High-throughput Protein Interactome Data: Minable or Not

Jake Y. Chen
Indiana University School of Informatics
Purdue School of Science Department of Computer and Information Science
Indianapolis, IN 46202
jakechen@iupui.edu

Andrey Y. Sivachenko
Prolexys Pharmaceuticals, Inc.
2150 W. Dauntless Ave
Salt Lake City, UT 84116
asivache@prolexys.com

Lang Li
Division of Biostatistics
Indiana University School of Medicine
Indianapolis, IN 46202
lali@iupui.edu

ABSTRACT

There is an emerging trend in post-genome biology to study the collection of thousands of protein interaction pairs (protein interactome) derived from high-throughput experiments. However, high-throughput protein interactome data, especially when derived from the Yeast 2-Hybrid (Y2H) method, have been generally believed to be irreproducible and unreliable, with an estimated high "noise ratio" of more than 50%. In this work, we performed a comprehensive study on approximately 70,000 protein interactions derived from a systematic yeast 2-hybrid (SY2H) method. We performed a comprehensive analysis of biases, reproducibility, statistical significance, and biologically significant patterns in this data set. Surprisingly, we found these protein interactions have a much higher quality. The data represented a comprehensive survey of the entire human proteome with no chromosomal location bias. The reproducibility rate of interactions among replicated searches was quite good, i.e., at 78.5%. The false positive rate, 5.5e-5, was two orders of magnitude better than that reported elsewhere. We further developed several statistical measures and concluded that a protein interaction only needs to appear in two different SY2H searches to become significant. We also developed techniques to show supporting evidence that "promiscuous" protein interactions were not random noises; instead, they could be "network hubs" of the cell signaling network. We also attributed the low noise in our data to the adoption of standard control in the experimental data generation process.

Keywords

Protein Interaction, Systematic Yeast 2-Hybrid, Reproducibility, Significance, Data Mining.

1.INTRODUCTION

In post-genome systems biology, the study of **protein interactomes**—comprehensive collections of all the expressed proteins and their interactions within cells of model organisms, has gained increasing popularity. Several protein interactome mapping projects, including those of *H. pylori* [1], *S. cerevisiae* [2, 3], *D. melanogaster* [4], *H. sapiens* [5], and *C. elegans* [6], have reported significant progress in recent years. In these projects, novel high-throughput experimental techniques, e.g., high-throughput yeast 2-hybrid (Y2H) screenings [7], protein arrays, and mass spectrometry, have been developed to measure physical bindings between proteins in parallel. This results in a steady influx of protein interaction data in the public domain. By understanding how proteins regulate each other through interaction, biologists can compile novel molecular pathway models, which they cannot normally derive from genomics techniques. The collection of thousands of protein interactions also enable system biologists to understand protein functions in a molecular network context, through which they may identify protein biomarkers or drug targets for diagnosing and treating human genetic diseases [8].

Nonetheless, there is a prevalent belief among many researchers that experimental protein-protein data generated from the high-throughput Y2H method equate to "high errors" and "poor reproducibility". Much doubt about Y2H data might have originated from a comparative analysis by Mrowka *et al* [9], who suggested that high-throughput Y2H experiments may have a false positive rate of greater than 50%. In a similar study, Bader *et al* analyzed high-throughput protein interaction data obtained from several sources and also concluded that these methods do not show enough internal consistency to warrant complete acceptance of the result [10]. Even more grim opinions exist [11]. Whether perceived or real, the high data "noise" has presented immense challenges for computational scientists to "mine" for biologically significant protein interactions and for biologists to trust data mining results from these efforts. Therefore, an imminent question for any researcher who will study the protein interactome data becomes,

(1) Can I trust the high-throughput protein interactome data at all?

(2) If so, how do I mine for significant protein interactions?

