research articles Journal of proteome research

An Optimized Predictive Strategy for Interactome Mapping†

Martin J. A. Aryee‡ and John Quackenbush*,‡,§

Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, Massachusetts 02115, and Department of Biostatistics and Computational Biology and Department of Cancer Biology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115

Received August 1, 2007

We present an optimized experimental strategy that can accelerate progress toward identifying the majority of pairwise protein interactions. Our method involves applying a predictive algorithm, based on the existing data, to identify protein pairs likely to interact and prioritizing these for screening. The approach is iterative as additional data allows one to refine predictions directing the next stage of experimentation.

Keywords: protein−protein interaction • interactome • yeast two-hybrid

Introduction

Although genome sequencing projects have provided preliminary catalogs of the protein coding genes within each organism, these "parts lists" are far from what is needed in order to understand how these proteins interact to mediate the wide range of cellular functions that exist. Interactome mapping seeks to identify the complete set of protein−protein interactions (PPIs) in an organism and as such represents a crucial element in moving beyond genomics to a systems-level understanding of biological systems.

There are two complementary approaches taken to high-throughput interactome mapping: the identification of protein complexes by pull-down experiments followed by mass spectrometry and the detection of pairwise interactions by the yeast two-hybrid (Y2H) system.[1] Although several projects are underway to produce interactome maps for a variety of organisms, these are far from complete. The very scale of interactome mapping presents logistical challenges that delay the availability of the most useful information. Assuming, for example, that the human genome contains 22 000 protein-coding genes, there are almost 250 million potential pairwise interactions that must be tested if one is to complete a comprehensive interactome map. When alternative splice forms and post-translational modifications are considered, the number of potential interactions that must be investigated increases dramatically. A further complication is that existing screening techniques require that each interaction be screened multiple times before high-confidence interactome maps can be produced.[2,3] Given the size of the search space, it is clear that interactome mapping is in its very early stages.

Because of the challenges in completing interactome maps, several approaches have been proposed to accelerate the process. A number of purely computational predictive methods that use existing data about proteins and their interactions to infer novel interactions have been proposed.[4−11] Unfortunately, these methods suffer from high false-positive or false-negative rates, limiting their usefulness in large-scale discovery of new interactions. An alternative approach would be to combine predictive and experimental methods to accelerate discovery through optimization of the strategy used to explore the search space. Here, we propose an iterative computational and experimental method that can accelerate mapping of pairwise interactions.

For the purposes of our discussion, we focus on high-throughput yeast two-hybrid (HT-Y2H) screens, although the method could equally well be adapted to mass-spectrometry based interactome techniques. Y2H analysis is based on the fact that, in most eukaryotic transcription factors, the activating and binding domains are modular and can function provided they are in close proximity. Fusion proteins are created in which two target proteins of interest are linked, respectively, to a DNA yeast transcription factor binding domain and a corresponding DNA transcription activation domain; these are typically referred to as the "bait" and "prey." If the bait and prey are cloned into a single yeast, an interaction between the two can initiate transcription of a selectable marker or "reporter gene"; if the two do not interact, transcription of the reporter does not occur. In this way, an interaction between proteins produces a detectable phenotype.

Since it would be nearly impossible to individually assay each of the possible protein pairs for potential interactions, high-throughput screens instead pool proteins and test them in a batch-wise fashion. Although early high-throughput Y2H analyses used large pools containing all available proteins, high false-positive and false-negative rates led to a change in strategy. Most studies now use a matrix approach that relies on relatively small pools of proteins (8−200).[12,13] The use of smaller pools improves sensitivity since rare or transient interactions are less likely to be masked by more prevalent ones. This matrix approach allows a natural partitioning of the full search space

‡ Harvard School of Public Health.
§ Dana-Farber Cancer Institute.
† Originally submitted and accepted as part of the "Statistical and Computational Proteomics" special section, published in the January 2008 issue of *J. Proteome Res.* (Vol. 7, No. 1).
* To whom correspondence should be addressed. E-mail: johnq@jimmy.harvard.edu. Tel: +1 617 582 8094.

10.1021/pr700858e CCC: $40.75

Published on Web 07/22/2008

into smaller sub-blocks that are selected in such a way that one works through all possible interactions in a systematic fashion. We aim to divide the search space into sub-blocks in a way that increases the rate of interaction discovery.

