones. On average, we observed a search positive rate of approximately three while counting all *null* searches, or approximately eight to

nine while not counting any *null* searches. In Figure 1, we showed a plot of “observed-and-picked” (or “picked” in short) search positive counts for all of the non-*null* searches.

Second, we calculated **interaction discovery rate**, which we define as the number of unique interaction pairs per search. We accumulated a total of 13,660 unique interaction pairs from 12,808 searches where both bait and prey loci have been identified. Among these 13,660 unique interaction pairs, 12,466 (~90%) were observed at least once within only one R2H search and 8,501 (~62%) were observed only once within only one R2H search. The high percentage (~90%) of novel searches suggested potentially high search efficiency of our process; the repeated detection (1-62%=38%) of the same interaction pairs more than once, on the other hand, enabled us to infer system errors of our R2H search process. While monitoring the accumulation of unique interaction pairs throughout the 17,371 non-*null* searches, we noticed a relatively stable “interaction discovery rate” of approximately 1.1 over the entire search period. This suggested that we might not have saturated all the interactions in the constructed human brain libraries.

Third and last, we calculated the **cDNA fragment size for bait and prey constructs**. Our bait and prey clone cDNA fragment insert size for all the positive clones followed a normal distribution with a mean of ~900bp and a standard deviation of ~250bp. This suggested that our human protein fragments enlisted into the interaction events were large enough to accommodate the majority of documented protein domains, 90% of which should be within 300 amino acid residues long [15].

We applied information visualization techniques, which enabled us to explore high-level data patterns and to follow up with queries of the underlying data. In Figure 2 we plotted a two-dimensional “heat map” (zoomed in to show a data subset consisting of 3,392 bait-prey protein interaction pairs) created within the Spotfire DecisionSite Browser. The heat map within the Spotfire software displayed protein-protein interaction pairs and, upon user selection of a set of data points on the plot, could display interacting details on a side panel (not shown). With this heat map, we could quickly detect “promiscuous interacting proteins” (proteins that tend to unselectively interact with many other protein partners) visually as either vertical lines or horizontal lines on the plot, and examine detailed protein descriptions by clicking on the dot representing a particular interaction pair. A vertical line often suggests that the bait involved is a “self activator”, a bait protein that can activate Y2H transcription without requiring a specific interacting prey protein. A horizontal line, on the other hand, suggests that the prey involved is a “false positive” or “sticky prey”, a prey protein that can engage in a wide spectrum of bait-prey interactions that activate Y2H transcription unselectively. For example, arrestin ARRB1 (pointed to by an arrow in the plot) showed a horizontal line pattern, suggesting that it might be a “false positive prey” in our R2H system. By browsing through ARRB1 interacting proteins, however, we also found that majority of the characterized observed interaction partners of ARRB1 were transmembrane receptors in accordance with the well established role of ARRB1 in dampening activated G-protein coupled receptor signals in cells [8]. Therefore, by providing biologists with a global interacting data set and a tool to “drill down” to the data details, we reap the benefits of our investment in “systems biology”.

We showed three high-level classifications of protein interaction pairs using gene ontology (GO) categories in Figures 3a-c [2]. Figure 3a showed a categorized “heat map” view of 9055 unique protein-protein interaction pairs (13,660 identified interaction pairs, minus 4605 pairs that contain molecular function annotation for neither the bait nor the prey protein) aggregated into

17x16 bins high-level GO molecular function terms (including 16 molecular function categories and an additional “unknown” categories for uncharacterized proteins^{*}). To accomplish this, we annotated all the proteins by tracing back their individual original GO annotation terms to their ancestor terms at the fixed level in the GO hierarchy. We made two observations from this visualization. First, we observed diverse but non-uniform and non-diagonal patterns in the categorized protein-protein interactions. Earlier studies (see, *e.g.* [10]) assumed that a diagonal pattern (i.e., proteins within the same category interacting with each other) was expected for most protein interactions found in the literature. However, we believe that this pattern should be difficult to observe due to a large percentage of unknown proteins and many proteins with multiple/incomplete GO category assignments. Besides, certain cross-category interactions are biologically plausible. For example, we did observe several significant and biologically interesting cross-category interaction patterns such as “enzyme—ligand binding molecule” interactions and “signal transducer—structural molecule” interactions. Our interaction data was also heavily concentrated around two functional categories of proteins—5,887 distinct interactions involving at least one ligand binding protein and 2,261 distinct interactions involving at least one enzyme—with each category interacting with proteins from all 17 GO categories. Second, we detected an opportunity to assess functions of previously uncharacterized proteins via their interactions. The figure showed that a total of 2,996 (67%) of all 4,474 observed distinct protein loci fell into the “unknown molecular function” category. However, 2,115 (71%) of these 2,996 proteins also interacted with at least one GO annotated protein. Overall, these “uncharacterized”-“characterized” protein interactions represented 49% (6,665 /13,660) of all unique interactions on the plot. This observation provides both an opportunity for inferring functions of uncharacterized proteins through their interaction context and a challenge for assessing the biological significance of the interaction pairs. Figures 3b and 3c show the same protein interaction pairs as Figure 3a categorized using high-level GO cellular component terms (Figure 3b) and biological process terms (Figure 3c).

