Structure-Templated Predictions of Novel Protein Interactions from Sequence Information

Doron Betel[1,2¤*], Kevin E. Breitkreuz[1], Ruth Isserlin[1,2], Danielle Dewar-Darch[1], Mike Tyers[1,3], Christopher W. V. Hogue[2]

1 Samuel Lunenfeld Research Institute, Mt. Sinai Hospital, Toronto, Ontario, Canada, **2** Department of Biochemistry, University of Toronto, Toronto, Ontario, Canada, **3** Department of Medical Genetics and Microbiology, University of Toronto, Toronto, Ontario, Canada

The multitude of functions performed in the cell are largely controlled by a set of carefully orchestrated protein interactions often facilitated by specific binding of conserved domains in the interacting proteins. Interacting domains commonly exhibit distinct binding specificity to short and conserved recognition peptides called binding profiles. Although many conserved domains are known in nature, only a few have well-characterized binding profiles. Here, we describe a novel predictive method known as domain–motif interactions from structural topology (D-MIST) for elucidating the binding profiles of interacting domains. A set of domains and their corresponding binding profiles were derived from extant protein structures and protein interaction data and then used to predict novel protein interactions in yeast. A number of the predicted interactions were verified experimentally, including new interactions of the mitotic exit network, RNA polymerases, nucleotide metabolism enzymes, and the chaperone complex. These results demonstrate that new protein interactions can be predicted exclusively from sequence information.

Citation: Betel D, Breitkreuz KE, Isserlin R, Dewar-Darch D, Tyers M, et al. (2007) Structure-templated predictions of novel protein interactions from sequence information. PLoS Comput Biol 3(9): e182. doi:10.1371/journal.pcbi.0030182

Introduction

The interaction between two proteins is a geometric and electrostatic match between two polypeptide surfaces that results in a stable set of bonds between amino acid side chains or backbone atoms. The interacting amino acids are often part of conserved sequence features such as domains or short linear motifs that constitute the interaction site between the two proteins. Despite the increased coverage and sensitivity of experimental techniques for detecting protein interactions [1–6] (reviewed in [7]), elucidating the precise interacting residues remains experimentally difficult. In most cases, all that is known about an interaction is the identity of the two interacting proteins, with little information about the underlying binding site. However, detailed knowledge of interaction specificity is important for understanding reaction mechanism, interaction prediction, and drug development.

Interacting domains are autonomous structural elements that exhibit distinct binding specificity to a multitude of target polypeptides. Such domains act as independent elements that can be "plugged" into a new protein and thereby introduce new functionality to the emerging protein [8]. From an evolutionary perspective, such rearrangements and the multiplication of existing conserved domains is a likely mechanism by which organisms generate new proteins, pathways, and novel functionalities [9,10]. Several protein interaction prediction methods exploit the conservation of protein-binding interfaces by identifying domain pairs that consistently co-occur in interacting proteins or coevolve, which are then used to predict new interactions [11–16]. Structure-based prediction methods use known protein complexes to model interactions between proteins that are homologous to the complex components [17,18]. Other prediction methods use integrative approaches that incorporate interaction experiments with additional functional information such as correlated expression level, common functional annotation [19,20], and cross-species comparisons [21]. Alternative approaches attempt to identify correlated sequence motifs that represent generic interacting sequence elements that may or may not be components of conserved domains [22–25]. In a few limited cases, detailed experimental data are used to generate high-resolution definition of domain binding profiles; however, such information is available only for a small number of domains [26,27].

Our primary objective is to predict interaction between proteins strictly from sequence information. Our approach is based on identifying the binding specificity of interacting domains that can then be used to predict new interactions. Here, we use existing physical interaction data to derive sequence profiles of the binding sequences that are presumed to determine the binding specificity of interacting domains. Our method, called domain–motif interactions from structural topology (D-MIST), is based on a two-step approach. First, potential domain-binding motifs are extracted from structural data. Second, these motifs are converted to

Editor: Luhua Lai, Peking University, China

Received April 4, 2007; **Accepted** August 2, 2007; **Published** September 21, 2007

Abbreviations: D-MIST, domain–motif interactions from structural topology; IP-MS, immunoprecipitation followed by mass spectrometry; IP-western, immunoprecipitation followed by Western blotting; PSSM, position-specific scoring matrix

* To whom correspondence should be addressed. E-mail: betel@cbio.mskcc.org

¤ Current address: Computational and Systems Biology Center, Memorial Sloan-Kettering Cancer Center, New York, New York, United States of America

Author Summary

Many functions performed within a living cell are mediated by specific interactions between proteins. Precise geometric and chemical matches between segments of the protein structures facilitate those interactions. Such binding surfaces are often evolutionarily conserved elements of protein structures known as conserved domains that recognize specific binding elements on the interacting proteins. Binding domains and their corresponding interacting profiles constitute basic interacting modules that are replicated in multiple protein pairs, where they mediate similar interactions. Although many conserved domains are identified, only a handful have known, well-characterized binding elements. This paper describes a computational method that aims to elucidate the binding specificity of many domains. The utility of the derived binding specificity is demonstrated by predicting new interactions between yeast proteins. The predictions are based solely on sequence information by identifying the conserved domains and their corresponding binding sequences. A number of the predicted interactions were confirmed experimentally, demonstrating the feasibility of this approach.

sequence profiles in the form of position-specific scoring matrices (PSSMs). These PSSMs are derived using a subset of experimentally determined binary interactions that contain the domain of interest (Figure 1). Gibbs sampling, seeded with the motif extracted from structural data, is used to generate a PSSM from similar sequences that occur in a subset of established interacting proteins. We used the domain-binding profiles to predict protein interactions in yeast. The predictions were compared to a hidden set of known interactions reported in the literature, and several predicted interactions were confirmed directly by in vivo coprecipitation experiments.

