Superfamily Assignments for the Yeast Proteome through Integration of Structure Prediction with the Gene Ontology

Lars Malmström[1], Michael Riffle[1], Charlie E. M. Strauss[2], Dylan Chivian[1¤], Trisha N. Davis[1], Richard Bonneau[3], David Baker[1,4*]

1 Department of Biochemistry, University of Washington, Seattle, Washington, United States of America, **2** Los Alamos National Laboratory, Los Alamos, New Mexico, United States of America, **3** Department of Biology, Department of Computer Science, and Center for Comparative Functional Genomics, New York University, New York, New York, United States of America, **4** Howard Hughes Medical Institute, University of Washington, Seattle, Washington, United States of America

Saccharomyces cerevisiae **is one of the best-studied model organisms, yet the three-dimensional structure and molecular function of many yeast proteins remain unknown. Yeast proteins were parsed into 14,934 domains, and those lacking sequence similarity to proteins of known structure were folded using the Rosetta de novo structure prediction method on the World Community Grid. This structural data was integrated with process, component, and function annotations from the *Saccharomyces* Genome Database to assign yeast protein domains to SCOP superfamilies using a simple Bayesian approach. We have predicted the structure of 3,338 putative domains and assigned SCOP superfamily annotations to 581 of them. We have also assigned structural annotations to 7,094 predicted domains based on fold recognition and homology modeling methods. The domain predictions and structural information are available in an online database at http://rd.plos.org/10.1371_journal.pbio.0050076_01.**

Citation: Malmström L, Riffle M, Strauss CEM, Chivian D, Davis TN, et al. (2007) Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology. PLoS Biol 5(4): e76. doi:10.1371/journal.pbio.0050076

Introduction

The yeast *Saccharomyces cerevisiae* is one of the most widely studied organisms, yet a large fraction of its proteins are of unknown structure and/or unknown function. Knowledge of the structure of a protein is critical to understand how it functions, and hence, a complete set of protein structures for yeast is desirable, but difficult to accomplish experimentally.

The accuracy of de novo structure prediction methods, although far from the accuracy of experimental structures, has improved in recent years. The Rosetta de novo structure prediction method [1–4] is currently one of the best methods available for predicting the structure of proteins lacking obvious homology to known structures [5–8]. Application of Rosetta to genome-wide annotation has been limited by the difficulty of distinguishing accurate from inaccurate predictions and the computational cost associated with scaling the procedure to whole genomes. Initial results have been encouraging, showing promise on subsets of protein families and prokaryotic genomes [9,10]. We have previously [11] predicted structures for short Pfam families without structural information, and showed that a simple confidence function could partially separate correct structure predictions from incorrect predictions.

There is a rich body of work on the relationship between superfamily (encoded in databases such as SCOP [12–14] and CATH [15]) and function (encoded in databases such as Kyoto Encyclopedia of Genes and Genomes [KEGG] [16] and Gene Ontology [GO] [17]). Although many superfamilies have been shown to carry out multiple functions, Hegyi and Gerstein [18] found that the majority of structure superfamilies carry out one or a few molecular functions, and conversely, that the majority of functions are carried out by one or a few SCOP superfamilies. This relationship can be exploited when predicting to which structure superfamily a protein belongs [9].

We describe an integrated approach for assigning protein domains to structure superfamilies that combines de novo structure predictions with GO function, process, and component annotations. We first parse all yeast proteins into putative structural domains using the Ginzu method [7,19]. Ginzu predicts domain boundaries by applying a hierarchy of sequence-based methods beginning with searching for homologs of known structure using PSI-BLAST [20] and ending by parsing block patterns in multiple sequence alignments (MSAs). After running Ginzu on the full proteome, we applied the Rosetta structure prediction method to domains shorter than 150 amino acids for which no homolog of known structure was found. The top structure predictions were compared to protein domains of known structure using the MAMMOTH protein structure comparison program [21]. The reliability of an assignment to a protein structure superfamily

Academic Editor: Andrej Sali, University of California San Francisco, United States of America

Received May 4, 2006; **Accepted** January 12, 2007; **Published** March 20, 2007

Abbreviations: CO, contact order; GO, Gene Ontology; HPF, Human Proteome Folding Project; MCM, MAMMOTH confidence metric; MSA, multiple sequence alignment; ORF, open reading frame; PDB, the Protein Data Bank