In this work, we restore confidence in high-throughput protein interactome data and the mining efforts, by investigating the biases, reproducibility, statistical significance, and functionally significant patterns of a human protein interactome data set. This data set consists of approximately 7,500 human proteins and 70,000 protein interactions, which was generated from a high-throughput **Systematic Yeast 2-Hybrid Method** (**SY2H**, refer to the Method section) [5]. Some of us have been curating and applying this data to biological discoveries for two years. We will show that by using a systematic method (SY2H), in which experimental conditions are enforced by standard protocols and the same robots, one can achieve reasonably good data reproducibility, keep false positive rate low, design reliable statistical hypothesis tests, discover statistically significant "interaction network hub proteins", and identify biologically significant interacting protein groups. We also show that "promiscuous" protein interactions should perhaps be regarded as "network hubs" instead of random noises—another explanation for the discrepancy between our analysis and the widely-held beliefs elsewhere. Our results may restore the confidence in similar high-throughput protein interactome data sets, and promote their application in subsequent molecular function studies. In Table 1, we have summarized some key features of the SY2H method by comparing it with the standard Y2H method. For a detailed description of this method, refer to the next section.

Table 1. A summary of comparisons between two Yeast 2-Hybrid (Y2H) methods. Refer to the Method section for an explanation of "baits", "preys", "searches", and "positives".

	Standard Y2H	Systematic Y2H
Bait Known Prior to a Search	Yes	No
Bait Sequence Enlisted in a Search	Whole or partial sequences	Short sequence fragments
Bait/Interaction Selection Bias	Yes (by design)	No (random sampling)
Possible Replicated Preys in a Search	Yes	Yes
Possible Replicated Same-bait Searches	No	Yes
Sequences to be Identified from Positives	Prey only	Bait and Prey
Global Assessment of Interactions	No	Possible

2.METHODS

Systematic Yeast 2-Hybrid (SY2H). First, two Y2H cDNA libraries from cDNA library samples from an organism are prepared using random internal primers. The hybrid proteins, which are derived by fusing a sample cDNA fragment with the yeast transcription factor DNA-binding domain or with the yeast transcription factor activation domain, are called "*bait*" and "*prey*", respectively. Second, haploid yeast bait and prey cDNA libraries are isolated into individual colonies, each containing a single bait or prey. Third, two types of haploid yeast cultures are mixed, one containing single bait colonies and the other containing colonies of the entire prey library, to allow mating to happen. Each such an experiment is called a "*search*". Fourth, mated diploid yeast cultures are placed on dishes that contain selective medium, which allows the mated yeast to grow only if bait and prey interact. Each grown diploid yeast colony is called a "*positive colony*", or a "*positive*". Fifth, up to a certain number of positive colonies is selected for picking ("*picked positives*"). Positives that are not picked are discarded. Sixth, DNA sequences from picked positives are amplified by PCR and DNA sequencing from both the 5' and 3' directions is performed. Seventh and lastly, interacting protein fragments are identified by comparing bait and prey DNA sequence fragments with annotated mRNA sequences from public sequence databases using the BLASTN software program.

Protein Interactome Data Collections. We collected the human protein interactome data from a high-throughput interactome mapping project using the above described SY2H system [5]. There were two major data collection milestones for this project in 2002-3. In the first major milestone, 13,656 unique protein interaction pairs were collected from approximately 50,000 SY2H searches against a prey cDNA library from mRNAs in homogenized human brain. This data set represented proteins from approximately 4,473 human gene loci, or approximately 5,000 unique proteins. In the second and a recent milestone in September 2003, approximately 70,000 unique protein interaction pairs were collected from more than 200,000 searches against a variety of human cDNA libraries. This interaction data set represented approximately 7,500 unique human proteins. We took a series of data snapshots between the milestones to perform our data analysis. The fact that we used slightly different data snapshots for each analysis is not a concern, since all these snapshots represented nearly random "samples" of the same protein interactome—a unique characteristic of the SY2H method (refer to Results).