Results

The library of 3-D structures of protein complexes contains a detailed description of the binding interfaces between interacting proteins that include atom contacts and residue side-chain interactions [28]. Using more than 10,000 structural complexes, we identified the domains in the binding sites and extracted their associated sequence motifs on the opposing chain. Interacting residues were defined as two residues on opposite polypeptide chains separated by a maximum of 5 Å (Figure 1A). On average, each domain had two spatially separated interacting sequence motifs per interaction. Most domains were present in multiple 3-D structures in a variety of conformations, resulting in varied interacting sequence motifs with different levels of similarities.

The binding specificity of a domain is determined by a combination of physiochemical properties and structural constraints at the binding site that can be satisfied by multiple variations of the consensus sequence motif [29]. The interacting sequence motifs extracted from the protein structures represent a first approximation of the binding specificity of the interacting domains, but do not represent the full evolutionary variations of the residue–residue interactions available in one binding topology. A more informative representation of the possible motif variations is a sequence profile in the form of a PSSM that captures the compositional variance by assigning probabilities to each amino acid at each position. These sequence variations of the binding profiles can be learned from proteins that are known to interact through the same domain.

We collected a set of 87,894 nonredundant protein interactions from four databases containing binary protein interactions from multiple species. Interactions derived from structural studies were excluded to preclude self-identification, as well as high-throughput protein complexes identification experiments [30,31] (see Methods). Gibbs sampling [32] was used to learn the PSSM binding profiles for a specific domain by sampling positions in the set of proteins that interact with proteins that contain the domain of interest. The majority of the proteins in the learning set are assumed to interact through the common domain, and the generated PSSM will represent its binding profile (Figure 1B). Gibbs sampling enables the incorporation of prior knowledge about the length and composition of the binding profiles. The motifs identified in the 3-D structural analysis were used as prior knowledge in seeding the profile detection step to bias the sampling towards similar sequence regions. The result is a set of sequence PSSMs that represent the binding profiles of the interacting domains (Text S1).

The learned PSSMs were used to predict interactions for 703 yeast proteins with domains for which we successfully derived binding profiles. A physical interaction was predicted between proteins containing interacting domains and proteins with one or more of the interacting profiles associated with those domains (Figure 1C). A total of 18,459 interactions were predicted between 2,313 proteins (Dataset S1). We compared the predicted interactions to a comprehensive list of physical and genetic yeast interactions extracted from the literature [33] and found that 609 predicted interactions have reported experimental evidence ($\sim$3%; $p = 1.0 \times 10^{-13}$; Figure S1). We note that 591 predicted interactions were found in both the 87,894 set of interactions used for the PSSM derivation and in the set of yeast literature curated interactions ($\sim$32,000). However, none of the 609 predicted interactions that have supporting evidence in the literature overlap with those common 591 interactions. We did not incorporate additional experimental information such as cellular localization, functional annotation, surface accessibility, or gene expression data that would likely improve our prediction accuracy given that our primary goal was to predict novel interactions exclusively from sequence information.

Experimental verification of a subset of the predicted interactions was performed by a one-step immunoaffinity purification of one of the two interaction partners, followed by mass spectrometric identification of associated proteins (IP-MS) as previously described [31]. The IP-MS method confirmed 37 predicted interactions, including 23 novel interactions (Figure 2). As a second means to experimentally verify our predictions, we immunoprecipitated one protein in the interacting pair, followed by antibody detection of the second protein (IP-western), also as described in [31]. The IP-western method reaffirmed five of the interactions confirmed by IP-MS (yellow edges; Figure 3) and identified an additional four novel interactions (green edges; Figure 3). We note that six interactions confirmed by the IP-MS approach were not detected by IP-western (red dashed edges; Figure 3); this

