32nd Annual International Conference of the IEEE EMBS
Buenos Aires, Argentina, August 31 - September 4, 2010

Unraveling the conundrum of seemingly discordant protein-protein interaction datasets

Shobhit Gupta, Anders Wallqvist, Rajkumar Bondugula, Joseph Ivanic, and Jaques Reifman

Abstract—**Most high-throughput experimental results of protein-protein interactions (PPIs) are seemingly inconsistent with each other. In this article, we re-evaluated these contradictions within the context of the underlying domain-domain interactions (DDIs) for two *Escherichia coli* and four *Saccharomyces cerevisiae* PPI datasets derived from high-throughput (yeast two-hybrid and tandem affinity purification) experimental platforms. For shared DDIs across pairs of compared datasets, we observed a remarkably high pair-wise correlation (Pearson correlation coefficient between 0.80 and 0.84) between datasets of the same organism derived from the same experimental platform. To a lesser degree, this concordance also held true for more general inter-platform and intra-species comparisons (Pearson correlation coefficient between 0.52 and 0.89). Thus, although varying experimental conditions can influence the ability of individual proteins to interact and, therefore, create apparent differences among PPIs, the physical nature of the underlying interactions, captured by DDIs, is the same and can be used to model and predict PPIs.**

I. Introduction

Protein-protein interaction (PPI) networks represent complex molecular relationships that broadly determine the activity of a functional cell. A comprehensive and accurate definition of various interactomes would, therefore, have enormous implications toward the understanding of cellular functions. In the past decade, high-throughput experiments to determine PPIs have been extensively pursued, buoyed by the introduction of yeast two-hybrid (Y2H) screens [1]. However, it is now well known that individual Y2H screens have low concordance with each other.

Recently, tandem affinity purification followed by mass spectrometry (AP-MS) techniques have been proposed as a new method to derive high-throughput PPI complexes [2]. Several PPI datasets based on the TAP-MS methodology are now publicly available. However, these datasets also record remarkably low overlap; only 209 interactions of a joint total of 15,810 interactions are shared.

In this article, we re-evaluated the consistency among PPI networks derived from both Y2H and TAP-MS experimental platforms in two organisms, *E. coli* and *Saccharomyces cerevisiae*, using protein domains. We considered the physical binding that occurs in a PPI as arising from interactions between domains and not the whole proteins *per se*. Therefore, we systematically broke down all proteins into domains and analyzed PPIs via domain components. In effect, the PPI data were converted into a set of domain-domain interactions (DDIs). The frequency of occurrence of domain pairs allowed us to estimate the propensity of domain interactions (association scores), and the resulting DDI profiles of association scores allowed us to evaluate the inherent similarities among sets of PPIs. By quantifying the similarity between DDI profiles, we demonstrated a significant, and perhaps underappreciated, coherence among the seemingly incoherent PPI datasets. Thus, even though pairs of datasets were highly discordant in terms of PPIs, the corresponding DDIs were highly consistent. Our findings have important implications in the perception of high-throughput PPI experiments and their application toward delineating an accurate and comprehensive picture of the interactome.

II. Methods

In this section, we briefly describe our approach to compare six high-throughput PPI datasets: two TAP-MS datasets for each of *E. coli* and *S. cerevisiae*, and two Y2H datasets for *S. cerevisiae*. To gauge the similarity of the underlying interactions in these datasets, we compared the DDI profiles between datasets.

A. PPI Datasets

We studied four datasets based on the TAP-MS experimental procedure: two *E. coli* datasets, described by Arifuzzaman et al. [3] and Butland et al. [4], and two *S. cerevisiae* datasets, described by Krogan et al. [5] and Gavin et al. [6]. In addition, we studied two datasets based on Y2H screens for *S. cerevisiae*, described by Uetz et al. [7] and Ito et al. [8], which were retrieved from the IntAct database [9].