Bioinformatics Data Analysis. We performed large-scale bioinformatics data analysis tasks to prepare and manage all the protein interaction data using Oracle9i server and genomic data modeling methods described in [12]. We integrated hundreds of gigabytes of biological data from more than 20 different sources. In particular, we integrated all the protein interaction pairs with public REFSEQ, LocusLink, and Gene Ontology annotations [13, 14]. In our data analysis, we used a combination of software tools, including the R statistical package and the Sportfire DecisionSite Browser for statistical data analysis and data visualizations. For this work, we developed several protein interaction data analysis methods, which we would describe along with the discussion of results next.

3.RESULTS

3.1 Comprehensive Protein Coverage and Bias

Due to the unique characteristics of the SY2H method and a homogenized human brain tissue library source, we expect to observe a wide spectrum of expressed proteins (a random sample of the entire "proteome") and interactions between them in the data. In principle, the data should represent a comprehensive survey of the entire human proteome with little sampling bias. In Figure 1, we confirmed this expectation by showing a relative frequency distribution for a snapshot of 5,619 proteins, binned by their chromosomal locations. While the relative distribution of all human REFSEQ proteins varies among different chromosomal and mitochondrial locations, the interacting proteins follow the varying distribution details quite well. Therefore, we can draw two inferences from this analysis. One is that the SY2H method indeed does a good job of randomly sampling the entire human proteome with no bias in coverage. The other is that all human proteins from different chromosomes and mitochondrion (perhaps except for chromosomes 6, 12, X, and Y) seem to share the same tendency to interact with each other.

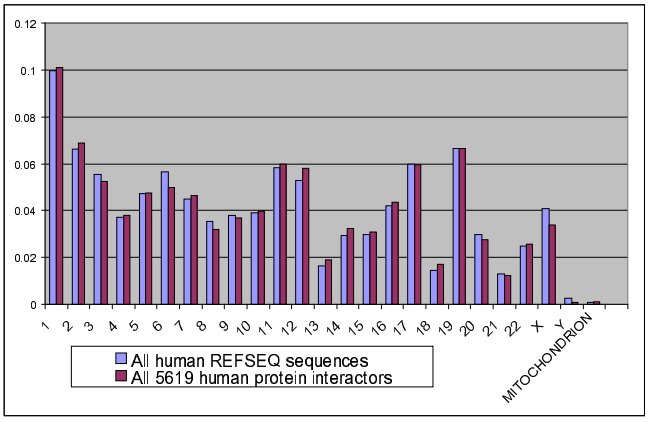

Figure 1. A comparison of the relative frequency distributions between all human REFSEQ sequences and all interacting proteins from the SY2H system, binned by their chromosomal and organelle locations.

Does a comprehensive coverage or a random sampling of the proteome suggest that there should be no bias in whichever proteins may become recruited in interactions? Not at all. If so, proteins of all 3-dimensional shapes would have interacted with each other equally. In Table 2, we showed an example of observed biases based on protein functional categories. Here, we listed eight protein functional categories. In each category, we listed a count of all human proteins from the LocusLink database, a count of interacting proteins identified with the SY2H method, and a percentage of coverage of identified protein for the category. In the last row of the table, we also showed several sums. This data shows that there were 18% of all 33,673 human proteins—a snapshot of 6,213 proteins derived from the SY2H system. However, "protein phosphatases" and "protein kinases" (CLASS I proteins) are highly enriched, at 29% and 26% respectively; "receptor" and "receptor : GPCR" proteins (CLASS II proteins), however, are scarce, at 6% and 5% respectively. We attribute this finding to a possible high **functional bias** towards proteins playing essential functional roles. For example, compared with other proteins, catalytic activities of enzymes (CLASS I proteins) are more frequently modulated by regulatory proteins through protein interactions; therefore, we observed an enrichment of CLASS I proteins. CLASS II proteins are poorly represented perhaps for a different reason--Y2H methods usually cannot capture protein interactions among membrane proteins (most CLASS II proteins).

Table 2. A breakdown of protein counts according to their functional categories.