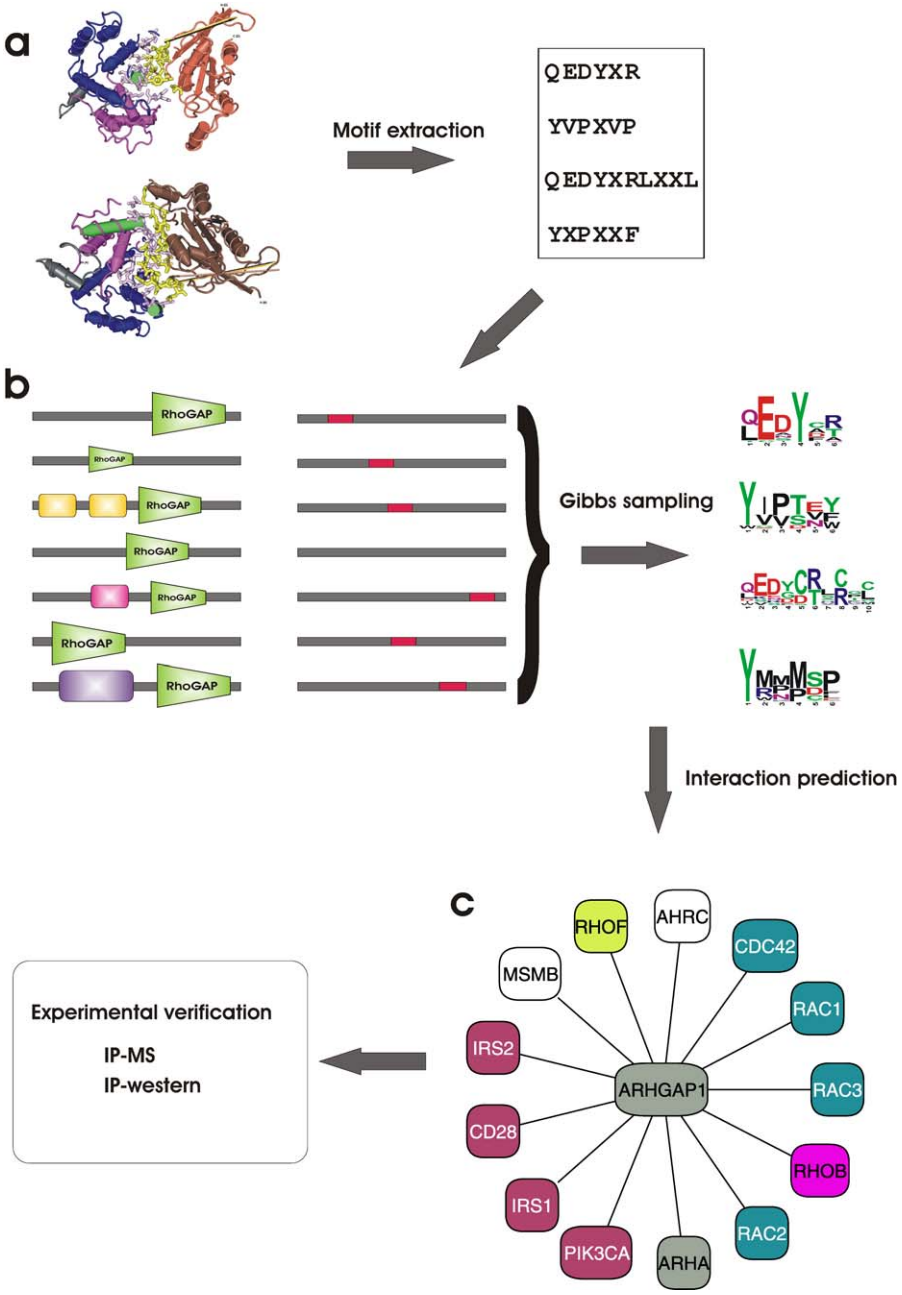

Figure 1. Outline of D-MIST Method for Predicting Protein Interactions by Learned Binding Profiles

Identification of domain-binding profiles begins by extracting the short sequence motifs from structural complexes that contain the domain of interest. (A) In this example, RhoGAP-interacting motifs are extracted from two structural complexes (PDB ID 1AM4, 1TX4) where RhoGAP is bound to small G proteins.

(B) Protein interactions containing the RhoGAP domain were collected from four databases to form the learning set for the Gibbs sampling to generate the binding profiles (shown here as sequence logos [57]). The sampling step is biased towards motifs that are similar to those found in the structural dataset.

(C) The resulting PSSMs are used to predict interactions for proteins with RhoGAP domains, such as the human ARHGAP1. A subset of the predicted interactions is subsequently tested by two experimental methods.

doi:10.1371/journal.pcbi.0030182.g001

discrepancy may be due either to nonspecific interactions detected by IP-MS or to interference of the second epitope tag with some interactions and/or expression levels in vivo. Of the 18 predictions that were tested by IP-western, nine novel interactions were confirmed, and a total of 30 new interactions were identified by both the IP-MS and IP-western methods.

Experimentally Confirmed Predictions

Among the experimentally confirmed predictions were interactions between the five components of the PRS complex, which together compose the 5-phosphoribosyl-1(a)-pyrophosphate synthetase enzyme (EC number 2.7.6.1). This complex is a key component in the production of the precursors for purine, pyrmidine, and pyridine nucleotides