Lastly, we investigated the topology of the protein-protein interaction subnetwork derived from our data. In Figure 4A (also in the inset), we plotted the distribution $P(k)$ of network “node degrees” k , where k represent the number of immediate interaction partners for the current protein as a network node. We showed that the distribution of protein node degrees exhibited a power-law dependence, i.e., $P(k) \propto k^{-\gamma}$ (with $\gamma \approx 1.7$). This distribution showed the existence of a large number of nodes with small node degrees, and a very small number of nodes with very large node degrees (with hundreds of connections). In contrast, in Figure 4B, the $P(k)$ distribution of a randomly constructed interaction network (by randomly choosing interacting pairs from a pool of available proteins) followed a very different Poisson distribution. The power-law $P(k)$ dependence is the key signature of “scale-free” networks, a type of network that characterize the World Wide Web and many social networks, and have been described only recently for certain metabolic networks and protein domain networks in biology [4, 13, 16]. A scale-free network implies the following property: the network is neither completely modular, *i.e.* it cannot be separated into a set of independent subcomponents, nor completely random, it is highly robust against errors caused by disruption of a randomly chosen node, and yet it is highly vulnerable to perturbations of the small number of highly connected protein nodes known as “network hubs”. The “scale-free” network property of our data gave us two insights. First, since

^{*} Note that the Figure 3a-c shows individual counts of pairs falling into each bin, and due to multiple annotations available for many loci, the total sum over all the bins exceeds the number of interacting pairs.

the power-law distribution $P(k)$ has a long tail with some nodes having a large k potentially serving “network hub” functions, we could no longer treat the problem of “promiscuous interacting proteins” simply by setting a fixed threshold k_0 and discarding all the proteins with $k > k_0$. In fact, these few “network hubs” might provide important clues as to how different molecular signals is broadcasted and dampened within the cell. Second, we also believe that the power-law distribution of our network implies a low system error rate (defined as the rate of pairing proteins randomly).

We plan to continue characterizing the human interactome data as more information accumulates. When combined with our data from tandem-affinity purification/mass spectrometry systems, and integrated with different types of genome, microarray, and disease pathway information, this data set will provide us with a detailed protein function roadmaps.

ACKNOWLEDGEMENT

We thank Dr. Christopher Martin for his invaluable critical comments during our manuscript preparations. We thank Dr. Manuel Rodriguez, Dr. Robert Hughes, and Alan James for their support throughout the project. We also thank Hsiao-kun Tu, Manjula Aliminati, Amit Phansalkar, Hisayoshi Zaima, and Mitsuhiro Kanazawa for their assistance in preparing the data.