	All Human Proteins from LocusLink	*Interacting Proteins Identified by SY2H*	*Percentage of Coverage*
Protein phosphatase	240	70	29%
Protein kinase	400	102	26%
Polymerase	161	39	24%
Transcription factor	372	77	21%
Channel protein	339	65	19%
Protease	233	33	14%
Receptor	3,294	203	6%
Receptor: GPCR	705	38	5%
Total	33,673	6,213	18%

3.2 Data Reproducibility

We assessed the reproducibility of interaction data derived from the SY2H system, and found it to be surprisingly good. For **reproducibility**, we refer to the capability of a high-throughput interaction discovery system to identify true interactions consistently. In Figure 2, we show that interaction reproducibility, calculated as the percentage of all interactions that can be replicated across different SY2H searches, is 78.5%. Comparing the our SY2H system with a standard Y2H system, we think it is possible that the reproducibility rate estimated in previous publications (from 10% to 50%) [11] were based collections of high-throughput data generated from different academic labs without the setup of robotic machineries for consistent controls. There is also a lack of report on replicated protein interaction pair data from public sources.

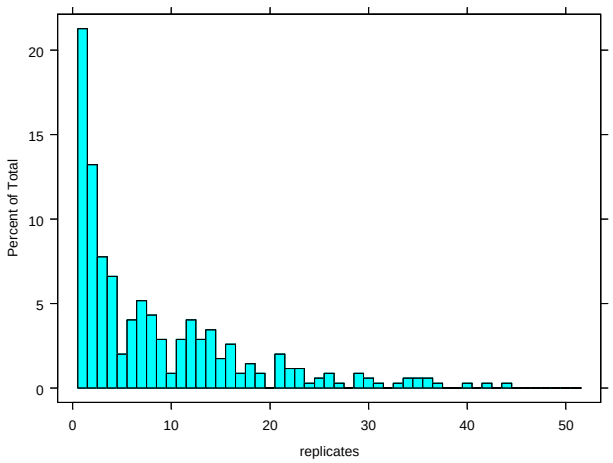

Figure 1. A relative frequency distribution of protein interaction "replicates". A "replicate" bin at x=1 indicates the percentage of interactions (21.5%) that are identified only once. All other "replicate" bins with x > 1 refer to the percentage of true replicated interaction (1-21.5%=78.5%) with an interaction replication count of x. The protein interaction data in this graph come from a random sample of 513 bait proteins, each of which is identified in at least two separate SY2H searches.

This 78.5% may still be an under-estimate of the true data reproducibly level for an SY2H system. This is because identifying protein interactions from searches is also a sampling process, in which the robots often pick a dozen top "positive colonies" for DNA amplification and sequencing. Therefore, one may not have exhaustively identified all replicated protein interactions from replicate searches. In other words, we expect the relative percentage for the "replicates"=1 bin becomes smaller than the 21.5% when the size of data increases.

3.3 Statistical Significance of Interactions

To identify statistically significant protein interactors (as preys) and protein interaction pairs (as bait-prey pairs), we describe a statistical data testing framework. First, we present a **null hypothesis**, in which we presume that the interactions happen randomly among all interactors. Therefore, the rate of interaction discovery, *p*, can be estimated by the following:

$$p = \frac{I}{N * M}, \qquad (1)$$

Here, *I* is the total number of observed unique interaction pairs, *N* is the total number of searches performed, and *M* is the total number of observed unique preys. To calculate *p*, for example, using a data snapshot taken from the first milestone (refer to Methods), we have *I* = 13,660, *N* = 50,000, *M* = 5,000, and therefore *p* = 13,660/50,000/5,000 = 5.5*e*-5. Similarly, using a recent data snapshot taken from the second milestone, we have *I* = 70,000, *N* = 200,000, *M* = 7,000, and therefore *p* =

70,000/200,000/7,000 = 5.0*e*-5. The two estimates are very close to each other.