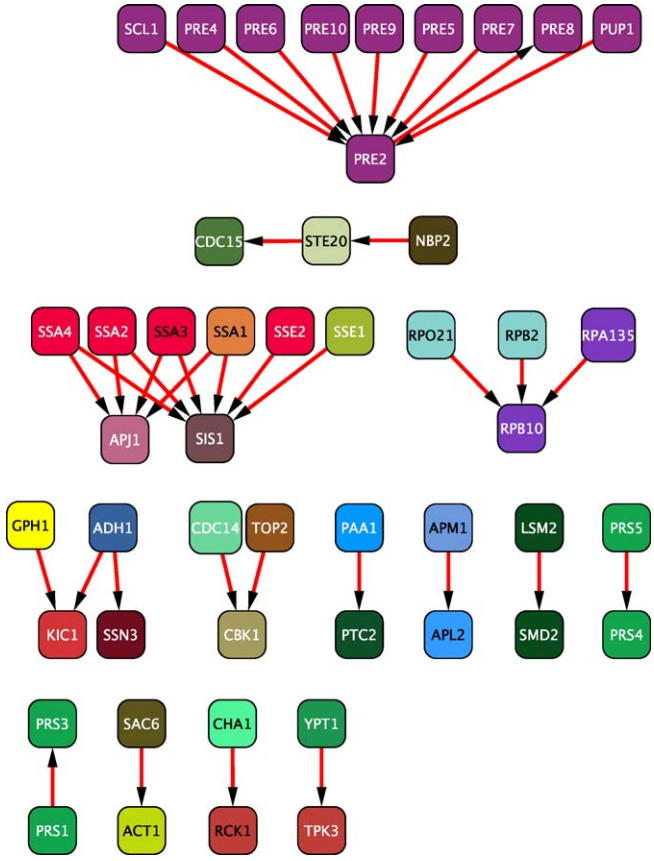

Figure 2. Predicted Interactions Verified by IP-MS

Immunoaffinity purification of bait proteins complexes followed by mass spectrometry identification of associated proteins confirmed 37 predicted interactions. Predictions between proteins that were both co-purified with the tagged bait protein (i.e., both proteins were prey) were not considered validated. Proteins are coloured according to their Gene Ontology biological process annotation.

doi:10.1371/journal.pcbi.0030182.g002

[34]. An additional interaction was confirmed between the alcohol dehydrogenase ($NADP^+$) Adh7 and Prs5, the latter being a member of the PRS complex. This result suggests a possible direct link between NADP/NADPH balance, which is controlled by Adh7 [35], and the biosynthesis of the purine and pyrimidine precursors. A predicted interaction between the histone H2A protein Hta1 and God1, a component of the SWR-C protein complex that incorporates Htz1 into the chromatin, was also confirmed. Chromatin remodelling by the exchange of Hta1 with Htz1 is thought to induce chromatin restructuring that favours gene transcription, RNA polymerase II recruitment, and gene expression induction near silent heterochromatin [36]. Another confirmed interaction is between a member of the HSP40 family (Apj1) with two HSP70 proteins (Ssa1, Ssa2). HSP40 family members form complexes with HSP70 chaperone proteins, which facilitate the folding of specific proteins at various cellular locations [37]. We also identified new interactions between the RNA polymerase II subunit Rpb2 with Rpb10, which is a common subunit of all three RNA polymerases [38]. Additional interaction was demonstrated between Rpc40, a known shared subunit of RNA polymerases I and III, and Rpb2, an exclusive component of RNA polymerase II. It is possible that some of these interactions are bridged or stabilized by other RNA polymerase subunits [39].

One might argue that the above successful predictions could be easily predicted from the orthology of the interacting proteins to the structural complexes used, such as the interactions between members of the PRS complex. We therefore tested several nonobvious predicted interactions that cannot be easily inferred from structural or sequence homology to other interacting pairs. The critical downstream effector of the mitotic exit network is the phosphatase Cdc14, which activates Clb degradation and Sic1 accumulation by dephosphorylation of key substrates [40]. We confirmed an unexpected predicted interaction between Cdc14 and the protein kinase Cbk1, which functions in a parallel pathway (called RAM [regulation of Ace2p activity and cellular morphogenesis]) at the end of mitosis to facilitate cytokinesis and mother–daughter abscission [41]. The Cdc14–Cbk1 interaction suggests that the activity of the mitotic exit network and RAM pathways may be coordinated via Cdc14-mediated dephosphorylation of RAM components and/or Cbk1-mediated phosphorylation of mitotic exit network components [42]. Other nonobvious interactions between known components of clathrin-associated (AP-1) complex Apm1 and Apl2, as well as between components of the RNA splicing complex Smd2 and Lsm2, were detected by the IP-MS experiments but not by IP-western under the conditions used. Given the strength of the D-MIST predictions for these latter interactions, further investigation using more sensitive reagents seems warranted. These confirmed predictions of nonobvious interactions illustrate the potential of the D-MIST approach to generate new biological hypotheses.

Discussion

As noted previously, we excluded additional experimental evidence, such as localization and expression data from our prediction method. Although additional experimental information and functional annotation would likely improve prediction accuracy, it may also limit predictions only to those proteins with prior experimental or functional information. In addition, the use of functional annotation such as Gene Ontology terms (assigned by human experts or predicted computationally) in a prediction method will penalize predicted interactions between proteins with unrelated functions. Therefore, it restricts the ability to predict interactions between apparently unrelated proteins that could illuminate new cellular functions [43].

The D-MIST method for identifying domain-binding modules is currently limited in a number of ways. The first limitation is the availability of detailed binding information, as attained primarily through structural studies and peptide-based approaches such as phage display [44] and random peptide libraries [45]. In addition, several studies have concluded that the repertoire of protein structures in the Protein Data Bank is significantly biased in that *trans*-membrane and disordered domains are underrepresented due to limitations in structure determination [46,47]. Consequently, D-MIST analysis that depends on structural representation of protein interactions is similarly biased. The existing detailed examples of interactions are therefore sparse and noncomprehensive, with only a small subset of all possible domains that is represented. The second limitation

