Supplemental:

Table of Contents:

S1. Methods

- **S1.1 Domain Prediction**
- **S1.2 Domain selection for *de novo* model prediction**
- **S1.3 *De novo* Structure Prediction**
- **S1.4 SCOP Superfamily Prediction**
  - **S1.4.1 BLAST and FFAS domains**
  - **S1.4.2 De Novo Structure Predictions**

S2. Additional Results

- **S2.1 Wash Complex (*Homo sapiens*)**
- **S2.2 Additional *De novo* Superfamily Prediction Analysis**

S3. Additional access and interface information

S4. Supplemental Figure Legends

S5. Supplemental Tables

S6. Supplemental References

S1. Extended Experimental Procedures

S1.1 Domain Prediction

We use the Ginzu pipeline(Chivian et al. 2005) to predict protein domains and domain boundaries. Query sequences are first annotated with secondary structure (PSIPRED), transmembrane helices (TMHMM), signal peptides (SignalP), coiled regions (COILS), and disordered regions (DISOPRED) predictions. A sequence profile is build for each sequence using a 6-pass iterative PSI-Blast search of NCBI's non-redundant protein database (NR) (versions circa 2005-01 and later 2007-10) with an expectation cutoff of 10^{-3}. The sequence profile is then used to search sequences in the PDB using PSI-Blast. FFAS profile-profile search against the PDB used a confidence cutoff of −9.5 (ORFeus (Ginalski et al. 2003) was also used for a small fraction of domains). Regions of the query sequence that meet these criteria are annotated with the matching PDB's structure. The next method in the Ginzu hierarchy searches for Pfam-A families using HMMER (Bateman et al. 1999), a hidden Markov model package and determines domains for matches with an expectation cutoff of 10^{-3} or less. We next use Ginzu's MSA protocol, which creates a multiple sequence alignment built by a 6-pass PSI-Blast search of NCBI's non-redundant sequence set using an expectation cutoff of 10^{-3} and assigns clusters of sequences within the multiple sequence alignment as likely protein domains. The final method, Heuristic, uses four rules to determine protein domain boundaries. First, a multiple sequence alignment built by MSA is searched to find the least occupied regions of the alignment. Second, the method finds portions of the alignment where sequences frequently begin or end. Third, regions of the alignment are identified whose secondary structure are predicted by PSIPRED to be loops. Fourth, confidence in predicted domain boundaries is increased if it is near a predicted PDB-Blast or FFAS03 region. Resulting domains from the genomes analyzed are available on request.

S1.2 Domain selection for *de novo* model prediction

A domain must meet the following requirements to be folded by Rosetta. First, the method which determined the domain must be Pfam, MSA or Heuristic and not PDB-Blast or FFAS03 ensuring no sequence homology to proteins in the PDB. Second, any signal peptides found by SignalP are removed. Third, transmembrane helices are removed if present as predicted by TMHMM. Fourth, the length of the domain must be between 40 and 150 amino acids. Fifth, there must be less than 50% disorder content throughout the domain's as predicted by DISOPRED. Finally, there must be less than 50% coil content throughout the domain's sequence as predicted by COILS. Only if all of these conditions are met is a *de novo* model predicted.

S1.3 *De novo* Structure Prediction

Domains that were not determined by PDB-Blast or FFAS03 and that passed the filtering step (described above) were *de novo* modeled using the Rosetta low resolution protocol (RevDate: 2004/07/09) or all-atom protocol (RevDate: 2005/07/13) to produce structure predictions. The Rosetta all-atom protocol was used for organisms labeled with an asterisk in table S1. The Rosetta algorithm has been described in detail previously (Rohl et al. 2004). Fragment libraries were built using the default "vall" database packaged with Rosetta. Rosetta is freely available to users at academic and nonprofit institutions from Rosetta Commons: http://www.rosettacommons.org/ . In all cases 10-20,000 low scoring conformations were made for each domain. These low scoring conformations are then clustered (Bonneau, 2004) based on root mean squared distance (RMSD) and the centers of the 25 largest clusters are ranked using the Mamoth Confidence Metric, as described below.

Sample low resolution command line: *./rosetta -series 09 -protein ed81 -chain 9 -nstruct 117 -constant_seed -jran 544988 -silent*

Sample all-atom command line: *./rosetta -abrelax -protein ms51 -chain 0 -series 00000 -nstruct 55 -silent -farlx -ex1 -ex2 -output_silent_gz -output_chi_silent -max_attempts 66 -vwu 15206159*

S1.4 SCOP Superfamily Prediction

S1.4.1 BLAST and FFAS domains

Superfamily assignments for BLAST and FFAS03 domains were inferred from sequence alignment overlap with the matched PDB's SCOP classification. We required the query domain to overlap at least 50% of the PDB's SCOP classified domain. Furthermore, many PDB chains are unannotated by SCOP. We currently are unable to provide superfamily annotations for these matches, although, we still report the matched PDB.