We interpret *p* as an upper-bound estimate of the false positive rate for protein interactions observed in an SY2H system. There are two reasons for us to believe that *p* may be conservatively estimated. First, many of these observed *I* interactions may not be discovered totally by chance (recall functional bias); therefore, *p* should be smaller. We choose to treat interactions as "random" events, also because we do not have sufficient "negative" control interaction data set, i.e., a set of known non-interacting protein pairs. Second, the estimated prey number, *M*, may be higher than we currently used, because a high-throughput SY2H system sometimes may fail to amplify and identify a DNA sequence. The conservative estimate of a *p* at 5.5*e*-5 is good, because we can be confident later that the calculation of *p*-values, which we base on *p*, will be reliable indicators of statistical significances for replicated interactions.

Note, however, much higher false positive rates have been estimated for several standard high-throughput Y2H systems. For example, Gavin *et al* [15] reported a *p=1.07e-3* in their recent study, and Ho et al also reported a *p=1.37e-3* [16]. We attribute the much smaller false positive rate for the SY2H system to the high data reproducibility described in an earlier section.

Next, we describe two **hypothesis test methods**. In both methods, we calculate the *p*-values, one for observing multiple preys interacting with the same bait, and the other for observing the same bait-prey interactions multiple times, given that the null hypothesis is true, i.e. interactions to happen randomly. Our methods are different from a Bayesian method recently developed by Gilchrist *et al* [17], in which only the bait-prey protein interaction hypothesis was discussed. Instead, our hypothesis test methods belong to a "frequentist method", which have the advantage of not requiring a prior protein interaction distribution for an alternative hypothesis.

In the first test method, we are concerned with whether a particular bait tends to interact with many different preys. We want to distinguish whether a bait protein "indiscriminately" chooses an interaction partner by chance or by an un-characterized statistically significant process. In this model, we use r $(r \geq 1)$ to indicate the number of times that a search is replicated, i.e., the number of times the same bait has been observed in different searches. We use l to indicate the number of preys from all the replicated searches sharing the same bait. We use p and M according to the previously described definition. Under the null hypothesis, we assume that every prey-bait interaction is an independent Bernoulli trial with a success rate of p. There are $r \times M$ trials among r replicated searches. Therefore, the probability to obtain l or more preys by chance among r searches (p-value) can be calculated through a binomial distribution,

$$pvalue_{IWTE} = \Pr(L \geq l) = 1 - \sum_{i=0,1,\ldots,l-1} \binom{r \times M}{i} p^i (1-p)^i \qquad (2)$$

Where $pvalue_{\mathrm{IWTE}}$ is the individual-wise type I error for a bait. However, as a total of *N* searches were performed, and each bait is assumed to have *r* replicated searches, the family-wise type I error can be controlled in (3) (Westfall),

$$pvalue_{FWTE} = \Pr(L \geq l) = 1 - \{1 - pvalue_{IWTE}\}^{N/r} \qquad (3)$$

In Table 3, we tabulated the *p-values* under all the scenarios, ranging in l=1 to 10 and r=1 to 4, where N=200,000, M=7,000, and p=5.0e-5. Each cell in the table contains a "family-wise" p-value, which measures the significance level for discovering l preys in r replicated searches. For example, when only six interactions or less are discovered in a non-replicated search (r=1 and l=6), this observation is not significant since p-value=0.314. However, when l increases from 6 to 7, 8, and ≥9 for a fixed r=1, the p-value decreases to 1.86e-02, 8.15e-04, and ≤3.16e-05 respectively, suggesting the data being increasingly significant. This table also confirms that for a fixed l number of preys, the less search replications r it takes to observe all of them, the more significant the observation becomes.

Table 2. A list of P-value that measures the significance of observing *l* number of preys in *r* different searches. The scenarios that are significant at a *p-value* threshold of ≤ 0.05 are highlighted by shade and a bold font.

	r=1	*r =2*	*r=3*	*r=4*
l =6	3.14e-01	1.00e-00	1.00e-00	1.00e-00
l =7	**1.86e-02**	5.88e-01	1.00e-00	1.00e-00
l =8	**8.15e-04**	7.39e-02	6.19e-01	9.95e-01
l =9	**3.16e-05**	**5.90e-03**	1.05e-01	5.57e-01
l =10	**1.10e-06**	**4.11e-04**	**1.15e-02**	1.06e-01
l =11	**3.49e-08**	**2.60e-05**	**1.09e-03**	**1.40e-02**
l =12	**1.02e-09**	**1.51e-06**	**9.48e-05**	**1.63e-03**