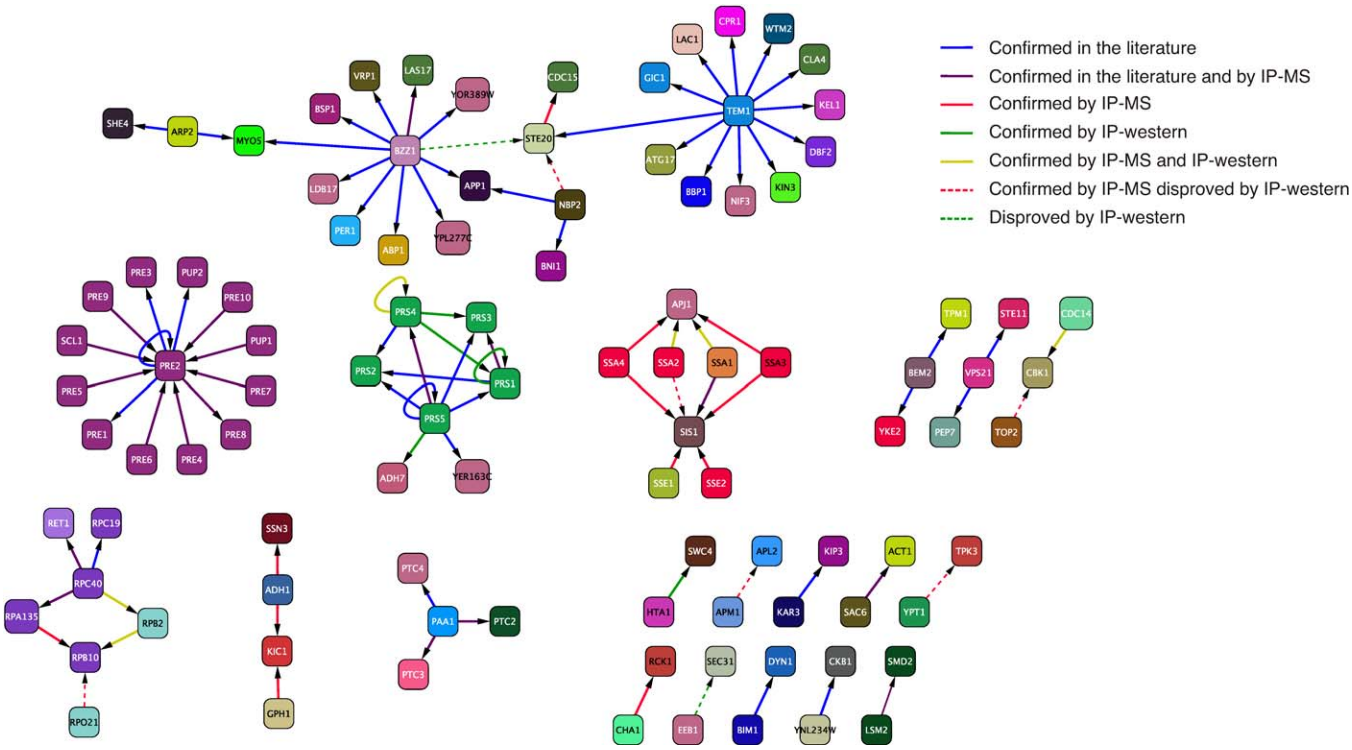

Figure 3. Predicted Interactions Confirmed by Experiments or by Previously Published Results in the Primary Literature

Interactions are coloured according to their verification source. Dashed red lines are predictions that were confirmed by IP-MS but not confirmed by IP-western; dashed green lines are predictions that failed experimental validation by IP-western.

doi:10.1371/journal.pcbi.0030182.g003

is that the derived motifs do not represent the entire repertoire of all possible domain-binding sequences, even for those domains where structural data exist. The third limitation arises from the statistical framework of the Gibbs sampling method that requires a sufficient number of proteins to sample from in order to converge towards a meaningful PSSM. We restricted the analysis to domains with five or more putative interactors, thereby excluding domains that are infrequently found in our set of protein interactions. Fourth, some domains are not amendable to this type of analysis due to the diverse nature of their binding motifs that lack sequence conservation [29]. Last, many interactions are governed by posttranslational modifications or precise physiological states, which may also hamper the accuracy of D-MIST predictions. Despite the above limitations, we have shown that novel protein interactions can be predicted strictly from primary sequence information. D-MIST not only predicts interactions between proteins but also provides sequence level predictions about the binding sites that can be verified experimentally. Predicting protein interactions without the need for additional information or prior experiments is particularly valuable when studying uncharacterized proteins and for predicting interactions in poorly studied organisms where typically only sequence information and predicted open reading frames are available. The sole dependence on sequence information allows for interaction prediction in other organisms without further modifications to the method or input datasets. With the advent of structural genomics initiatives [48], the power of the D-MIST approach will certainly increase.

Methods

Extracting motifs. The domain-binding motifs were extracted from BIND protein interaction records that were generated from 10,064 structures [28]. Interactions were filtered for crystal-packing artifacts using the PQS server [49], and all the interactions are available as a subset of the BIND database. Domain annotation was assigned to the protein structures using our in-house adaptation of CDD [50] with an e-value cutoff of 10×10^{-6} and then converted to InterPro identifiers [51]. Binding motifs are defined as polypeptide segments of five residues or longer in which the amino acids side chains are <5 Å from the interacting domain's side chains on the opposing protein. Two motif residues that are in direct contact with the interacting domain can be separated by a maximum of two noncontacting residues. For example, the first motifs in Figure 1A contain a tyrosine and an arginine that are within 5 Å from the side chains of the RhoGAP domain separated by a distal residue, marked by X, that is not within contact range with the RhoGAP domain.