REFERENCE

- [1] *Myriad Proteomics WWW Site*, <http://www.myriad-proteomics.com/>. 2002.
- [2] Ashburner, M., et al., *Gene ontology: tool for the unification of biology*. *The Gene Ontology Consortium*. Nat Genet, 2000. **25**(1): 25-9.
- [3] Auerbach, D., et al., *The post-genomic era of interactive proteomics: facts and perspectives*. Proteomics, 2002. **2**(6): 611-23.
- [4] Barabasi, A.L. and R. Albert, *Emergence of scaling in random networks*. Science, 1999. **286**(5439): 509-12.
- [5] Bartel, P. and S. Fields, eds. *The Yeast Two-Hybrid System*. Advances in Molecular Biology. 1997, Oxford University Press.
- [6] Chen, J.Y. and J.V. Carlis, *Genomic Data Modeling*. Information Systems, 2003. **28**(4): 287-310.
- [7] Drees, B.L., et al., *A protein interaction map for cell polarity development*. J Cell Biol, 2001. **154**(3): 549-71.
- [8] Ferguson, S.S., et al., *Molecular mechanisms of G protein-coupled receptor desensitization and resensitization*. Life Sci, 1998. **62**(17-18): 1561-5.
- [9] Ito, T., et al., *Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins*. Proc Natl Acad Sci U S A, 2000. **97**(3): 1143-7.
- [10] Mering, C.v., et al., *Comparative Assessment of Large-scale Data Sets of Protein-protein Interactions*. Nature, 2002. **417**: 399-403.
- [11] Pruitt, K.D. and D.R. Maglott, *RefSeq and LocusLink: NCBI gene-centered resources*. Nucleic Acids Res, 2001. **29**(1): 137-40.
- [12] Rain, J.C., et al., *The protein-protein interaction map of Helicobacter pylori*. Nature, 2001. **409**(6817): 211-5.
- [13] Ravasz, E., et al., *Hierarchical organization of modularity in metabolic networks*. Science, 2002. **297**(5586): 1551-5.

- [14] Valencia, A. and F. Pazos, *Computational methods for the prediction of protein interactions*. *Curr Opin Struct Biol*, 2002. **12**(3): 368-73.
- [15] Wheelan, S.J., A. Marchler-Bauer, and S.H. Bryant, *Domain size distributions can predict domain boundaries*. *Bioinformatics*, 2000. **16**(7): 613-8.
- [16] Wuchty, S., *Scale-free behavior in protein domain networks*. *Mol Biol Evol*, 2001. **18**(9): 1694-702.
- [17] Yarmush, M.L. and A. Jayaraman, *Advances in proteomic technologies*. *Annu Rev Biomed Eng*, 2002. **4**: 349-73.