If the statistical significant level is set as 0.05 the family-wise p-value, there are many significant conclusions that we can derive from this test result. For example, we can conclude that if we observe at least 7 preys interacting with a single bait in any search, the event is statistically significant (p-value=0.0186). For another example, if the same bait has appeared 3 times in different searches, we have to observe an additional 3 preys for these 10 (=7+3) preys to be taken as statistically significant (p-value=0.0115). In a final example, if a bait interacts with hundreds of other preys in a few different SY2H searches, according to the above result, we say the bait must be selected with a significant bias. This result supports many earlier findings of "sticky proteins" and highly interacting proteins serving as "interaction network hubs" [5, 18].

In the second test, we are concerned with the significance of identifying a protein interaction pair from experimental results. We want to know whether or not a protein interaction can be "trusted" for use in subsequent knowledge discovery tasks. We use t to indicate the number of times that a prey is discovered, and use r, p, and M as described early in this section. Using a binomial model, we can calculate the individual-wise type I error, $pvalue_{IWTE}$, for seeing the same prey appearing t times by chance in r replicated searches as the following:

$$pvalue_{IWTE} = \Pr(T \geq t) = 1 - \sum_{i=0,1,\ldots,t-1} \binom{r}{i} p^i (1-p)^i \qquad (4)$$

However, as all M preys are available, the family-wise type I error can be controlled in (5) (Westfall),

$$pvalue_{FWTE} = \Pr(L \geq l) = 1 - \{1 - pvalue_{IWTE}\}^{M} \quad (5)$$

In Table 4, we tabulated the *p*-values under scenarios for t=1 to 4 and for r=1, 2, 3, 4, where N=200,000, M=7,000, and p=5.0e-5. Each cell in the table contains a family-wise type I error *p*-value, which measures the significance level for discovering the same bait-prey interaction for t times in r replicated searches. For example, when an interaction is discovered only once in a single non-replicated search (r=1 and t=1), this observation is not significant since *p-value*=0.295. However, any replicated interaction identified from at least two different SY2H searches ($r \geq 2$ and t=2 to 4) is going to be significant, because the calculated p-value in all these scenarios are less than 0.05.

Table 3. A list of P-value that measures the significance of observing the same interaction pair *t* times in *r* different searches. The scenarios that are significant at a *p-value* threshold of ≤ 0.05 are highlighted by shade and a bold font.

	r=1	*r=2*	*r=3*	*r=4*
t =1	2.95e-01	5.03e-01	6.50e-01	7.53e-01
t =2	--	**1.74e-05**	**5.25e-05**	**1.05e-05**
t =3	--	--	**8.75e-10**	**3.50e-09**
t =4	--	--	--	**0.00e-00**

3.4 Biological Significance of Interactions

Following the discovery of statistically significant patterns in the "raw" data set , the next question arises, "How does one identify biologically significant protein interactions?" Not all statically significant interactions discovered in the previous section are biologically sensible, because a falsely identified human interacting protein can appear in many searches simply because this protein interacts with the yeast transcription factor in the SY2H system. To address this issue eventually, significant research efforts beyond this work is necessary, including efforts to perform complementary or validation experimental studies, conduct manual knowledge curations, and incorporate different types of biological data into the current computational analysis. In [19], we summarized the challenges and opportunities of integrating biological data such as gene expression information, functional annotations, homology information, and interaction network modules.