To allow comparisons between different datasets, all protein identifiers for *E. coli* were converted into Swiss-Prot identifiers [10] (October 10, 2007 freeze) and all protein identifiers for *S. cerevisiae* were converted into open reading

The authors are with the Department of Defense Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Materiel Command, Fort Detrick, MD, USA (corresponding author: Jaques Reifman, Phone: 302-619-8130; Fax: 301-619-1983; E-mail: jaques.reifman@us.army.mil).

SG is currently with the Johnson School at Cornell University, Ithaca, NY. JI is currently with the Information Systems Program, Advanced Biomedical Computing Center, SAIC, Frederick, MD.

The authors were supported, in part, by the U.S. Army's Network Science Initiative through the Military Operational Medicine Research Program of the U.S. Army Medical Research and Materiel Command, Ft. Detrick, MD.

The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army or the U.S. Department of Defense. This paper has been approved for public release with unlimited distribution.

978-1-4244-4124-2/10/$25.00 ©2010 IEEE

TABLE 1. STATISTICS OF SUPERFAMILY DOMAIN DISCOVERY IN PROTEINS FROM DIFFERENT PPI DATASETS

Organism/Platform/Dataset	Number of Proteins	Number of Interactions	Number of Proteins with Domains, *n*(%)	Number of Distinct Domains	Number of Domains per Protein
S. cerevisiae/Y2H/Uetz	1,315	1,389	795(60)	333	1.27
S. cerevisiae/Y2H/Ito	3,241	4,367	1,792(55)	553	1.27
S. cerevisiae/TAP-MS/Gavin	1,441	6,918	1,049(73)	438	1.47
S. cerevisiae/TAP-MS/Krogan	2,694	7,089	1,502(56)	502	1.39
E. coli/TAP-MS/Arifuzzaman	2,942	11,161	1,969(67)	658	1.41
E. coli/TAP-MS/Butland	1,158	4,858	828(72)	475	1.52

Y2H, yeast two-hybrid screen; TAP-MS, tandem affinity purification followed by mass spectrometry.

frame identifiers using the Saccharomyces Genome Database [11]. Table 1 shows the number of proteins and interactions of the six PPI datasets after the removal of self-interactions and duplicate interactions. We did not distinguish interactions between the same pair of proteins but with different bait/prey roles and considered such interactions as a single interaction.

B. *Domain Annotation*

The downloaded protein sequences were scanned for Superfamily (version 1.69) [12] domains using the default E-value cutoffs of 0.02 implemented in InterPro (version 17.0) [13]. We used the high-performance computing pipeline for protein function annotation (PIPA) [14] to perform these computations.

C. *Construction and Comparison of DDI Profiles*

For a given pair of domains *x* and *y*, their association score (A_{xy}) was defined as the ratio of the number of observed interacting protein pairs containing domains *x* and *y* (O_{xy}) and the number of all possible protein pairs containing domains *x* and *y* (P_{xy}) [15], as follows:

$$A_{xy} = \frac{O_{xy}}{P_{xy}}. \qquad (1)$$

The association score A_{xy} was calculated for all possible pairs of domains of each dataset, and provided an empirical measure of the likelihood of two proteins to interact given that each protein contained one of the two domains. For a given dataset, each domain pair that was not observed in interacting protein pairs was assigned an association score of zero for that dataset. The matrix of association scores from Equation 1 constituted the DDI profile that was used to compare two PPI datasets.