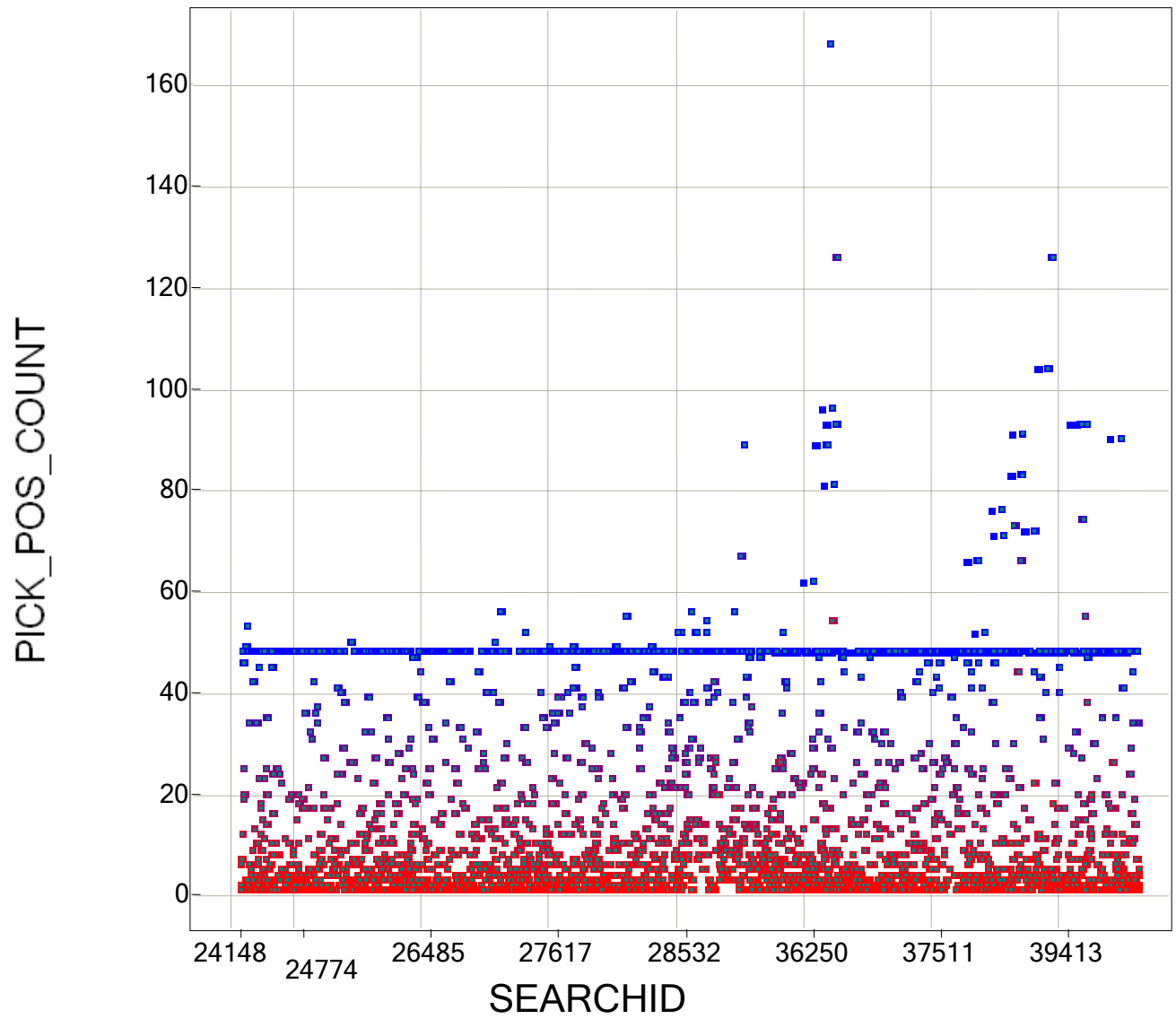


Figure 1. A scatter plot of “picked positive count” for all the ~17,000 non-*null* R2H searches performed on human brain libraries. The numbers along *x* axis are unique numbers, SEARCHID, identifying each R2H search. Pick_pos_count numbers along *y* axis are the count of positive clones picked in a search. Note the horizontal line of pick_pos_count = 48 indicated that at one decision point, we decided to pick at most 48 positive clones from each positive search.