S1.4.2 De Novo Structure Predictions

We produce novel superfamily predictions using the clustered representatives from *de novo* structure predictions and the Mammoth Confidence Metric (MCM)(Malmström et al. 2007). Top cluster center structures were compared to SCOP Astral domains(Chandonia et al. 2004) (SCOP version 1.75) using Mammoth (Ortiz et al. 2002). The MCM function classifies a given protein into a SCOP superfamily based on quality of the Mammoth match (z-score), Rosetta convergence, protein contact order in the mammoth matched region, and length ratio between compared sequences. One of three sets of coefficients (primarily α, primarily β, and mixed content) trained using a logistic regression model was applied based on the query protein's secondary structure predictions. The coefficients can be found in Malmström et al. (Malmström et al. 2007). Coefficients were optimized using a benchmark of PDB proteins

grouped based on helix and beta strand content. The resulting MCM score estimates the probability of the query domain being a member of the matched superfamily.

S2. Additional Results

S2.1 Wash Complex (*Homo sapiens*)

The actin cytoskeleton is important for cellular structure and morphogenesis. The actin nucleator, Arp2/3 organizes the actin cytoskeleton when activated by a Nucleation Promoting Factor (NPF). The human Wash Complex (WC) is a seven component NPF of the Arp2/3 complex and is involved in rearranging actin networks during fission of endosomes (Derivery et al. 2009). Wash Complex mediated fission of endosomes also involves microtubules which provide tension during fission. Additionally, in mass spectrometry pulldown experiments, WC associates with tubulin. Human KIAA1033 is an uncharacterized, 1173 residue member of Wash Complex and has no known sequence homology to the PDB. KIAA1033 was predicted by the PFP to have 9 domains, several of which had confident SCOP superfamily classifications (figure S6). Domain 2 is classified as a "t-snare" which may be involved with vesicle binding. Domain 3 is classified as a "PABC (PABP)" which is involved in protein-protein interactions and therefore may be necessary for complex formation. Finally, domain 8 is classified as "Tubulin chaperone cofactor A" which may be partially responsible for the Wash Complex's association with tubulin. This example shows PFP predictions are specific at the protein domain level allowing development of focused experiments for testing in the lab even in large multi-domain proteins where several domains (even after the application of our pipeline) elude annotation.

(http://www.yeastrc.org/pdr/viewProtein.do?id=878270)

S2.2 Additional *De novo* Superfamily Prediction Analysis

It has been shown that the Rosetta *de novo* protocol performs differently on proteins of different secondary structure content. For example, predictions on proteins with large β-sheet content and/or high contact order tend to be much less accurate (Bonneau et al. 2002a). To determine the accuracy of our method on different fold classes, we separated the SAP set based on the predicted SCOP class. At its topmost level SCOP divides protein structures into four classes based on the arrangement of secondary structure: class A (α proteins), class B (β proteins), class C (β-α-β proteins) and class D (segregated α and β). As shown in figure 2 (top and middle panels) and table 2 of the main text, superfamily classification accuracies and yields vary

between SCOP classes. Specifically, classes A and D have similar yields of high confident predictions, 45.5% (138 high confidence / 303 total predicted) and 41.8% (142/340), respectively. Class B yields less high-confident predictions, 11% (11/97) but 9 of those predictions were correct. The low yield but high accuracy for class B shows our classifier is robust to Rosetta's limited ability for folding β proteins. Furthermore, high confident predictions for classes A, C and D are correct a majority of the time, 69%, 73% and 88% respectively. These results show the superfamily classifier is accurate and yields substantial numbers of high quality structural assignments to further proteome annotation.

A previous study found smaller proteins to have higher absolute *de novo* model accuracy but poor superfamily predictive power relative to larger proteins (Bonneau et al. 2002b). We split our SAP set into domains greater than and less than 100 amino acids to determine if this holds true for our data (tables S3a and S3b). Consistent with this earlier result, domains between 100 and 150 residues in length (83%) outperform smaller domains (66%) in high confident superfamily prediction accuracy. Although the sample size is small, this supports the previous finding that the length of a protein domain is an important predictor of successful superfamily assignment.