Here, we propose a directed strategy that involves iteratively screening groups of proteins and using the resulting interaction data to develop predictions of which proteins not yet screened are likely to interact. These predictions allow selections of sub-blocks for the next round of screening that have the greatest predicted return. This strategy is then iterated, with the data from each round being used to improve and refine predictions for the next round of screening. In this way, we hope to more rapidly approach the asymptotic limit of detectable interactions, delivering the majority of the interactome in the shortest possible time. To illustrate the value of such an approach, we use data from several large high-throughput interactome data sets and estimate the increased rate at which protein−protein interactions are discovered by an optimized search strategy. While useful in any interactome screen, such an approach will be particularly useful when the search space of interest is too large to permit an exhaustive search since such a directed approach will provide the greatest number of protein−protein interactions early on in the screen.

To the best of our knowledge, this is the first application of computational PPI predictions to optimize a high-throughput pairwise screen. It should be noted that our proposal is complementary to and distinct from 'smart pooling' techniques such as PI-deconvolution proposed by Jin et al.[14] or the shifted transversal design proposed by Thierry-Mieg.[15] Smart pooling combines baits and/or preys into efficiently designed pools to improve coverage while simultaneously reducing the number of screens required to screen a given search space. We propose an approach that combines our proposed method with smart pooling by first dividing the search space into a sequence of large (say 1000 × 1000) sub-blocks using the algorithm we present here, and then screening each sub-block in an optimal manner using a smart pooling technique. Finally, while, Lappe et al.[16] have proposed a method to accelerate discovery of protein complexes in high-throughput pull-down/mass spectrometry experiments, the method we propose here represents a unique application to high-throughput analysis of pairwise interactions.

Methods

PPI Data Sources. We define the search space of an organism as all potential pairwise interactions between its proteins. This is conveniently visualized as a two-dimensional matrix in which proteins are represented as individual rows and columns; the matrix is largely populated by zeros as most proteins do not interact with each other and interactions between two proteins are represented as ones in the corresponding locations in the matrix.

To illustrate the benefit of integrating prediction with systematic mapping, we used a data set from a large human interactome screen to assess the impact on the rate of interactome mapping. The data set generated by the Dana-Farber Cancer Institute Center for Cancer Systems Biology is described in detail in Rual et al.[12] It contains 2754 protein−protein interactions between 1549 proteins. The high-throughput yeast two-hybrid screen tested all two-way pairings between Gateway-cloned open reading frames representing ~7000 unique genes. Assuming ~22 000 genes in the human genome and ignoring splice forms and post-translational modifications the ~7000 × 7000 gene pairs screened represent about 10% of the entire human interactome space. We created simulated search spaces to test our proposed algorithm by starting with the 1549 proteins in the data set and adding randomly chosen proteins to create a 7000 × 7000 space. All simulations were repeated 10 times.

We used other PPI data sets to provide initial training data, and as independent test sets. The data set from a second large human Y2H screen[13] and other Y2H data sets from the yeast,[17,18] worm[19] and fly[20] interactomes were obtained from the EBI IntAct repository.[21] A large compendium of non-Y2H human PPI data was obtained from UniHI,[22] a resource that combines data from a variety of other PPI data sources.

Identifying Likely Protein−Protein Interactions. The predictive method used to infer likely protein−protein interactions is a central component of the optimized search strategy. Several computational methods have been developed including predictions based on gene fusion events,[6,7] orthology across species,[23–25] and domain−domain interactions.[5,8,9,26] For the purpose of directing a high-throughput screen of millions of potential interactions, it is ideal to rank order as much of the search space as possible. While methods such as cross-species orthology can predict a small number of likely interactions, they provide no information about the likelihood of the vast majority of potential interactions. The strength of domain based methods in this setting is that they can assign a rough interaction probability to the majority of potential interactions making them ideal for screening the search space in a directed manner. These methods rely on the assumptions that (1) protein interactions are mediated by interactions between their constituent domains, and (2) disparate data sets and organisms share common domain−domain interactions. Figure 1a visualizes the domain−domain interaction space of two PPI search spaces. The shaded overlapping region contains domain−domain interactions (DDIs) that can be inferred from one data set and that are useful for making predictions about PPIs in the second.