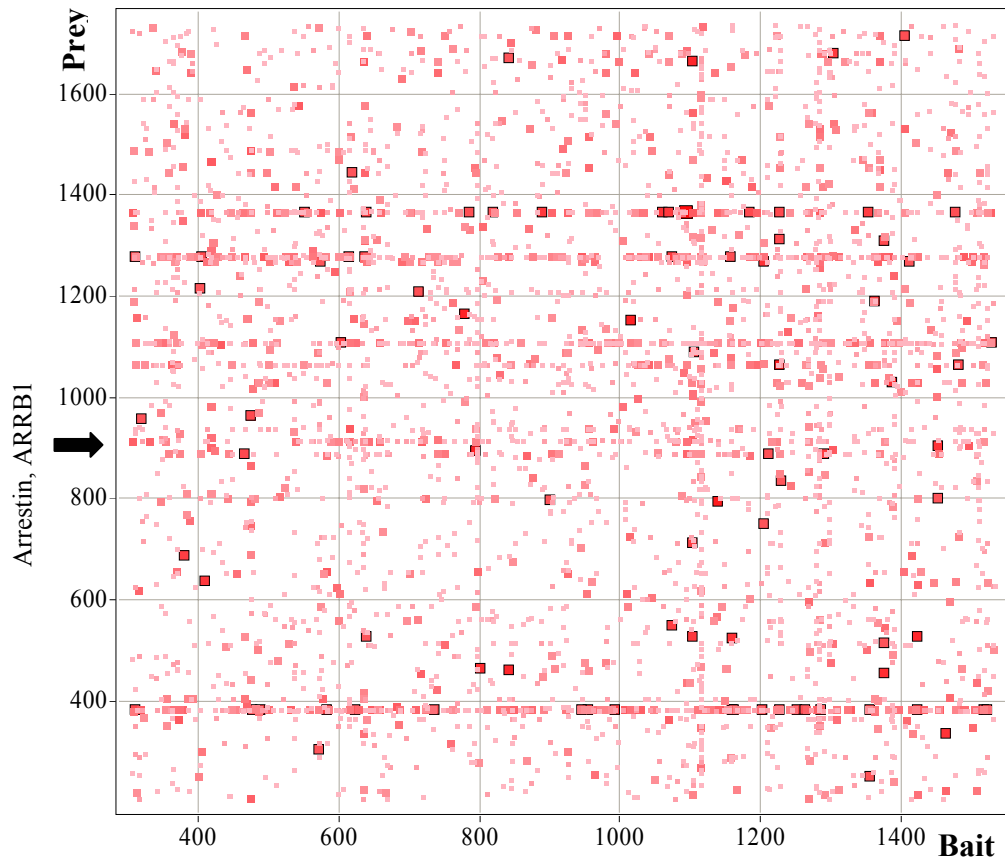


Figure 2. An interaction heat map showing interactions (dots) between a set of 1,200 bait proteins and a set of 1,600 prey proteins. The numbers along x and y dimensions are arbitrarily assigned protein identity labels (not REFSEQ accession numbers). The size and color intensity of a dot at (x, y) in the plot represent, in logarithmic proportion, the observed interaction frequency between bait protein x and prey protein y .

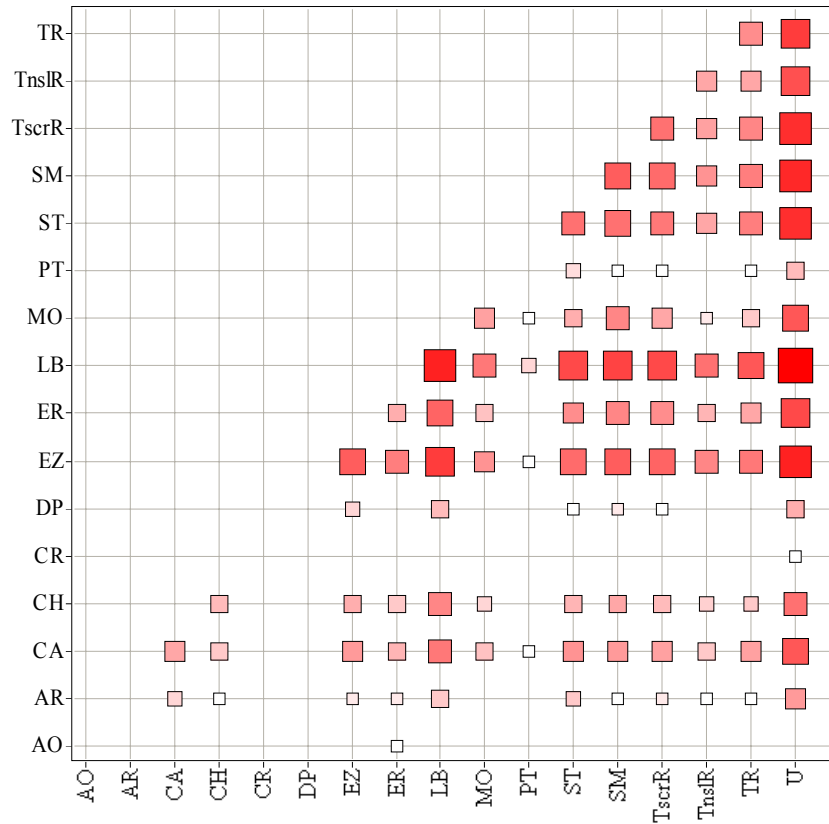


Figure 3a. Observed interactions binned into pairs of GO annotation terms in Molecular Function category at level 2 in the GO hierarchy. The size and color saturation of the squares are proportional to the logarithm of the number of interactions falling into the given term pair (the lowest and the largest numbers being 1 and 3946 respectively). The axis labels are: (AO) antioxidant, (AR) apoptosis regulator, (CA) cell adhesion molecule, (CH) chaperone, (CR) cytoskeletal regulator, (DP) defense/immunity protein, (EZ) enzyme, (ER) enzyme regulator, (LB) ligand binding or carrier, (MO) motor, (PT) protein tagging, (ST) signal transducer, (SM) structural molecule, (TscrR) transcription regulator, (TnsIR) translation regulator, (TR) transporter, and (U) molecular function unknown.