In this section, we describe an example of such integrative data analysis based on the annotations of protein's interaction partners. The null hypothesis is that proteins do not have specific functional or localization preferences when choosing their interaction partners. To collect statistics under the null hypothesis, we used a numerical re-sampling method, in which we randomly rewired the interaction network. Our randomization procedure, however, preserved each node's degree of connectivity and thus the overall network node degree distribution. For each randomly rewired network, we retrieved the interaction partners $v(n)$ of each protein n and calculated the frequency of occurrences of each annotation term among all the annotations, $A[v(n)]$, available for the proteins in $v(n)$. From this data, we computed the distribution functions for the fractions of each annotation term among all the terms assigned to protein's interaction partners. Since we were interested in statistically significant *co-occurrences* of the annotation terms, we actually calculated a conditional probability, $p[|N(t)|=k|t]$, to observe k occurrences of term t among $A[v(n)]$, given $t \in A[v(n)]$. The continuous approximation (calculating fractions instead of counts) is helpful for analyzing very small number of proteins with very large number of interaction partners, which would otherwise require expensive full network re-sampling to analyze.

Table 4. A summary report showing that highly interacting proteins, binned by their node degree range, have significantly high shares of interaction partners in diverse annotation categories (sampled).

Node Degree Range	20-30	31-40	41-80	>80
Development	24	26	22	26
Chaperone activity	15	7	15	25
Catalytic activity	48	16	4	2
Transporter activity	37	21	19	13
Motor activity	12	9	23	27
Signal transducer	42	16	14	15
Translation regulator activity	9	10	15	28
Extra-cellular	44	30	25	30
Enzyme regulator activity	14	15	6	9
Transcription regulator activity	24	32	29	27
Structural molecule activity	24	19	44	118
Defense/immune activity*	5	2	2	0
Cell adhesion molecule activity*	24	23	21	12
Apoptosis regulator activity*	7	2	5	1
Significant / Total	239/596 =40%	177/269 = 66%	188/398 = 47%	196/281 =70%

***** This term is obsolete in the current version of GO. Our analysis is still consistent since we used the concurrent versions of GO and the protein annotation mapping.

For annotations we used a vocabulary derived from the Gene Ontology (GO) database. Since the GO has directed acyclic graph quasi-hierarchical structure, for each annotation term we perform a "roll-up" similar to [5] by tracing the term back to all its "ancestors" at the GO level l=2 (l=*0* is for the *root*, l=*1* is for "molecular function", "biological process", and "cellular component" labels). With this operation we generally avoid the problem of many terms being too narrow and not sufficiently represented to enable building robust statistics. It is also plausible biologically to expect multiple interactions with related proteins, belonging to the same general group (*e.g.* 'structural' or

'transcription factor'), rather than with a number of same very specific functional modules.

In Table 5, we present a summary of our results. Here, we are primarily interested in characterizing potentially self-activating and "sticky" false p ositive proteins—**highly interacting proteins** that we define here as those having >20 interaction partners. We selected four node degree ranges (20-30, 31-40, 41-80, and >81) and calculated significance levels for each annotation term according to the method outlined above. For instance, in the group of proteins with 20 to 30 interaction partners (first row) that consists of 596 proteins (see the "Significant/Total" column), there are 24 proteins that interact with significantly high numbers of proteins annotated with the "Development" GO term (or its more specific descendants), 15 interact with the significantly high numbers of proteins involved in "Chaperon Activities" *etc*. Total of 239 (40% of 596) proteins in this group have at least one annotation term overrepresented among their interaction partners (note that many proteins have more than one significant term in *A(v(n))*). From our result, we can conclude that 70% of the **promiscuously interacting proteins** (node degree >80) have statistically significant interaction patterns and thus can be biologically significant and active "functional hubs". Combined with the evidence from previous work and previous results in this work, we believe that they should not be recklessly dismissed, but rather thoroughly analyzed with all the biological evidence available. It should be noted that some proteins are involved in various activities under different conditions, so that when the whole set of interaction partners is analyzed regardless of the source tissue, developmental stage, *etc.*, no particular functional category may seem to be overrepresented. Thus, our estimates can be conservative.