Our proposed algorithm is not specific to any particular PPI prediction algorithm. PPI prediction is an active area of research and newer algorithms can be substituted as they become available. For this work, we chose the association method[8] and obtained domain composition for proteins from the InterPro database.[27]

As an example, consider two proteins where protein i contains domains m and n, and protein j contains domain o, as shown in Figure 1b. Our goal is to estimate the likelihood, P_{ij}, that these two proteins interact based on the observed likelihood, λ, that the domains that comprise those proteins, interact in a pairwise fashion:

$$\begin{aligned} P_{ij} &= P(\text{protein i and protein j interact}) \\ &= 1 - P(\text{protein i and protein j do not interact}) \\ &= 1 - P(\text{no domain pairs interact}) \\ &= 1 - (1 - \lambda_{mo})(1 - \lambda_{no}) \end{aligned} \quad (1)$$

We estimate the domain interaction probabilities, λ, from the available protein interaction data. For example, the likelihood that domains m and o interact, λ_{mo} is given by equation 2.

$$\lambda_{mo} = \frac{I_{mo}}{N_{mo}} \quad (2)$$

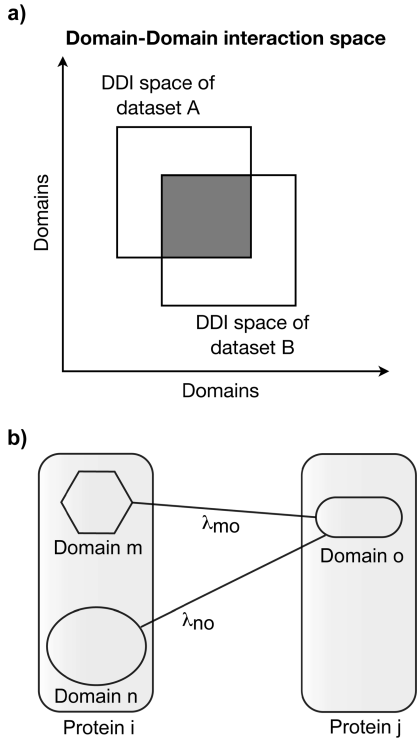

Figure 1. (a) The domain−domain interaction spaces of two PPI data sets. The overlapping region contains interactions present in both data sets. (b) Two proteins, i and j, with two and one domain, respectively. The λ's represent the observed probability that any two domains bind to each other; we assume that the proteins will interact if at least one of the two domain pairs interact. Given the domain interaction probabilities, we estimate P_{ij}, the probability of protein i interacting with protein j.

where N_{mo} is the number of training data set experiments where domains m and o were present. I_{mo} is the number of these experiments that resulted in binding. λ_{mo} is then an estimate of the interaction probability between domains m and o. Essentially, this requires that we decompose the matrix of protein−protein interactions into a matrix of domain−domain interactions to estimate the probabilities. This model assumes that interactions are additive, that domain order does not contribute to the interaction probability, and that there are no inhibitory domain interactions.

To improve predictive power, we can also take advantage of the evolutionary conservation of protein structure and function and use existing interaction data sets from other species to expand the training set. This has been shown to improve performance in a study where success in predicting yeast protein interactions was considerably enhanced by adding worm and fly data to the training set.[26] For the analysis we present here, we include training data from yeast,[17,18] worm,[19] and fly.[20]

Iterative Interactome Screening Algorithm. In demonstrating the utility of our optimization procedure, we use a large human Y2H interactome data set[12] in simulated experiments in which subsets of proteins are sequentially screened. We can envision the search space as a ~7000 × 7000 grid where each cell represents a protein pair. To partition this grid into sub-blocks, we divide the proteins into smaller sets by reorganizing the rows and columns and selecting intersection regions within the grid. This allows us to determine an optimal group of proteins for screening in each iteration, maximizing the potential to discover new interactions.

Initially, we train the model using previously available protein−protein interaction data from other data sets to estimate domain−domain interaction probabilities. We then calculate an interaction probability for each protein pair in the search space and create the first sub-block to be screened by selecting the section of the search space grid with the largest expected number of interactions. After this and each subsequent screening step, we add the newly obtained interaction data to the training set and update our predictive model by recalculating the domain interaction probabilities in equation 2. The refined models are then used to select the next set of proteins and to choose the next sub-block for screening. This process is repeated iteratively until the entire search space has been screened or the rate of discovery of new interactions falls below some threshold. The size of the sub-blocks can be chosen to suit the workflow of the laboratory, depending on how often it is convenient to interrupt screening in order to update the predictive model. For our main simulation, we divided the search space into 49 sub-blocks of 1000 × 1000 proteins. The approach is illustrated in Figure 2. The simulation software implementing the optimized screening algorithm is available at: http://compbio.dfci.harvard.edu/publications.html.