* To whom correspondence should be addressed. E-mail: dabaker@u.washington.edu

¤ Current address: Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America

Author Summary

The three-dimensional structure of a protein can reveal much about that protein's evolutionary relationships and functions. Such information about all the proteins in an organism—the proteome—would offer a more global view of these relationships, but solving each structure individually would be a formidable task. In this study, we have parsed all *Saccharomyces cerevisiae* proteins into nearly 15,000 distinct domains and then used de novo structure prediction methods together with worldwide distributed computing to predict structures for all domains lacking sequence similarity to proteins of known structure. To overcome the uncertainties in de novo structure prediction, we combined these predictions with data on the biological process, function, and localization of the proteins from previous experimental studies to assign the domains to families of evolutionarily related proteins. Our genome-wide domain predictions and superfamily assignments provide the basis for the generation of experimentally testable hypotheses about the mechanism of action for a large number of yeast proteins.

derived from these structure comparisons was evaluated using a logistic regression–based confidence function optimized on a large training set of Rosetta models for proteins of known structure. Superfamily predictions of increased accuracy were obtained by integrating GO function, component, and process annotations [17,22] from the *Saccharomyces* Genome Database [23] with the structure prediction data using a simple Bayesian approach. We predicted structures for 3,338 domains and have annotated 581 of them with novel SCOP superfamily assignments. The domain predictions, the predicted structures, and superfamily assignments are accessible at http://rd.plos.org/10.1371__journal.pbio.0050076__01.

Results

Predicting Structural Domains

A total of 6,238 open reading frames (ORFs) were parsed into structural domains using Ginzu [7,19]. Ginzu was used successfully in Critical Assessment of Techniques for Protein Structure Prediction 6 (CASP6) to delineate domains within query proteins by sequentially searching for (1) sequence-detectable homology to the Protein Data Bank (PDB) using PSI-BLAST [20], (2) more-remote fold recognition hits to PDB structures [24,25], (3) hits to Pfam-conserved sequence family domains [26,27], and (4) block patterns in MSAs. This hierarchical application of methods is organized so that methods providing more reliable information are applied first, thus accuracy is not sacrificed as we apply multiple methods in an attempt to maximize comprehensive coverage of the genome. A total of 14,934 domains were predicted, of which 38% had a sequence-detectable homolog of known structure, and an additional 9% could confidently be annotated by fold recognition methods. A summary of the genome-wide domain parses is presented in Table 1, and a complete list of domain predictions are presented in Table S1.

Fold Recognition

Although the confident fold recognition results generated as part of this study are not the main focus of this paper, they provide a wealth of information on proteins for which there are no detectable sequence homologs of known structure. The results for 1,361 domain annotations using fold recognition are detailed in Table S1 and are available at http://rd.plos.org/10.1371__journal.pbio.0050076__01.

Protein Structure Prediction

A total of 4,006 yeast protein domains shorter than 150 amino acids (a practical length limit for the Rosetta method) and not linked to known structures by PSI-BLAST or fold recognition methods were identified by Ginzu: 668 of these contained predicted transmembrane helices and were omitted; the remaining 3,338 domains were folded using the Rosetta de novo method. Ten thousand structure models were generated for each of these remaining 3,338 domains using the Rosetta de novo method [1,2,28] and then condensed to 30 representative models by clustering. The size of the calculation is significant and is estimated at 12 million CPU hours, or 1,350 CPU years. This calculation was performed on the World Community Grid (WCG) parallel grid computing facility provided by IBM (http://wcgrid.org).

Superfamily Assignment by Structure Comparison

The 30 representative models for each domain were compared to a database of experimentally determined protein structure domains (based on ASTRAL; see Materials and Methods) with representatives from all SCOP (version 1.67) superfamilies and evaluated using a confidence function (referred to as the MAMMOTH Confidence Metric [MCM]) described below. The confidence of a given prediction for a given protein-domain is estimated based on features resulting from the Rosetta structure prediction, clustering, and structure–structure matching steps (using MAMMOTH [21]). The primary improvement in the confidence function over our previous work [11] is the inclusion of the contact order

Table 1. Summary of Domain Assignments for the Yeast Genome Made Using the Ginzu Method