Learning the binding modules. A total of 87,894 nonredundant protein interactions were collected from 204 species from four database sources: BIND [52], DIP [53], Mint [54], and IntAct [55]. We excluded all interactions that were derived from 3-D studies, high-throughput protein complex identification studies [30,31], or interactions inferred from synthetic lethal experiments. The interactions were indexed in a relational database by domain annotation such that a single query can provide the full list of proteins that interact with a domain of interest (Figure 1B). We used Gibbs sampling [32] seeded with sequence motifs identified in the structural studies to compute a PSSM using the subset of pairwise protein interactions that contain the domain to which the motif was bound in the 3-D structure. The length of the structural motifs was used to approximate the length of the PSSMs. The frequency of residue j at position i in the PSSM (the i,j entry in the matrix) is computed as follows:

$$q_{i,j} = \frac{c_{i,j} + b_j}{N + B} \quad (1)$$

where $c_{i,j}$ is the observed counts of residue j at position i in the sampled proteins, b_j is pseudocounts for residue j, N is the number of

sequences sampled, and B is the total number of pseudocounts for all residues. By increasing the pseudocount term (b_j) for specific positions in the PSSM, the sampling algorithm is biased to favour positions where the residue at position i in the sampled protein is similar to the residue at position i of the structural motif. We set the pseudocounts to equal 62% of the residue counts in the sampled proteins.

Predicting new interactions. Two proteins were predicted to interact if one protein had a domain and a second protein matched one or more of the binding profiles for that domain (Figure 1C). We attempted to predict interactions between all yeast proteins by searching for domain-binding profiles as described in [56] using PSSMs with a score cutoff >10.0 (as scored by the Gibbs sampler) and a cutoff >0.20 for the match between the PSSM and the protein. Potential interactors among the yeast proteome were identified for 703 domain-containing proteins with derived binding profiles. In total, 18,459 interactions were predicted between 2,313 proteins based on the presence of a domain and its binding profile in the interacting pair.

Experimental verification. Recombination-based cloning, culture growth, and protein complex isolation were performed essentially as described [31] with minor modifications. Each uncharacterized open reading frame was tagged at the 3′-end with the FLAG-tag epitope using the Gateway recombination-based cloning system (Invitrogen, http://www.invitrogen.com). Bait complexes were immunopurified on anti-FLAG M2 antibody resin, resolved by denaturing gel electrophoresis, and visualized by colloidal Coomassie stain. Protein identification by automated liquid chromatography tandem mass spectrometry on a Finnigan LCQ DECA ion trap (Thermo Finnigan, http://www.thermo.com) mass spectrometer was as described previously [31]. Predicted protein interactions were also confirmed by IP-western [31] using interaction partners tagged either as C-terminal HA or Myc_3 epitope fusions and detection with 12CA5 anti-HA or 9E10 anti-Myc monoclonal antibodies, respectively (Figure S2).

Overlap with literature. The predicted interactions were compared to a new set of yeast curated interactions collected from more than 50,000 abstracts and publications [33] (available at www.thebiogrid.org). The probability of the observed overlap between the predicted interactions and the literature curated is approximated by a Poisson distribution. A random variable Y has a Poisson distribution if

$$P(y) = \frac{e^{-\lambda}\lambda^{y}}{y!}$$

where $\lambda = Np$, N is the sample size, and p is the probability of a single event; i.e., the probability of selecting a true interaction by random chance. In the current analysis, N is the number of predicted interactions (18,459), y is the number of literature-validated predictions (609), and p is the probability of predicting a correct interaction by random chance for the 703 proteins for which interactions were predicted. The value of p is approximated as the frequency of true interactions among all possible protein pairs that were considered. Since there is no known complete set of interactions for any reference organism, we cautiously assume an upper bound of 100 physiological interactions per bait protein. This number is likely an order of magnitude larger than the true value. Potential interactors for the 703 proteins containing domains with derived binding profiles were identified by scanning the entire yeast proteome (~6,000 proteins) for proteins that matched the domains binding profiles. Hence, the total number of proteins pairs that were considered (i.e., the entire search space) is $703 \times 6{,}000$. The value of p is then $\frac{703 \times 100}{703 \times 6{,}000} \approx 0.017$. Given these parameters P ($y \geq 609$) under a Poisson distribution is 1.0×10^{-13}. Similar calculation using a hypergeometric distribution (sampling without replacement) yields a p-value of 1.0×10^{-8}.

Supporting Information

Dataset S1. Cytoscape Session File Containing the Validated and Predicted Protein Interactions

A Cytoscape session file containing the complete set of predicted interactions as well as the networks in Figures 2, 3, S1, and S2. The networks can be viewed using the Cytoscape network visualization tool freely available at http://www.cytoscape.org.

Found at doi:10.1371/journal.pcbi.0030182.sd001 (2.0 MB ZIP).

Figure S1. The Overlap between the Predicted Interaction Network and a Comprehensive Set of Literature-Curated Interactions [33]

The predicted interactions were compared to a new and exhaustive set of curated interactions extracted from the literature that includes physical interactions from both high-throughput and directed studies as well as genetic interactions. The overlap contains 609 interactions that represent ~3% of the predicted interactions. Proteins are coloured according to Gene Ontology biological process annotation.