It is well-known that interactions between yeast two-hybrid baits and preys are not reciprocal.[28] This is not surprising since the system involves fusing the proteins of interest to an additional domain which may alter structure, function and/or binding properties. This introduces a difficulty for directed search algorithms since we are unable to predict in which direction a particular interaction will be detected. Our solution is to screen the search space in a symmetrical fashion whenever possible. When we identify a sub-block of preys and baits to screen, we also screen the reciprocal block, using the baits as preys and vice versa.

Results

We tested the directed search strategy using several independent data sets covering parts of the human,[12] worm,[19] fly,[20] and yeast[17,18] interactomes. Here, we present the results using the human screen of ~7000 proteins described in Rual et al.[12] as test data with the other data sets used as initial training data.

Figure 3 compares the progress toward the human Y2H interactome map of Rual et al. using an undirected screen (dotted line) and the expected progress had an iterative prediction method been applied (solid line). In the case of an undirected screen, the fraction of interactions found at any point will be roughly equal to the fraction of the search space that has been screened. For example, after 10% of the search space has been systematically screened in the standard fashion, we would have found about 10% of the interactions that would eventually be revealed by the completed screen. By comparison, had we directed the search with the suggested prediction strategy, we would expect to have found over 20% of the interactions, more than doubling the size of the known set at this time. On the basis of this simulated experiment, we can predict how the increased rate of discovery will translate into gains in future large-scale screens. As our current knowledge has been estimated to represent only a small fraction[2] of true interactions, we find ourselves on the early part of the curve shown in Figure 3 where the return from experimental optimization is highest.

By initially directing the screen toward interactions that are predicted to be more likely, we obtain an increase in the rate of interaction discovery. Figure 4 shows the difference in the