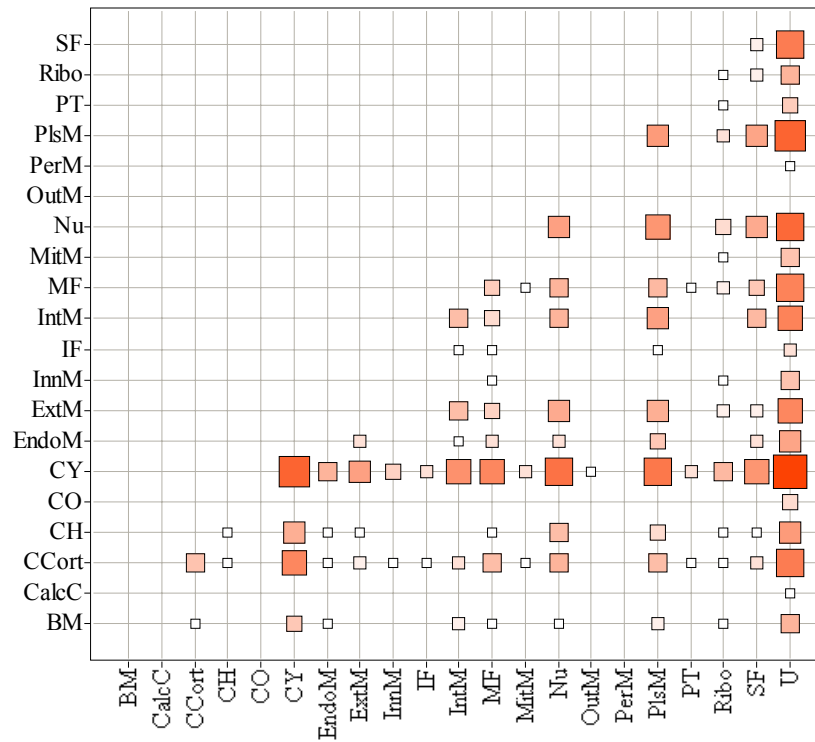


Figure 3b. Observed interactions binned into pairs of GO annotation terms in Cellular Component category at level 5 in the GO hierarchy. The axis labels are: (BM) basement membrane, (CalcC) calcineurin complex, (CCort) cell cortex, (CH) chromosome, (CO) collagen, (CY) cytoplasm, (EndoM) endomembrane system, (ExtM) extrinsic membrane protein, (InnM) inner membrane, (IF) insoluble fraction, (IntM) integral membrane protein, (MF) membrane fraction), (MitM) mitochondrial membrane, (Nu) nucleus, (OutM) outer membrane, (PerM) peroxisomal membrane, (PlsM) plasma membrane, (PT) proton-transporting ATP synthase complex, (Ribo) ribonucleoprotein complex, (SF) soluble fraction, and (U) cellular localization unknown