Detection Method	# ORFs	# Domains	Number of Residues per Domain			
			Average	Standard Deviation	Maximum	Minimum
Fold recognition	779	1,361	181.5	107.0	792	40
MSA	1,721	2,286	229.1	143.0	1,325	52
Pfam	797	973	219.2	170.3	1,427	33
PSI-BLAST	2,912	5,733	185.7	109.6	1,561	33
Unassigned	3,855	4,581	175.9	163.2	2,220	4

The hierarchy of domain detection methods used by Ginzu is listed in the first column.
doi:10.1371/journal.pbio.0050076.t001

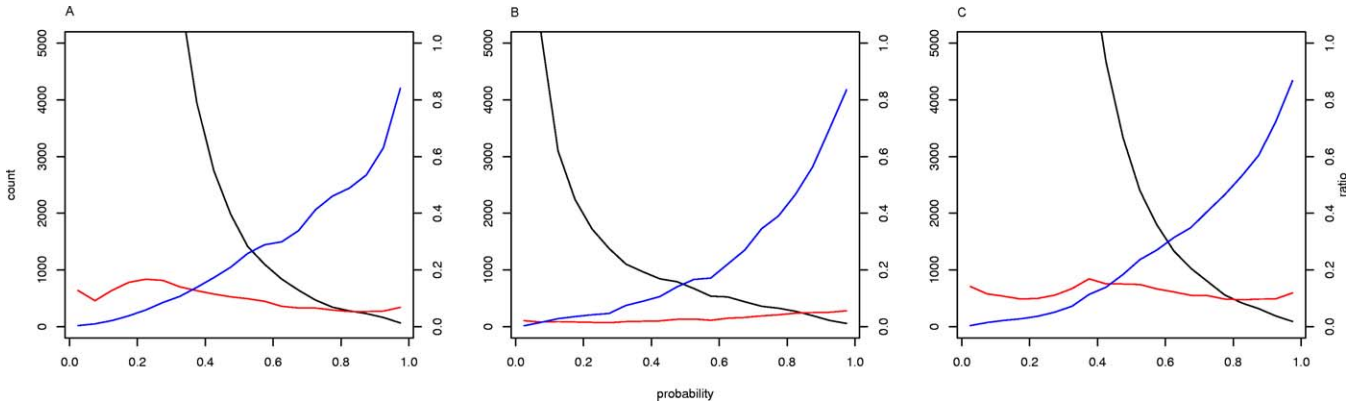

Figure 1. Performance of MAMMOTH Confidence Metric on Benchmark Set

(A) Alpha-helical protein model, (B) beta-sheet protein model, and (C) alpha-helical/beta-sheet protein model. The *x*-axis is the P_{MCM}; the *y*-axis the number of matches that are incorrect (black) and correct (red). The blue line is the ratio (correct/incorrect).

doi:10.1371/journal.pbio.0050076.g001

(CO; average sequence separation of contacting amino acids [29]) of the residues superimposed in the MAMMOTH structure–structure alignment of the predicted structure with the matched structure; this CO term penalizes less-significant matches dominated by local contacts such as single long alpha helices. Figure 1 describes the performance of the MCM on a large benchmark set developed for this study (see Materials and Methods). The MCM score, P_{MCM}, ranges between 0 and 1 and is an estimate of the probability of the identification of the correct structure superfamily identification. A total of 404 domains in the yeast dataset (see Table 2) have a P_{MCM} above 0.8 (considered significant for the purpose of this discussion) and can be found in Table S2 (additional domains are annotated with superfamily via integration with GO as described below).

The confidence estimates derived from our SCOP benchmark set are likely to be somewhat inflated when applied to the yeast protein set for two reasons; first, as discussed in the following section, the domain boundaries are derived directly from experimental structures in our SCOP benchmark, but are subject to error for the yeast proteins, and second, in the SCOP benchmark set, there is by construction always at least one closely related structure in the correct superfamily, whereas proteins with novel folds in yeast may not belong to any pre-existing superfamily. Below and in Materials and Methods, we describe tests on two additional validation sets that include the above sources of error (and thus allow for the estimation of the effects of such errors on structure superfamily prediction). Although there is a non-negligible presence of errors in domain parsing and superfamily assignment, our results show that the superfamily assignments generated herein (see Table S2) should be valuable for stimulating the generation of experimentally testable hypotheses about the structure and often the mechanism of action of these proteins.