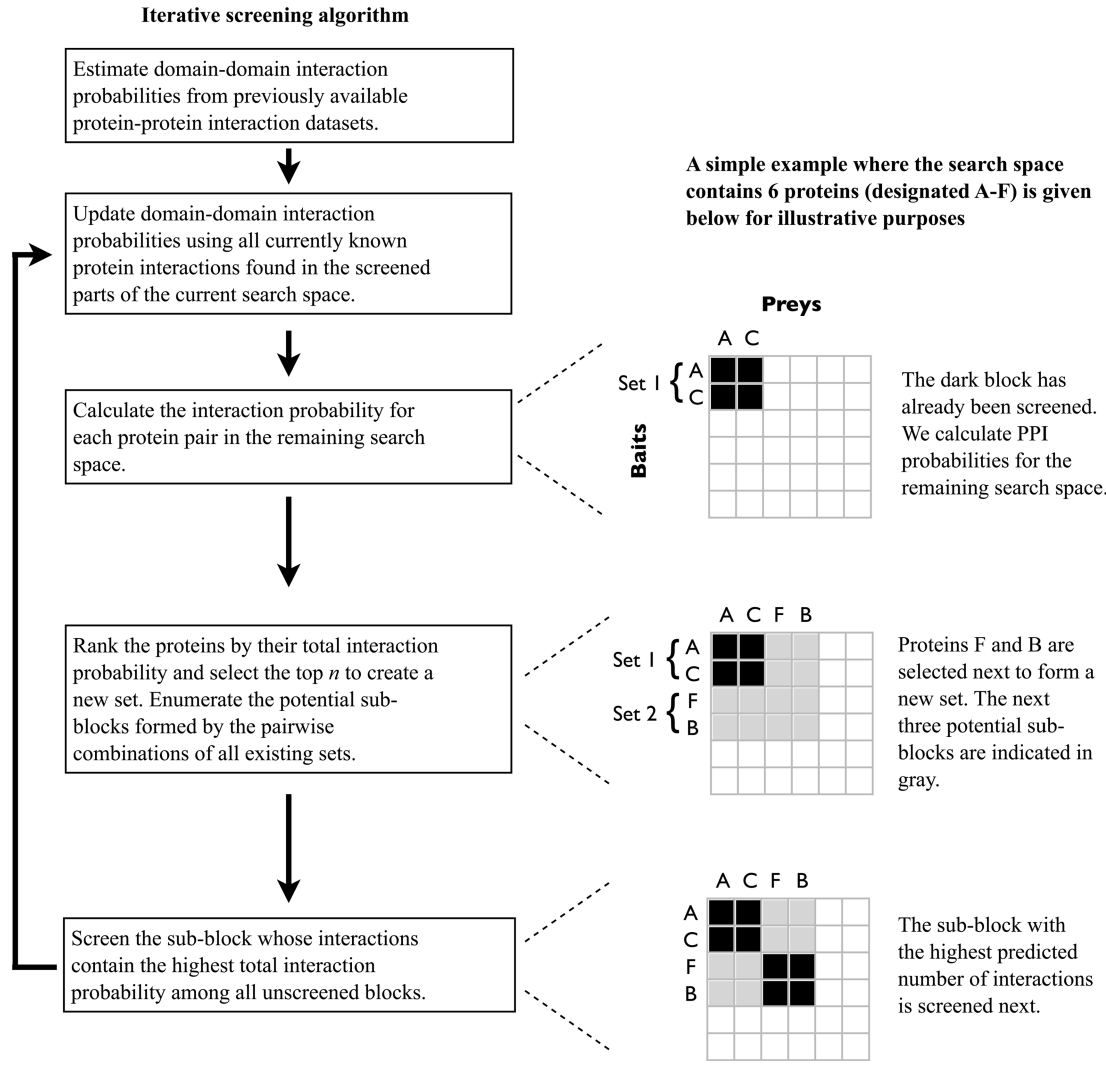

Figure 2. An overview of the iterative screening process that represents the core of our proposed method. A model is trained on previous data sets and available data from the ongoing screen to identify those sub-blocks of the remaining search space that are most likely to contain interactions. As each sub-block is screened, the newly found interactions are used to update the predictive model and the process is iterated.

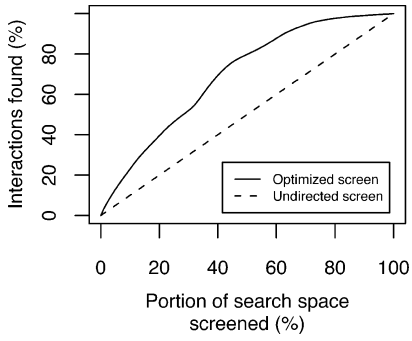

Figure 3. A plot showing the fraction of interactions found as a function of the amount of search space that has been screened. The dashed line represents progress with a standard systematic screen where the number of interactions found is roughly proportional to the fraction of the search space that has been completed. The solid line shows the increased rate of progress when the screen is supplemented by a prediction strategy.

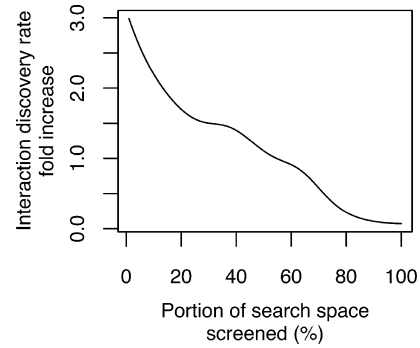

Figure 4. The interaction discovery rate fold increase provided by our proposed strategy as compared to an undirected screen. The greatest benefit from our proposed optimization method is obtained during the early stages of the screen.

rates of protein interaction discovery between an undirected screen and an optimized screen. Using the predictive strategy gives an increased rate of new interaction discovery until about 50% of the search space has been screened.

An interesting finding is that the best predictive power and hence greatest gains in screening efficiency are obtained when the training data consists exclusively of other Y2H data sets.

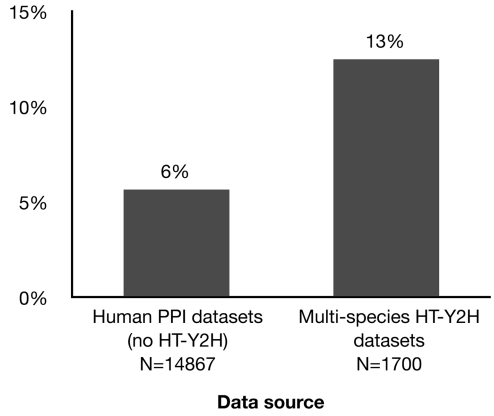

Figure 5. Percentage of domain−domain interactions in different training data sets that are useful for predicting the human PPIs detected by HT-Y2H in the Rual et al. screen. There is considerably greater concordance of domain−domain interactions between two HT-Y2H data sets than between a HT-Y2H data set and general PPI data.