S3. Additional access and interface information

We provide three main interfaces to data produced by the PFP. First, the web database (http://www.yeastrc.org/pdr/) displays general information (e.g. name, organism, GO annotations, sequence) of individual proteins as well as the graphical representations of primary sequence annotations (e.g. PSIPRED, DISOPRED) and domain predictions (figure S5 B-F). For each predicted domain, the detection method used (e.g. FFAS03, Pfam), a confidence and when available a graphic of the matched three-dimensional structure from the PDB is shown. Confident GO molecular function predictions are also displayed per domain. The web database is searchable by standard protein identifiers (e.g. gi, refseq); Gene Ontology terms (e.g. search all proteins with an annotation of cellular component: coated membrane GO:0048475) and also provides capability to limit queries by an organism of interest. Additionally, we provide a Blast interface search page (http://pfp.bio.nyu.edu/blast/index), which searches sequences processed through the PFP.

Another interface provided is BioNetBuilder (http://err.bio.nyu.edu/cytoscape/bionetbuilder/), a Cytoscape plugin, which is intended to give users a high level view of predicted annotations in relation to other related proteins (figure S5A). Users can create protein protein interaction networks based on popular databases (e.g. DIP, BIND) and view results from the PFP

represented as varying sizes, shapes and colors. Different visual styles can be applied to view structure annotations and function predictions. Using the structure annotation visual style (hpf_structure), the size of the node represents the total number of domains predicted for the protein. The shape of a node represents whether there is complete coverage of structure annotations for all domains (i.e. diamonds) or incomplete coverage (i.e. circle). Finally, the color of the node represents the quality of the top prediction where blue is high quality, light blue is medium quality and brown is low quality. The function prediction visual style (hpf_function) allows the user to view GO function predictions available for the proteins in the network. This style colors the nodes with a gradient from brown (low confidence) to blue (high confidence) based on the log likelihood ratio produced by the function prediction method. The shape of the node is a square if the function predicted is general (> ~2% of all proteins), a circle if the function predicted is specific (< ~2% of all proteins) or a diamond if very specific (<~0.04% of all proteins). The size of the node is proportional to a value metric, which is based on both the confidence and specificity of the function prediction. Value, V is calculated by V = P(confidence)*-log(P(function)) where P(confidence) is the confidence represented as a probability and P(function) is the prior probability of the molecular function. This metric weights very specific terms that have been predicted with high confidence as having high value and general terms predicted with low confidence as having low value. Each node in the BioNetBuilder network has, as attributes, predicted structure and functions including scores, values, coverage and a link to the web database for more detailed information. A tutorial (http://err.bio.nyu.edu/pfp/tutorial/) is provided to walk new users through the steps of building a BioNetBuilder network, selecting visual styles, linking to and navigating the web database.

S4. Supplemental Figure Legends

Figure S1: Coverage of Ginzu domain types, Related to Figure 1. Proteomes from 94 organisms were processed using the domain prediction protocol Ginzu. Shown are organisms from both eukaryotes and prokaryotes including two eukaryotic parasites, *T. cruzi* and *P. vivax*. Approximately 50%-70% of domains in the model organisms can be assigned a structural annotation based on the sequence based methods of PDB-Blast (light blue) and fold recognition (dark blue).

Figure S2: Precision/Yield on the Solved After Prediction (SAP) benchmark, classified with SCOP v1.67 (bottom line) and v1.75 (top line), Related to Figure 2. The plot shows the percentage of protein domains in the SAP set classified using SCOP v1.67 and v1.75 for varying precisions.

Figure S3: MCM score correlates with native structural alignment z score, Related to Figure 2. The top Rosetta decoy for 875 domains in the SAP set was structurally compared to its recently solved structure in the PDB using Mammoth. This plot shows that Rosetta decoys with high confident MCM scores align well with their native PDB structures. The lower panel is a bar chart of structure counts in each MCM score bin. The bar chart shows a substantial portion of domains is predicted with high confidence MCM scores.

Figure S4: Examples of model structure similarity to recently solved PDB structures, Related to Figure 2. (**A**) *M. musculus* Rwdd1 & 2ebma, (**B**) *T. maritima* TM1266 & 2nzca, (**C**) *S. aureus* SAAV_0614 & 3ct6a, (**D**) *D. psychrophila* DP2708 & 1yyv, (**E**) *Listeria monocytogenes* lmo035 & 3b48a, (**F**) *Geobacter sulfurreducens* AAR34733 & 2a3qa. Rosetta models and native PDB structures are displayed in the first and second columns respectively. The third column shows the PDB structure where red represents structural alignment with the model (portions of the solved structure that correctly align to the predicted model within 4 Å) and gray represents incorrectly predicted regions. The fourth column is the z score returned by Mammoth.