Superfamily Assignment through Integration of Structure Predictions with Function

There is a strong relationship between the function of a protein and its structural superfamily [18]. Most commonly, proteins in the same superfamily carry out one or a few functions. The reverse is also true; often only one or a few superfamilies are found to carry out a specific function. We derived probability distributions, $P(\mathrm{GO}|\mathrm{SF})$, that relate SCOP superfamily (SF) to molecular function, biological process, and cellular component (GO). We also constructed probability distributions, $P(\mathrm{SF}|\mathrm{D})$, that give the probability of a given superfamily, given the predicted structures (D), that is derived from the distributions of P_{MCM} for a target, as described in Materials and Methods. These distributions were integrated to determine the degree to which a superfamily prediction is simultaneously compatible with the structure predictions and the functional annotation available for a given protein, using:

$$P\left(\mathrm{SF}|\vec{\mathrm{D}},\mathrm{GO}\right) = \frac{P(\mathrm{GO}|\mathrm{SF}) \cdot P(\mathrm{SF}|\vec{\mathrm{D}})}{P(\mathrm{GO})} \qquad (1)$$

where $P(\mathrm{SF}|\mathrm{D},\mathrm{GO})$ is the probability that the domain belongs to SCOP superfamily SF, given the predicted structures, D, and the GO terms, GO, for the protein. The independence assumption underlying Equation 1 is described in Materials and Methods.

The superfamily distributions derived from the structure prediction data alone ($P(\mathrm{SF}|\mathrm{D})$), the GO annotations ($P(\mathrm{SF}|\mathrm{GO})$), and from the two together ($P(\mathrm{SF}|\mathrm{D},\mathrm{GO})$), are compared in Figure 2 for four proteins for which the true SCOP superfamilies are known, showing the synergy between the two sources of information. The ambiguities in $P(\mathrm{SF}|\mathrm{D})$ (red line) and $P(\mathrm{SF}|\mathrm{GO})$ (blue line) are reduced upon integration $P(\mathrm{SF}|\mathrm{D},\mathrm{GO})$ (black line), resulting in less ambiguous predictions for many difficult-to-annotate domains. The overall performance for the $P(\mathrm{SF}|\mathrm{D},\mathrm{GO})$ over the benchmark

Table 2. Overview of Domain Annotations Using All Methods Employed in This Work

Domain Annotations	# Domains
Domains annotated by PSI-BLAST	5,733
Domains annotated by fold recognition	1,361
Domains annotated with $P_{GI} \geq 0.8$	177
Domains annotated with $P_{MCM} \geq 0.8$	404