Found at doi:10.1371/journal.pcbi.0030182.sg001 (519 KB PDF).

Figure S2. IP-Western Results for the Novel Interactions Predicted by D-MIST

Bait proteins were purified using FLAG antibodies, and their interacting proteins were detected by antibodies specific to C-terminal HA or Myc3 epitopes.

Found at doi:10.1371/journal.pcbi.0030182.sg002 (325 KB PDF).

Text S1. The Domain-Binding Profiles Derived by D-MIST

Each domain-binding profile is specified as a list of sequence motifs. The sequence motifs are used as input to a PSSM search program [56]. Source code available at http://www.people.fas.harvard.edu/~junliu/index1.html#Computational__Biology.

Found at doi:10.1371/journal.pcbi.0030182.sd002 (2.7 MB TXT).

Acknowledgments

We thank Mai Vo, Brett Larsen, Pavel Metalnikov, and Howard Feldman for technical assistance.

Author contributions. DB and MT conceived and designed the experiments and wrote the paper. KEB and DDD performed the experiments. DB and RI analyzed the data. RI contributed reagents/materials/analysis tools. DB conceived and designed the project, and performed the computational work. MT and CWVH directed the study.

Funding. DB was supported by Ontario Graduate Scholarship, and KB was supported by a Canadian Institute of Health Research (CIHR) Training Grant. MT's research is supported by CIHR and Genome Canada; MT holds a Canada Research Chair in Bioinformatics and Functional Genomics. CWVH's research is funded by the Ontario R&D Challenge Fund and by Genome Canada through the Ontario Genomics Institute.

Competing interests. The authors have declared that no competing interests exist.

References

1. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature. 440: 631–636.
2. Jones RB, Gordus A, Krall JA, MacBeath G (2006) A quantitative protein interaction network for the ErbB receptors using protein microarrays. Nature 439: 168–174.
3. Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, et al. (2005) Global analysis of protein phosphorylation in yeast. Nature 438: 679–684.
4. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, et al. (2005) Towards a proteome-scale map of the human protein–protein interaction network. Nature 437: 1173–1178.
5. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, et al. (2005) A human protein–protein interaction network: A resource for annotating the proteome. Cell 122: 957–968.
6. Uetz P, Dong YA, Zeretzke C, Atzler C, Baiker A, et al. (2006) Herpesviral protein networks and their interaction with the human proteome. Science 311: 239–242.
7. Shoemaker BA, Panchenko AR (2007) Deciphering protein–protein interactions. Part I. Experimental techniques and databases. PLoS Comp Biol 3: e42.
8. Pawson T, Nash P (2003) Assembly of cell regulatory systems through protein interaction domains. Science 300: 445–452.
9. Dueber JE, Yeh BJ, Bhattacharyya RP, Lim WA (2004) Rewiring cell signaling: The logic and plasticity of eukaryotic protein circuitry. Curr Opin Struct Biol 14: 690–699.
10. Scott JD, Pawson T (2000) Cell communication: The inside story. Sci Am 282: 72–79.
11. Deng M, Mehta S, Sun F, Chen T (2002) Inferring domain–domain interactions from protein–protein interactions. Genome Res 12: 1540–1548.
12. Ng SK, Zhang Z, Tan SH, Lin K (2003) InterDom: A database of putative