Figure S5: PFP Results are easily accessible through several interfaces, Related to Figure 1. BioNetBuilder (**A**, http://err.bio.nyu.edu/cytoscape/bionetbuilder/) is a Cytoscape plugin that allows the construction of protein-protein interaction networks and the automatic formatting of node size, shape and color which correspond to PFP structure classifications and available function predictions. Active hyperlinks allow access to the web database with a single click (**B**). The web database (http://www.yeastrc.org/pdr/) is searchable directly using many gene identifiers and displays domain information, structure results and function predictions (**C**). Individual domains can be analyzed further along with any function predictions available (**D & F**) and *de novo* structures can be viewed using an online molecular viewer (**E**). A tutorial is available at http://err.bio.nyu.edu/pfp/tutorial/ .

Figure S6: Domain layout for *H. sapiens* KIAA1033 of the Wash Complex, Related to Figure 4. Domain boundaries of KIAA1033 were predicted using Ginzu resulting in 9 distinct domains. Three of the domains, domains 2 (yellow), 3 (magenta) and 8 (green), had moderate to high confident *de novo* predictions. Domain 2, 3 and 8 are classified into "t-snare", "PABC (PABP)" and "Tubulin chaperone cofactor A" superfamilies and the *de novo* model is aligned to the matched superfamily representative (blue) below with structural alignment z scores of 12.09, 9.35 and 10.09 respectively. Also shown above is predicted secondary structure (SS) where red is helix and blue is strand, disorder (D), transmembrane regions (TM), coiled coil (CC), signal peptides (SP) and conservation (PS)

Figure S7: Precision vs Recall graph for function predictions made for domains in the SAP *de novo* set, Related to Figure 3. The graph shows the precision of function predictions versus recall for sequences that were structurally classified by *de novo* and were members of the solved after predicted (SAP) set. The red lines represent the prediction method using GO Process and Component (PC). The green lines represent the prediction method using GO Process, Component

and Structure (PCS). Although this benchmark has a limited number of domains present, the graph shows adding structure information from *de novo* improves precision for function prediction.

S5. Supplemental Tables

See excel spreadsheet labeled tableS1.xls.

Table S1: Protein and Domain Structural Annotation by Organism

SCOP class	Total (%)	Total correct (%)	MedConf (%)	MedConf correct (%)	Yield MedConf	HighConf (%)	HighConf correct (%)	Yield HighConf
A	339 (38.7%)	47 (13.9%)	146 (43.1%)	44 (30.1%)	16.7%	77 (22.7%)	37 (48.1%)	8.8%
B	113 (12.9%)	6 (5.3%)	22 (19.5%)	4 (18.2%)	2.5%	7 (6.2%)	2 (28.6%)	0.8%
C	140 (16.0%)	17 (12.1%)	50 (35.7%)	12 (24.0%)	5.7%	22 (15.7%)	11 (50.0%)	2.5%
D	256 (29.3%)	22 (8.6%)	91 (35.5%)	16 (17.6%)	10.4%	38 (14.8%)	9 (23.7%)	4.3%
Other	27 (3.1%)	0 (0.0%)	7 (25.9%)	0 (0.0%)	0.8%	2 (7.4%)	0 (0.0%)	0.2%
All	875	92 (10.5%)	316 (36.1%)	76 (24.1%)		146 (16.7%)	59 (40.4%)	

Table S2: Superfamily classifications for SAP structures using previous version of SCOP (v1.67). Percents are in parentheses. MedConf and HighConf columns have MCM scores > 0.8 and > 0.9 respectively.

(a) protein domains < 100aa length

SCOP class	Total (%)	Total correct (%)	MedConf (%)*	MedConf correct (%)*	Yield MedConf*	HighConf (%)**	HighConf correct (%)**	Yield HighConf**
A	129 (37.4%)	28 (21.7%)	64 (49.6%)	23 (35.9%)	18.6%	40 (31.0%)	17 (42.5%)	11.6%
B	48 (13.9%)	21 (43.8%)	18 (37.5%)	11 (61.1%)	5.2%	10 (20.8%)	9 (90.0%)	2.9%
C	19 (5.5%)	2 (10.5%)	11 (57.9%)	2 (18.2%)	3.2%	3 (15.8%)	2 (66.7%)	0.9%
D	133 (38.6%)	78 (58.6%)	79 (59.4%)	60 (75.9%)	22.9%	44 (33.1%)	35 (79.5%)	12.8%
Other	16 (4.6%)	7 (43.8%)	8 (50.0%)	6 (75.0%)	2.3%	8 (50.0%)	6 (75.0%)	2.3%
All	345	136 (39.4%)	180 (52.2%)	102 (56.7%)		105 (30.4%)	69 (65.7%)	