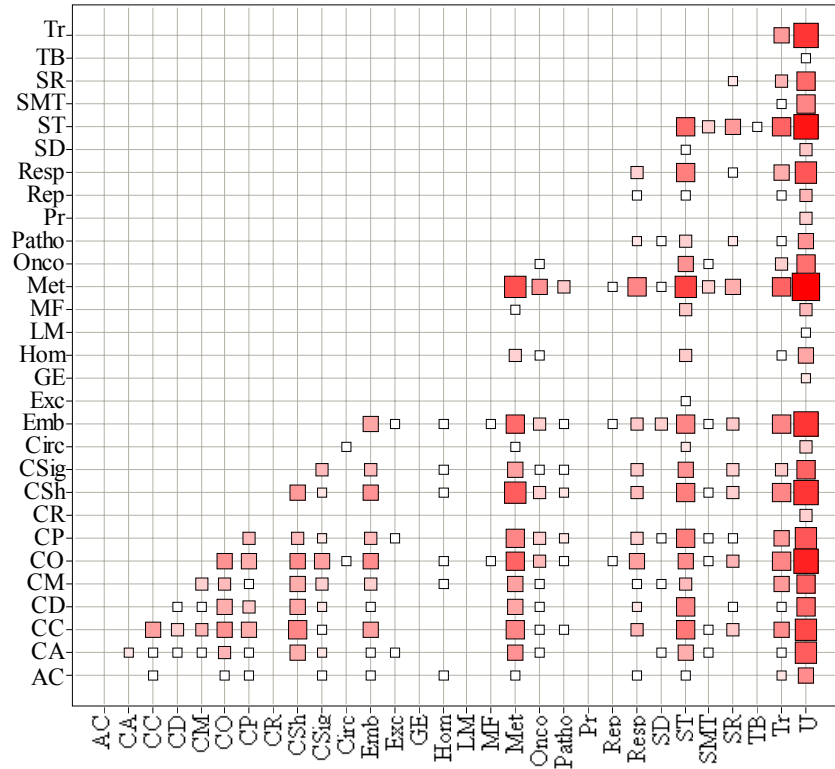


Figure 3c. Observed interactions binned into pairs of GO annotation terms in Biological Process category at level 4 in the GO hierarchy. The axis labels are: (AC) actin cytoskeleton reorganization, (CA) cell adhesion, (CC) cell cycle, (CD) cell death, (CM) cell motility, (CO) cell organization and biogenesis, (CP) cell proliferation, (CR) cell recognition, (CSn) cell shape and cell size control, (CSig) cell-cell signaling, (Circ) circulation, (Emb) embryogenesis and morphogenesis, (Exc) excretion, (GE) genetic exchange, (Hom) homeostasis, (LM) learning and memory, (MF) membrane fusion, (Met) metabolism, (Onco) oncogenesis, (Patho) pathogenesis, (Pr) pregnancy, (Rep) reproduction, (Resp) response to external stimulus, (SD) sex determination, (ST) signal transduction, (SMT) small molecule transport, (SR) stress response, (TB) telomere binding, (Tr) transport, and (U) biological

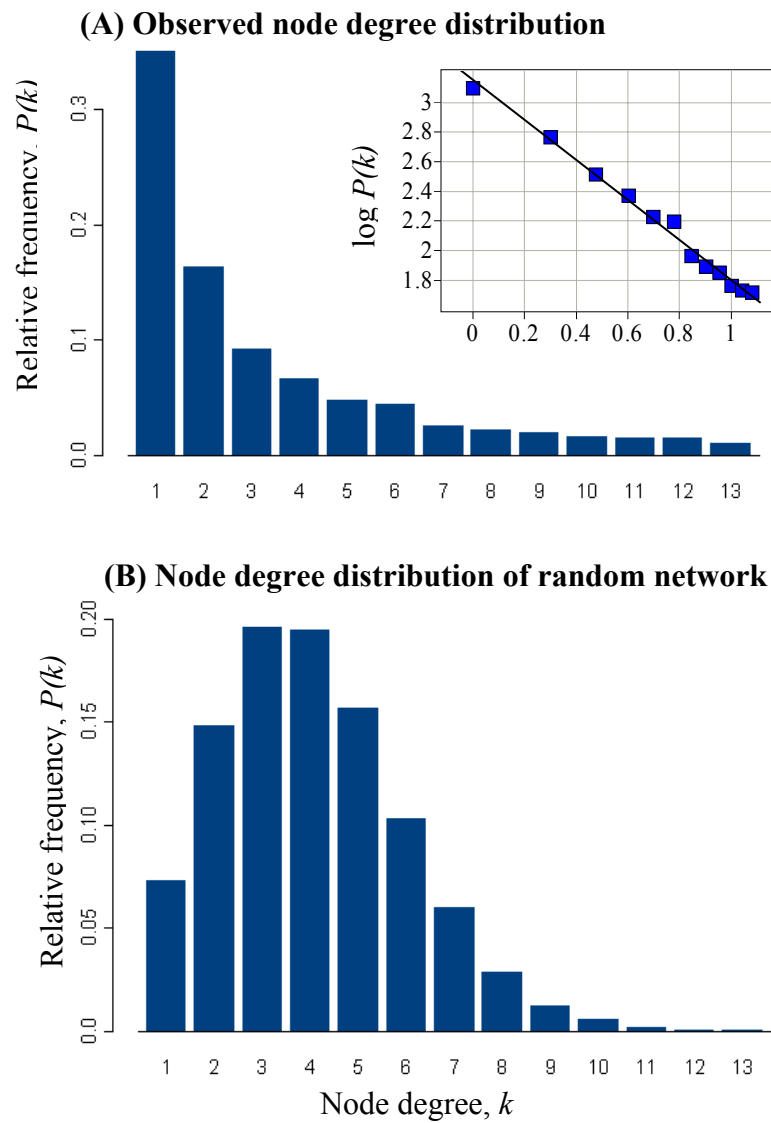


Figure 4. Node degree distribution $P(k)$ of (A) observed protein-protein interaction subnetwork (inset: the same distribution in log-log scale with best linear fit); (B) random network (see text).