Expanding the training set to include other (presumably higher-quality) data, including literature-curated sources, results in decreased performance. In this case, the area under the curve (AUC) shown in Figure 3 drops from 70% to 61% (Supplemental Figure S1). When using an existing PPI data set to make predictions about a future screen, we rely on their shared domain−domain interaction rules. Figure 1a visualizes the domain−domain interaction space of two PPI search spaces. The shaded overlapping region contains the DDIs useful for making predictions across the PPI search spaces. Prediction based on DDIs works best when this shaded region has a large overlap of domain interactions between the training and test PPI data sets. As shown in Figure 5, HT-Y2H PPI data sets contain a distinct subset of the domain interaction observed in general PPI data. If our goal is to predict Y2H interactions, we obtain better predictive performance by basing our model only on DDIs observed in Y2H data.

To verify that the results are not specific to the data set used above, we repeated the simulation using an independent HT-Y2H human data set from a different laboratory.[13] The performance gains were similar and are presented in Supporting Material Figure S2. To demonstrate that the procedure is robust to the choice of different block sizes, we repeated the simulation with a block size of 500 × 500 proteins. There was little impact on performance as can be seen in Supporting Material Figure S3.

Discussion and Conclusion

It is likely that any prediction method that uses high-throughput training data will be biased toward predicting only those interactions detectable by the method used to produce the training data set(s). This bias is actually a benefit in the setting of experimental optimization as we are trying to direct the search only toward those interactions that can be detected by the experimental system of interest. As can be seen in Figure 5, the HT-Y2H assay preferentially detects PPIs mediated by certain domain interactions. As a result, general PPI data sets provide poor training material for directing a HT-Y2H screen. Therefore, it is advantageous to restrict the training data to only HT-Y2H data sets.

Using prediction to guide the search order in a systematic screen will accelerate progress, particularly in the early stages of a screening project. While this strategy will not provide us with a complete interactome map any sooner than an undirected approach (since the entire space must, by definition, still be searched), it can provide a more comprehensive catalogue of pairwise interactions at an earlier stage in the project. Given that we find ourselves at the very early stages of interactome mapping and that a vast amount of experimental work lies ahead, there are considerable gains to be had by using a prioritization approach such as that described here. Other studies focused at deducing cellular networks would benefit from a larger experimentally supported catalogue of protein−protein interactions.

Given the current rate of screening, completing the interaction maps of human and other organisms will take many years to complete. If the search space were expanded to include various splice-forms and protein variants representing genetic polymorphisms, the task becomes practically insurmountable with current technology. Using a directed approach would require more effort to prioritize clones prior to pooling and screening and involve some investment in additional laboratory information management systems development to track the clones that have been screened. However, the approaches used to date have largely been semidirected already, in the sense that subsets of genes have been prioritized for screening. Thus, using this predictive strategy would add relatively little additional time and effort to the process.

The demonstrated value of interactome data for understanding the cellular networks that ultimately dictate cellular phenotypes suggests that large-scale, long-term projects such as the mapping of the complete human interactome should be optimized to provide the greatest amount of useful data as soon as possible. The method presented here represents an efficient and effective means of achieving this goal.