We evaluated the concordance between any two DDI profiles by computing the Pearson correlation coefficient (*r* value) between the association scores A_{xy} for the set of domain pairs (*x*, *y*) in the two profiles. Because PPI datasets provide only positive instances of interactions and are largely incomplete, it was not possible to compute an "unbiased" correlation between the corresponding DDIs. Here, we use the term "unbiased correlation" to mean the correlation coefficient that would have resulted if an association score could be assigned to each pair of domains in both datasets, i.e., if there were no missing data. Since PPI datasets are largely incomplete, we estimated a range of correlations based on two different criteria for including domain-domain association scores in the computation. In *criterion I*, we only computed correlations for scores of $A_{xy} > 0$ in both DDI profiles being compared. $A_{xy} = 0.0$ in a given profile implied that domains *x* and *y* were not observed to interact in the particular dataset. Hence, the correlation coefficients attained under *criterion I* provided a measure of consistency between shared or commonly observed DDIs in two independent PPI datasets. In *criterion II*, we additionally considered association scores of $A_{xy} = 0$ in one of the profiles as long as both domains *x* and *y* were observed in the dataset and $A_{xy} > 0$ in the other profile. *Criterion II* represented the case where the interaction of a particular pair of domains *x* and *y* was observed exclusively in one dataset but the two domains were observed, albeit not interacting, in the other dataset. The correlation coefficients computed based on *criterion II* amplified the inconsistencies between the datasets. Notwithstanding the lack of a true gold standard for either PPIs or DDIs, these correlation coefficients provided a range of similarities based on the observed interactions.

III. RESULTS

In this section, we present the results of our domain-level evaluations of the six PPI datasets. We estimated the correlation coefficients between DDI profiles for different pairs of datasets to gauge the consistency of different datasets when analyzed at the domain level.

A. *PPI Datasets and Their Overlap*

The overlap between the three pairs of datasets corresponding to the same species and the same experimental platform (*S. cerevisiae* PPIs determined using Y2H, *S. cerevisiae* PPIs determined using TAP-MS, and *E. coli* PPIs determined using TAP-MS) revealed that, despite the large number of shared proteins between datasets, the actual number of shared PPIs was low. The Venn diagrams in Figure 1 show protein and PPI overlaps between the pairs of compared datasets. In all instances, even though a large fraction of proteins was common to both datasets, the overlap between interactions was low.

However, even within this small shared set of interactions, we observed biologically relevant and verifiable interactions. Figure 2 shows the sub-network containing only the interactions that were shared in the two (Butland/Arifuzzaman) E. coli PPIs determined using the TAP-MS technique. The complex highlighted in red identified the interaction map consisting of a number of

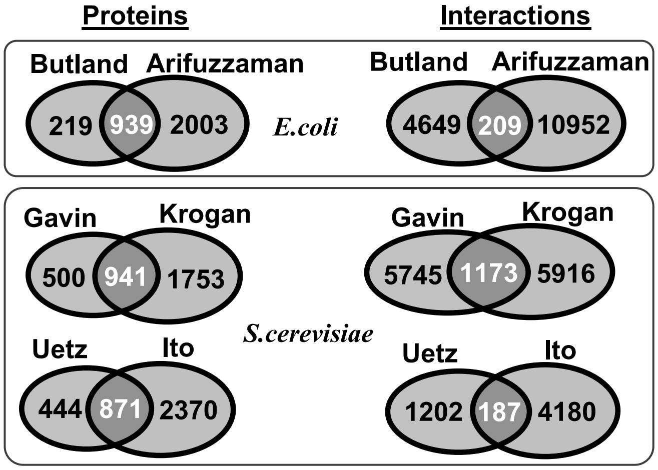

Figure 1: Shared proteins (*left*) and shared interactions (*right*) between the two *Escherichia coli* (*top*) and four *Saccharomyces cerevisiae* (*bottom*) protein-protein interaction (PPI) datasets. The Venn diagrams are not drawn to scale.