4.DISCUSSION

In this study, we performed a comprehensive assessment of the human protein interactome data, which were derived from the SY2H method. We showed that this data set comprehensively surveyed the human proteome without an apparent bias in source chromosomal locations. We also showed that the data had a good reproducibility above 78% and a low false positive rate at approximately 5.5e-5. We developed several statistical data mining techniques to assess both the statistical and biological significance of interactions, especially for those data with replications and annotated GO term labels. We showed evidence were not random noises; instead, they could be "network hubs" of the cell signaling network. We also attributed the low noise in our data to the adoption of standard control in the experimental data generation process.

Several factors may have prevented similar insights into the protein interactome data "noise" issue from being developed until this work. First, protein interactome data were scarce until very recently. Not many researchers have access to this type of data; even fewer have first-hand experience trying to extract useful information from it. Second, there is a genuine lack of biological understanding in how Y2H method works. This may have presented the development of new concepts such as protein "network hubs". Third, many public data are produced in labs without the modern robot equipments or proper enforcement of standard operation protocols. Therefore, experimental variations are prevalent. Fourth, even today, the public protein interactome data set does not contain essential information such as protein interaction regions, interaction strengths, or replicated interactions under similar experimental conditions. With an ongoing surge of protein interactome data in the next few years, we hope our early results will restore some assurance to forthcoming data miners of this information.

5. REFERENCES

[1] Rain, JC, et al., *The protein-protein interaction map of Helicobacter pylori.* Nature, 2001. **409**(6817): p. 211-5.

[2] Ito, T, et al., *A comprehensive two-hybrid analysis to explore the yeast protein interactome.* Proc Natl Acad Sci U S A, 2001. **98**(8): p. 4569-74.

[3] Uetz, P, et al., *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.* Nature, 2000. **403**(6770): p. 623-7.

[4] Giot, L, et al., *A protein interaction map of Drosophila melanogaster.* Science, 2003. **302**(5651): p. 1727-36.

[5] Chen, JY, et al. *Initial Large-scale Exploration of Protein-protein Interactions in Human Brain*. in *IEEE CSB Bioinformatics 2003*. 2003. Stanford, California.

[6] Li, S, et al., *A map of the interactome network of the metazoan C. elegans.* Science, 2004. **303**(5657): p. 540-3.

[7] Bartel, P and S Fields, eds. *The Yeast Two-Hybrid System*. Advances in Molecular Biology. 1997, Oxford University Press.

[8] Auerbach, D, et al., *The post-genomic era of interactive proteomics: facts and perspectives.* Proteomics, 2002. **2**(6): p. 611-23.

[9] Mrowka, R, A Patzak, and H Herzel, *Is there a bias in proteome research?* Genome Res, 2001. **11**(12): p. 1971-3.

[10] Bader, GD and CW Hogue, *Analyzing yeast protein-protein interaction data obtained from different sources.* Nat Biotechnol, 2002. **20**(10): p. 991-7.

[11] Legrain, P, J Wojcik, and JM Gauthier, *Protein--protein interaction maps: a lead towards cellular functions.* Trends Genet, 2001. **17**(6): p. 346-52.

[12] Chen, JY and JV Carlis, *Genomic Data Modeling.* Information Systems, 2003. **28**(4): p. 287-310.

[13] Pruitt, KD and DR Maglott, *RefSeq and LocusLink: NCBI gene-centered resources.* Nucleic Acids Res, 2001. **29**(1): p. 137-40.

[14] Ashburner, M, et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.

[15] Gavin, AC, et al., *Functional organization of the yeast proteome by systematic analysis of protein complexes.* Nature, 2002. **415**(6868): p. 141-7.

[16] Ho, Y, et al., *Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry.* Nature, 2002. **415**(6868): p. 180-3.

[17] Gilchrist, MA, LA Salter, and A Wagner, *A statistical framework for combining and interpreting proteomic datasets.* Bioinformatics, 2004. **20**(5): p. 689-700.

[18] Hoffmann, R and A Valencia, *Protein interaction: same network, different hubs.* Trends Genet, 2003. **19**(12): p. 681-3.

[19] Chen, JY and AY Sivachenko, *Data Mining Challenges for Protein Interactomics Studies (accepted).* IEEE Magazine in Biology and Medicine, 2004.