Supporting Information Available: Figures of simulated screens where the training data included both HT-Y2H and non-HT-Y2H data, using a different human test data set described in Stelzl et al., and where the search space is divided into ~200 500 × 500 blocks. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Fields, S.; Song, O. A novel genetic system to detect protein-protein interactions. *Nature* **1989**, *340* (6230), 245–6.
(2) Hart, G. T.; Ramani, A. K.; Marcotte, E. M. How complete are current yeast and human protein-interaction networks. *GenomeBiology* **2006**, *7* (11), 120.
(3) Goll, J.; Uetz, P. The elusive yeast interactome. *GenomeBiology* **2006**, *7* (6), 223.
(4) Bock, J. R.; Gough, D. A. Predicting protein-protein interactions from primary structure. *Bioinformatics* **2001**, *17* (5), 455–60.
(5) Deng, M.; Mehta, S.; Sun, F.; Chen, T. Inferring domain-domain interactions from protein-protein interactions. *Genome Res.* **2002**, *12* (10), 1540–8.
(6) Enright, A. J.; Iliopoulos, I.; Kyrpides, N. C.; Ouzounis, C. A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **1999**, *402* (6757), 86–90.
(7) Marcotte, E. M.; Pellegrini, M.; Ng, H. L.; Rice, D. W.; Yeates, T. O.; Eisenberg, D. Detecting protein function and protein-protein interactions from genome sequences. *Science* **1999**, *285* (5428), 751–3.
(8) Sprinzak, E.; Margalit, H. Correlated sequence-signatures as markers of protein-protein interaction. *J. Mol. Biol.* **2001**, *311* (4), 681–92.
(9) Wojcik, J.; Schachter, V. Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics* **2001**, *17* (Suppl. 1), S296–305.
(10) Jansen, R.; Yu, H.; Greenbaum, D.; Kluger, Y.; Krogan, N. J.; Chung, S.; Emili, A.; Snyder, M.; Greenblatt, J. F.; Gerstein, M. A Bayesian