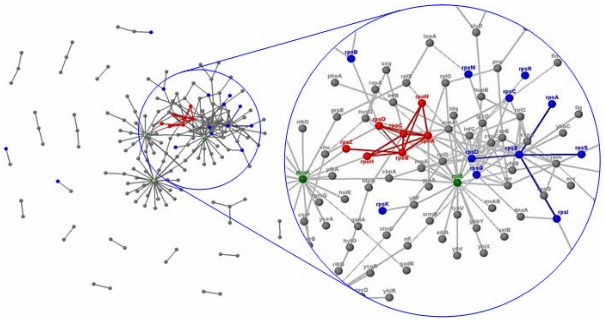

Figure 2: Protein-protein interactions (PPIs) shared by both the Butland and Arifuzzaman datasets in E. coli using TAP-MS revealed two biologically relevant PPI sub-networks. The first network, highlighted in red, corresponded to RNA polymerase subunits (rpoZ, rpoH, rpoB, rpoC, rpoA, rpoN, and rpoD). The second, slightly dispersed network, highlighted in blue, was composed of ribosomal proteins (rpsD, rpsE, rpsS, rpsA, and rpsI). Also, the chaperone/heat shock proteins groL and dnaK, highlighted in green, exhibited high connectivity or the ability to interact with many other proteins.

subunits of RNA polymerase. Another sub-network, highlighted in blue, shows interactions between ribosomal proteins that are known to interact with each other and form the functional ribosome during protein translation. Also, in agreement with other PPI datasets [9, 16, 17], the chaperone/heat shock proteins groL and dnaK, highlighted in green, exhibited high connectivity or the ability to interact with many other proteins.

B. Domain Discovery Statistics

Table 1 shows the statistics of the domains identified in *E. coli* and *S. cerevisiae*. On average, 64% of proteins had Superfamily domains. We also observed a higher number of domains per protein in the TAP-MS datasets for both *S. cerevisiae* and *E. coli* compared with the *S. cerevisiae* Y2H datasets, suggesting a relative bias in the TAP-MS experiments toward evaluating multi-domain proteins compared with the Y2H experiments.

C. Correlation With DDI Profiles

We computed DDI profiles using Superfamily domain definitions for each of the six PPI datasets. Table 2 shows the *r* values and the associated *P*-values for the 15 possible pair-wise comparisons between the 6 datasets. Using *criterion I*, we assumed that domain interactions observed in only one of two datasets being compared were false observations and exclusively compared DDIs that were shared between datasets. This scenario considered a smaller number of interactions (*N*) and led to higher correlations. The *top* panel in Table 2 shows the *r* values obtained with *criterion I*. We observed high *r* values ($0.52 \leq r \leq 0.89$) for all intra-species comparisons (unshaded cells), indicating that the binding propensities observed in different experimental techniques were comparable with each other. The *r* values were consistently high ($0.80 \leq r \leq 0.84$) for intra-species comparisons with the same experimental technique (marked in bold). For inter-species comparisons (shaded cells), the results obtained with the TAP-MS technique indicated good levels of correlation ($0.42 \leq r \leq 0.73$). Similar comparisons between *S. cerevisiae* Y2H and *E. coli* TAP-MS techniques were inconclusive (Table 2). The high correlation values obtained when using *criterion I* were not an artifact of special cases where a DDI occurs only once in both datasets, i.e., $O_{xy} = 1$ and $P_{xy} = 1$ in both profiles. Very few such cases were observed, and when these DDIs were excluded from the datasets we observed very similar *r* values (results not shown).

Criterion II necessarily considered a larger number of domain interactions (*N*) and yielded a smaller correlation than *criterion I*, because the zero and the corresponding non-zero association scores were included in the computation of the *r* values. As shown in the *bottom* panel in Table 2, *criterion II* yielded weak correlations between the *S. cerevisiae* DDIs derived from the same experimental technique (marked in bold). All other comparisons indicated a lack of correlation. These results represented the lower range of the correlations between DDI profiles.

We repeated these analyses using Pfam (version 21.0) [18] domain annotations of the proteins in the PPIs instead of Superfamily domains and obtained very similar trends in the correlations (results not shown).

Given the perceived notion that high-throughput measurements of PPIs tend to be biased and, in general, unreliable due to the low concordance between different datasets, the analyses here showed a potentially higher concordance by analyzing the interactions in terms of DDIs instead of PPIs. Irrespective of the experimental biases, the DDI analysis was able to cut through some of the underlying inconsistencies and expose an inherent propensity for certain types of domains to associate with each other. The DDI profiles can be downloaded from http://bhsai.org /downloads/ddi_profiles.