(b) protein domains >= 100aa length

SCOP class	Total (%)	Total correct (%)	MedConf (%)*	MedConf correct (%) *	Yield MedConf *	HighConf (%)**	HighConf correct (%)**	Yield HighConf**
A	174 (32.8%)	94 (54.0%)	122 (70.1%)	83 (68.0%)	23.0%	98 (56.3%)	78 (79.6%)	18.5%
B	49 (9.2%)	20 (40.8%)	9 (18.4%)	1 (11.1%)	1.7%	1 (2.0%)	0 (0.0%)	0.2%
C	94 (17.7%)	33 (35.1%)	48 (51.1%)	28 (58.3%)	9.1%	30 (31.9%)	22 (73.3%)	5.7%
D	207 (39.1%)	124 (59.9%)	130 (62.8%)	110 (84.6%)	24.5%	98 (47.3%)	90 (91.8%)	18.5%
Other	6 (1.1%)	0 (0.0%)	2 (33.3%)	0 (0.0%)	0.4%	1 (16.7%)	0 (0.0%)	0.2%
All	530	271 (51.1%)	311 (58.7%)	222 (71.4%)		228 (43.0%)	190 (83.3%)	

***MCM score > 0.8**

****MCM score > 0.9**

Table S3: Superfamily classifications for SAP structures. (a) Protein domains in SAP < 100 amino acids. (b) Protein domains in SAP >= 100 amino acids. Comparison of (a) and (b) show, in general, a higher accuracy of superfamily prediction when protein domains are >= 100 amino acids.

See excel spreadsheet labeled tableS4.xls.

Table S4:Gene Ontology Molecular Function Predictions by Organism

See excel spreadsheet labeled tableS5.xls.

Table S5:Solved After Predicted (SAP) set

S6. Supplemental References

Bateman, A., E. Birney, R. Durbin, S.R. Eddy, R.D. Finn, and E.L. Sonnhammer. 1999. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res* **27:** 260-262.

Bonneau, R., I. Ruczinski, J. Tsai, and D. Baker. 2002a. Contact order and ab initio protein structure prediction. *Protein Sci* **11:** 1937-1944.

Bonneau, R., C. Strauss, C. Rohl, D. Chivian, P. Bradley, L. Malmstrom, T. Robertson, and D. Baker. 2002b. De novo prediction of three-dimensional structures for major protein families. *Journal of Molecular Biology* **322:** 65-78.

Chandonia, J.M., G. Hon, N.S. Walker, L. Lo Conte, P. Koehl, M. Levitt, and S.E. Brenner. 2004. The ASTRAL Compendium in 2004. *Nucleic Acids Res.* **32:** D189--192.

Chivian, D., D.E. Kim, L. Malmström, J. Schonbrun, C.A. Rohl, and D. Baker. 2005. Prediction of CASP6 structures using automated Robetta protocols. *Proteins* **61 Suppl 7:** 157-166.

Derivery, E., C. Sousa, J.J. Gautier, B. Lombard, D. Loew, and A. Gautreau. 2009. The Arp2/3 Activator WASH Controls the Fission of Endosomes through a Large Multiprotein Complex. *Developmental Cell* **17:** 712-723.

Ginalski, K., J. Pas, L.S. Wyrwicz, M. von Grotthuss, J.M. Bujnicki, and L. Rychlewski. 2003. ORFeus: Detection of distant homology using sequence profiles and predicted secondary structure. *Nucleic Acids Res* **31:** 3804-3807.

Malmström, L., M. Riffle, C.E.M. Strauss, D. Chivian, T.N. Davis, R. Bonneau, and D. Baker. 2007. Superfamily Assignments for the Yeast Proteome through Integration of Structure Prediction with the Gene Ontology. *PLoS Biol* **5:** e76.

Ortiz, A.R., C.E. Strauss, and O. Olmea. 2002. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.* **11:** 2606--2621.

Rohl, C.A., C.E.M. Strauss, K.M.S. Misura, and D. Baker. 2004. Protein structure prediction using Rosetta. *Methods Enzymol* **383:** 66-93.