networks approach for predicting protein-protein interactions from genomic data. *Science* **2003**, *302* (5644), 449–53.
(11) Riley, R.; Lee, C.; Sabatti, C.; Eisenberg, D. Inferring protein domain interactions from databases of interacting proteins. *GenomeBiology* **2005**, *6* (10), R89.
(12) Rual, J. F.; Venkatesan, K.; Hao, T.; Hirozane-Kishikawa, T.; Dricot, A.; Li, N.; Berriz, G. F.; Gibbons, F. D.; Dreze, M.; Ayivi-Guedehoussou, N.; Klitgord, N.; Simon, C.; Boxem, M.; Milstein, S.; Rosenberg, J.; Goldberg, D. S.; Zhang, L. V.; Wong, S. L.; Franklin, G.; Li, S.; Albala, J. S.; Lim, J.; Fraughton, C.; Llamosas, E.; Cevik, S.; Bex, C.; Lamesch, P.; Sikorski, R. S.; Vandenhaute, J.; Zoghbi, H. Y.; Smolyar, A.; Bosak, S.; Sequerra, R.; Doucette-Stamm, L.; Cusick, M. E.; Hill, D. E.; Roth, F. P.; Vidal, M. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **2005**, *437* (7062), 1173–8.
(13) Stelzl, U.; Worm, U.; Lalowski, M.; Haenig, C.; Brembeck, F. H.; Goehler, H.; Stroedicke, M.; Zenkner, M.; Schoenherr, A.; Koeppen, S.; Timm, J.; Mintzlaff, S.; Abraham, C.; Bock, N.; Kietzmann, S.; Goedde, A.; Toksoz, E.; Droege, A.; Krobitsch, S.; Korn, B.; Birchmeier, W.; Lehrach, H.; Wanker, E. E. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **2005**, *122* (6), 957–68.
(14) Jin, F.; Hazbun, T.; Michaud, G.; Salcius, M.; Predki, P.; Fields, S.; Huang, J. A pooling-deconvolution strategy for biological network elucidation. *Nat. Methods* **2006**, *3* (3), 183–9.
(15) Thierry-Mieg, N. A new pooling strategy for high-throughput screening: the Shifted Transversal Design. *BMC Bioinf.* **2006**, *7*, 28.
(16) Lappe, M.; Holm, L. Unraveling protein interaction networks with near-optimal efficiency. *Nat. Biotechnol.* **2003**, *22* (1), 98–103.
(17) Ito, T.; Chiba, T.; Ozawa, R.; Yoshida, M.; Hattori, M.; Sakaki, Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98* (8), 4569–74.
(18) Uetz, P.; Giot, L.; Cagney, G.; Mansfield, T. A.; Judson, R. S.; Knight, J. R.; Lockshon, D.; Narayan, V.; Srinivasan, M.; Pochart, P.; Qureshi-Emili, A.; Li, Y.; Godwin, B.; Conover, D.; Kalbfleisch, T.; Vijayadamodar, G.; Yang, M.; Johnston, M.; Fields, S.; Rothberg, J. M. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **2000**, *403* (6770), 623–7.
(19) Li, S.; Armstrong, C. M.; Bertin, N.; Ge, H.; Milstein, S.; Boxem, M.; Vidalain, P. O.; Han, J. D.; Chesneau, A.; Hao, T.; Goldberg, D. S.; Li, N.; Martinez, M.; Rual, J. F.; Lamesch, P.; Xu, L.; Tewari, M.; Wong, S. L.; Zhang, L. V.; Berriz, G. F.; Jacotot, L.; Vaglio, P.; Reboul, J.; Hirozane-Kishikawa, T.; Li, Q.; Gabel, H. W.; Elewa, A.; Baumgartner, B.; Rose, D. J.; Yu, H.; Bosak, S.; Sequerra, R.; Fraser, A.; Mango, S. E.; Saxton, W. M.; Strome, S.; Van Den Heuvel, S.; Piano, F.; Vandenhaute, J.; Sardet, C.; Gerstein, M.; Doucette-Stamm, L.; Gunsalus, K. C.; Harper, J. W.; Cusick, M. E.; Roth, F. P.; Hill, D. E.; Vidal, M. A map of the interactome network of the metazoan C. elegans. *Science* **2004**, *303* (5657), 540–3.
(20) Giot, L.; Bader, J. S.; Brouwer, C.; Chaudhuri, A.; Kuang, B.; Li, Y.; Hao, Y. L.; Ooi, C. E.; Godwin, B.; Vitols, E.; Vijayadamodar, G.; Pochart, P.; Machineni, H.; Welsh, M.; Kong, Y.; Zerhusen, B.; Malcolm, R.; Varrone, Z.; Collis, A.; Minto, M.; Burgess, S.; McDaniel, L.; Stimpson, E.; Spriggs, F.; Williams, J.; Neurath, K.; Ioime, N.; Agee, M.; Voss, E.; Furtak, K.; Renzulli, R.; Aanensen, N.; Carrolla, S.; Bickelhaupt, E.; Lazovatsky, Y.; DaSilva, A.; Zhong, J.; Stanyon, C. A.; Finley, R. L., Jr.; White, K. P.; Braverman, M.; Jarvie, T.; Gold, S.; Leach, M.; Knight, J.; Shimkets, R. A.; McKenna, M. P.; Chant, J.; Rothberg, J. M. A protein interaction map of Drosophila melanogaster. *Science* **2003**, *302* (5651), 1727–36.
(21) Kerrien, S.; Alam-Faruque, Y.; Aranda, B.; Bancarz, I.; Bridge, A.; Derow, C.; Dimmer, E.; Feuermann, M.; Friedrichsen, A.; Huntley, R.; Kohler, C.; Khadake, J.; Leroy, C.; Liban, A.; Lieftink, C.; Montecchi-Palazzi, L.; Orchard, S.; Risse, J.; Robbe, K.; Roechert, B.; Thorneycroft, D.; Zhang, Y.; Apweiler, R.; Hermjakob, H. IntAct-open source resource for molecular interaction data. *Nucleic Acids Res.* **2007**, *35* (Database issue), D561–5.
(22) Chaurasia, G.; Iqbal, Y.; Hanig, C.; Herzel, H.; et al. UniHI: an entry gate to the human protein interactome. *Nucleic Acids Res.* **2007**, *35* (Database issue), D590–4.
(23) Brown, K. R.; Jurisica, I. Online Predicted Human Interaction Database. *Bioinformatics* **2005**, *21* (9), 2076–82.
(24) Lehner, B.; Fraser, A. G. A first-draft human protein-interaction map. *GenomeBiology* **2004**, *5* (9), R63.
(25) Persico, M.; Ceol, A.; Gavrila, C.; Hoffmann, R.; et al. HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinf.* **2005**, *6* (Suppl. 4), S21.
(26) Liu, Y.; Liu, N.; Zhao, H. Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics* **2005**, *21* (15), 3279–85.
(27) Mulder, N. J.; Apweiler, R.; Attwood, T. K.; Bairoch, A.; et al. New developments in the InterPro database. *Nucleic Acids Res.* **2007**, *35* (Database issue), D224–8.
(28) Aloy, P.; Russell, R. Potential artefacts in protein-interaction networks. *FEBS Lett.* **2002**, *530* (1−3), 253–4.

PR700858E