interacting protein domains for validating predicted protein interactions and complexes. Nucleic Acids Res 31: 251–254.
13. Riley R, Lee C, Sabatti C, Eisenberg D (2005) Inferring protein domain interactions from databases of interacting proteins. Genome Biol 6: R89.
14. Sprinzak E, Margalit H (2001) Correlated sequence-signatures as markers of protein–protein interaction. J Mol Biol 311: 681–692.
15. Liu Y, Liu N, Zhao H (2005) Inferring protein–protein interactions through high-throughput interaction data from diverse organisms. Bioinformatics 21: 3279–3285.
16. Guimaraes KS, Jothi R, Zotenko E, Przytycka TM (2006) Predicting domain–domain interactions using a parsimony approach. Genome Biol 7: R104.
17. Aloy P, Bottcher B, Ceulemans H, Leutwein C, Mellwig C, et al. (2004) Structure-based assembly of protein complexes in yeast. Science 303: 2026–2029.
18. Lu L, Arakaki AK, Lu H, Skolnick J (2003) Multimeric threading-based prediction of protein–protein interactions on a genomic scale: Application to the *Saccharomyces cerevisiae* proteome. Genome Res 13: 1146–1154.
19. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, et al. (2003) A Bayesian networks approach for predicting protein–protein interactions from genomic data. Science 302: 449–453.
20. Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, et al. (2005) Probabilistic model of the human protein–protein interaction network. Nat Biotechnol 23: 951–959.
21. Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, et al. (2005) Conserved patterns of protein interaction in multiple species. Proc Natl Acad Sci U S A 102: 1974–1979.
22. Li H, Li J, Tan SH, Ng SK (2004) Discovery of binding motif pairs from protein complex structural data and protein interaction sequence data. Pac Symp Biocomput: 312–323.
23. Neduva V, Linding R, Su-Angrand I, Stark A, de Masi F, et al. (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. PLoS Biol 3: e405.
24. Wang H, Segal E, Ben-Hur A, Koller D, Brutlag D (2005) Identifying protein–protein interaction sites on a genome-wide scale. In Saul LK, Weiss Y, Bottou L, editors. Proceedings of the Conference Advances in Neural Information Processing Systems (NIPS); 13–18 December, 2004; Cambridge, Massachusetts, United States. Vancouver, Canada. pp. 1465–1472.
25. Shen J, Zhang J, Luo X, Zhu W, Yu K, et al. (2007) Predicting protein–protein interactions based only on sequences information. Proc Natl Acad Sci U S A 104: 4337–4341.
26. Hou T, Chen K, McLaughlin WA, Lu B, Wang W (2006) Computational analysis and prediction of the binding motif and protein interacting partners of the Abl SH3 domain. PLoS Comput Biol 2: e1.
27. Reiss DJ, Schwikowski B (2004) Predicting protein–peptide interactions via a network-based motif sampler. Bioinformatics 20 (Supplement 1): I274–I282.
28. Salama JJ, Donaldson I, Hogue CW (2001) Automatic annotation of BIND molecular interactions from three-dimensional structures. Biopolymers 61: 111–120.
29. DeLano WL, Ultsch MH, de Vos AM, Wells JA (2000) Convergent solutions to binding at a protein–protein interface. Science 287: 1279–1283.
30. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415: 141–147.
31. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, et al. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 415: 180–183.
32. Thompson W, Rouchka EC, Lawrence CE (2003) Gibbs recursive sampler: Finding transcription factor binding sites. Nucleic Acids Res 31: 3580–3585.
33. Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hon GC, et al. (2006) Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. J Biol 5: 11.
34. Hernando Y, Carter AT, Parr A, Hove-Jensen B, Schweizer M (1999) Genetic analysis and enzyme activity suggest the existence of more than one minimal functional unit capable of synthesizing phosphoribosyl pyrophosphate in *Saccharomyces cerevisiae*. J Biol Chem 274: 12480–12487.
35. Larroy C, Pares X, Biosca JA (2002) Characterization of a *Saccharomyces cerevisiae* NADP(H)–dependent alcohol dehydrogenase (ADHVII), a member of the cinnamyl alcohol dehydrogenase family. Eur J Biochem 269: 5738–5745.
36. Krogan NJ, Keogh MC, Datta N, Sawa C, Ryan OW, et al. (2003) A Snf2 family ATPase complex required for recruitment of the histone H2A variant Htz1. Mol Cell 12: 1565–1576.
37. Fan CY, Lee S, Cyr DM (2003) Mechanisms for regulation of Hsp70 function by Hsp40. Cell Stress Chaperones 8: 309–316.
38. Cramer P (2002) Multisubunit RNA polymerases. Curr Opin Struct Biol 12: 89–97.
39. Lalo D, Carles C, Sentenac A, Thuriaux P (1993) Interactions between three common subunits of yeast RNA polymerases I and III. Proc Natl Acad Sci U S A 90: 5524–5528.
40. Visintin R, Craig K, Hwang ES, Prinz S, Tyers M, et al. (1998) The phosphatase Cdc14 triggers mitotic exit by reversal of Cdk-dependent phosphorylation. Mol Cell 2: 709–718.
41. Nelson B, Kurischko C, Horecka J, Mody M, Nair P, et al. (2003) RAM: A conserved signaling network that regulates Ace2p transcriptional activity and polarized morphogenesis. Mol Biol Cell 14: 3782–3803.
42. Bidlingmaier S, Weiss EL, Seidel C, Drubin DG, Snyder M (2001) The Cbk1p pathway is important for polarized cell growth and cell separation in *Saccharomyces cerevisiae*. Mol Cell Biol 21: 2449–2462.
43. Brenner S (2002) Life sentences: Ontology recapitulates philology. Genome Biol 3: COMMENT1006. doi:10.1186/gb-2002-3-4-comment1006
44. Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, et al. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science 295: 321–324.
45. Yaffe MB, Cantley LC (2000) Mapping specificity determinants for protein–protein association using protein fusions and random peptide libraries. Methods Enzymol 328: 157–170.
46. Liu J, Rost B (2002) Target space for structural genomics revisited. Bioinformatics 18: 922–933.
47. Peng K, Obradovic Z, Vucetic S (2004) Exploring bias in the Protein Data Bank using contrast classifiers. Pac Symp Biocomput 435–446.
48. Chandonia JM, Brenner SE (2006) The impact of structural genomics: Expectations and outcomes. Science 311: 347–351.
49. Henrick K, Thornton JM (1998) PQS: A protein quaternary structure file server. Trends Biochem Sci 23: 358–361.
50. Marchler-Bauer A, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, et al. (2003) CDD: A curated Entrez database of conserved domain alignments. Nucleic Acids Res 31: 383–387.
51. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, et al. (2003) The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res 31: 315–318.
52. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, et al. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res 33: D418–D424.
53. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, et al. (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 32: D449–D451.
54. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, et al. (2002) MINT: A Molecular INTeraction database. FEBS Lett 513: 135–140.
55. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, et al. (2004) IntAct: An open source molecular interaction database. Nucleic Acids Res 32: D452–D455.
56. Neuwald AF, Liu JS, Lawrence CE (1995) Gibbs motif sampling: Detection of bacterial outer membrane protein repeats. Protein Sci 4: 1618–1632.
57. Schneider TD, Stephens RM (1990) Sequence logos: A new way to display consensus sequences. Nucleic Acids Res 18: 6097–6100.