TABLE 2. PEARSON CORRELATION COEFFICIENTS, CORRESPONDING P-VALUES, AND NUMBERS OF CONSIDERED SUPERFAMILY DDIS FOR EACH OF THE 15 POSSIBLE PAIR-WISE COMPARISONS BETWEEN THE 6 DATASETS

Organism/Platform/Dataset	*S. cerevisiae* Y2H/Ito			*S. cerevisiae* TAP-MS/Gavin			*S. cerevisiae* TAP-MS/Krogan			*E. coli* TAP-MS/Arifuzzaman			*E. coli* TAP-MS/Butland		
	r	*P*	*N*	*r*	*P*	*N*	*r*	*P*	*N*	*r*	*P*	*N*	*r*	*P*	*N*
							Criterion I								
S. cerevisiae/Y2H/Uetz	**0.84**	**P***	**144**	0.89	P*	143	0.52	P*	149	0.32	$1.3e^{-2}$	61	0.06	$6.6e^{-1}$	49
S. cerevisiae/Y2H/Ito				0.87	P*	229	0.86	P*	247	0.66	P*	194	0.35	$8.6e^{-5}$	121
S. cerevisiae/TAP-MS/Gavin							**0.80**	**P***	**795**	0.42	P*	535	0.52	P*	464
S. cerevisiae/TAP-MS/Krogan										0.70	P*	430	0.73	P*	329
E. coli/TAP-MS/Arifuzzaman													**0.80**	**P***	**1,396**
							Criterion II								
S. cerevisiae/Y2H/Uetz	**0.13**	**P***	**1,252**	-0.09	$3.8e^{-4}$	1,556	-0.10	$8.8e^{-5}$	1,583	-0.09	$2.7e^{-4}$	1,685	-0.14	P*	1,111
S. cerevisiae/Y2H/Ito				-0.01	$7.3e^{-1}$	3,255	-0.07	$1.1e^{-4}$	2,954	-0.08	P*	4,369	-0.14	P*	2,948
S. cerevisiae/TAP-MS/Gavin							**0.18**	**P***	**4,217**	-0.11	P*	5,468	-0.17	P*	3,795
S. cerevisiae/TAP-MS/Krogan										-0.05	$6.3e^{-5}$	5,383	-0.10	P*	3,692
E. coli/TAP-MS/Arifuzzaman													**-0.12**	**P***	**1,400**

r, Pearson correlation coefficient; *N*, number of considered Superfamily domain-domain interactions (DDIs). *Criterion I* only considered DDIs that were observed in both datasets. *Criterion II* considered DDIs that were observed in only one dataset as long as the involved domains were observed in the other dataset. Intra-species and intra-platform correlations are shown in bold, whereas inter-species comparisons are shown by shaded cells. P* denotes $P \leq 1.0e^{-5}$

IV. DISCUSSION AND CONCLUSIONS

High levels of incoherence among PPIs of the same organisms have been documented in multiple studies involving high-throughput experimental techniques. Technological advancements have not resolved such inconsistencies. In this article, we demonstrated that, even though different high-throughput experimental determinations of PPIs were seemingly inconsistent with each other in terms of common interactions, there was a significant similarity at the level of DDIs. When considering only those DDIs that were observed in both datasets being compared, we observed a remarkably high pair-wise correlation ($0.80 \leq r \leq 0.84$) between datasets derived from the same experimental platform (TAP-MS or Y2H) on the same organism (*E. coli* or *S. cerevisiae*). To a lesser degree, this concordance held true even for inter-species comparisons between *E. coli* and *S. cerevisiae* PPIs based on the same TAP-MS experimental platform. However, different studies identified largely different sets of DDIs. Therefore, the correlation between DDI profiles was lost upon the inclusion of non-shared (as in *criterion II*) DDIs for all such comparisons except for those involving *S. cerevisiae* DDIs derived from the same experimental platform, which showed a weak but statistically significant correlation.

We further noted that a large fraction of PPIs evaded DDI-based modeling due to a lack of domain annotation in the involved proteins. To address this limitation, a more comprehensive and consistent annotation of domains is needed.

ACKNOWLEDGMENTS

We thank Dr. Chenggang Yu for assistance in identifying Superfamily domains via PIPA (www.bhsai.org). We also acknowledge fruitful discussions with Dr. Nela Zavaljevski.

REFERENCES

[1] S. Fields and O. Song, "A novel genetic system to detect protein-protein interactions," Nature, vol. 340, pp. 245-6, Jul 1989.

[2] A. Tsai and R. P. Carstens, "An optimized protocol for protein purification in cultured mammalian cells using a tandem affinity purification approach," Nat Protoc, vol. 1, pp. 2820-7, 2006.

[3] M. Arifuzzaman, M. Maeda, A. Itoh, et al., "Large-scale identification of protein-protein interaction of Escherichia coli K-12," Genome Res, vol. 16, pp. 686-91, May 2006.

[4] G. Butland, J. M. Peregrin-Alvarez, J. Li, et al., "Interaction network containing conserved and essential protein complexes in Escherichia coli," Nature, vol. 433, pp. 531-7, Feb 2005.

[5] N. J. Krogan, G. Cagney, H. Yu, et al., "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae," Nature, vol. 440, pp. 637-43, Mar 2006.

[6] A. C. Gavin, P. Aloy, P. Grandi, et al., "Proteome survey reveals modularity of the yeast cell machinery," Nature, vol. 440, pp. 631-6, Mar 2006.

[7] P. Uetz, L. Giot, G. Cagney, et al., "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae," Nature, vol. 403, pp. 623-7, Feb 2000.

[8] T. Ito, T. Chiba, R. Ozawa, et al., "A comprehensive two-hybrid analysis to explore the yeast protein interactome," Proc Natl Acad Sci USA, vol. 98, pp. 4569-74, Apr 2001.

[9] S. Kerrien, Y. Alam-Faruque, B. Aranda, et al., "IntAct-open source resource for molecular interaction data," Nucleic Acids Res, vol. 35, pp. D561-5, Jan 2007.

[10] A. Bairoch, B. Boeckmann, S. Ferro, et al., "Swiss-Prot: juggling between evolution and stability," Brief Bioinform, vol. 5, pp. 39-55, Mar 2004.

[11] J. M. Cherry, C. Ball, S. Weng, et al., "Genetic and physical maps of Saccharomyces cerevisiae," Nature, vol. 387, pp. 67-73, May 1997.

[12] M. Madera, C. Vogel, S. K. Kummerfeld, et al., "The SUPERFAMILY database in 2004: additions and improvements," Nucleic Acids Res, vol. 32, pp. D235-9, Jan 2004.

[13] N. J. Mulder, R. Apweiler, T. K. Attwood, et al., "New developments in the InterPro database," Nucleic Acids Res, vol. 35, pp. D224-8, Jan 2007.

[14] C. Yu, N. Zavaljevski, V. Desai, et al., "The development of PIPA: an integrated and automated pipeline for genome-wide protein function annotation," BMC Bioinformatics, vol. 9, pp. 52, Jan 2008.

[15] E. Sprinzak and H. Margalit, "Correlated sequence-signatures as markers of protein-protein interaction," J Mol Biol, vol. 311, pp. 681-92, Aug 2001.

[16] C. Su, J. M. Peregrin-Alvarez, G. Butland, et al., "Bacteriome.org--an integrated protein interaction database for E. coli," Nucleic Acids Res, vol. 36, pp. D632-6, Jan 2008.

[17] I. Xenarios, L. Salwinski, X. J. Duan, et al., "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions," Nucleic Acids Res, vol. 30, pp. 303-5, Jan 2002.

[18] R. D. Finn, J. Mistry, B. Schuster-Bockler, et al., "Pfam: clans, web tools and services," Nucleic Acids Res, vol. 34, pp. D247-51, Jan 2006.