doi:10.1371/journal.pbio.0050076.t002

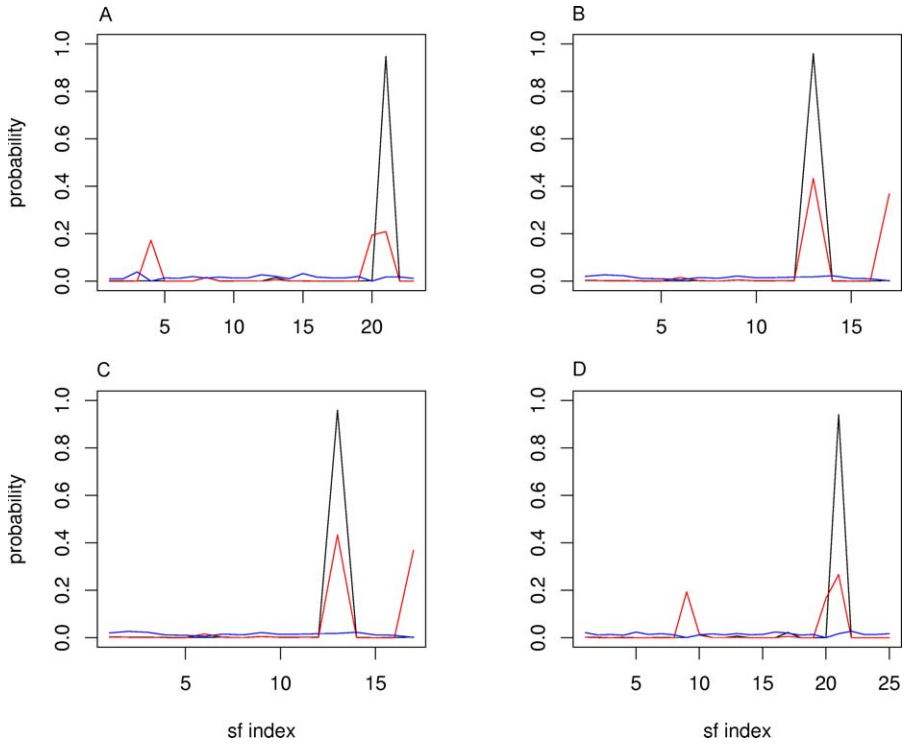

Figure 2. Integration of Structure Prediction with GO Annotations

Red line represents the superfamily distribution for the predicted structures, *P*(SF|D); blue line, the superfamily distribution based on GO annotations, *P*(SF|GO). Black line represents the Bayesian combination (*P*(SF|D,GO); Equation 2). Only superfamilies with a probability over 0.001 in either category are displayed. The names of the proteins and the GO annotations for which the black line is derived are (A) 1KMDA (Vam7p Px Domain)/Golgi to vacuole transport (process), (B) 1IOUA (v-SNARE)/vesicle fusion (process), (C) 1F32 (*Ascaris* pepsin inhibitor-3)/endopeptidase inhibitor activity (function), and (D) 1DUJA (Spindle Assembly Checkpoint protein Human Mad2)/Chromosome (component).

doi:10.1371/journal.pbio.0050076.g002

set (see Materials and Methods) is shown in Figure 3. A total of 177 yeast domains (see Table 2) were assigned a structural superfamily with a *P*(SF|D,GO) over 0.8 (Table S3).

Internal Standards—Additional Validation of Confidence Metric Using Proteins Solved after Calculations Were Completed

True performance of these technologies cannot be assessed on the benchmark dataset because the domain boundaries of this set are perfect (derived from known structures in the ASTRAL database). A subset of the proteins without links to known structure at the start of this project now have strong homology to a structure that has since been solved, see Figure 4 for examples. These recently solved structures give us an opportunity to assess the performance of our technology without bias in the selection of the proteins, with real domain prediction error incorporated, and without the contamination of the results by weak homology to known structures.

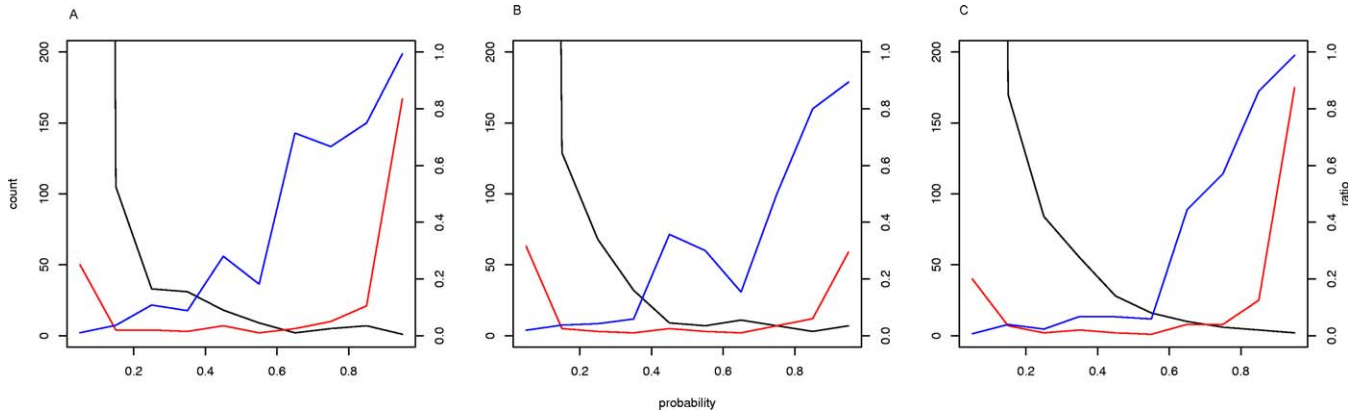

Figure 3. Performance of the GO Integration on Benchmark Set

(A) Alpha-helical proteins, (B) beta-sheet proteins, and (C) alpha-helical/beta-sheet proteins. The *x*-axis is *P*(SF|D,GO); the *y*-axis the number of matches with that score that are incorrect (black) and correct (red). The blue line is the ratio (correct/incorrect).

doi:10.1371/journal.pbio.0050076.g003

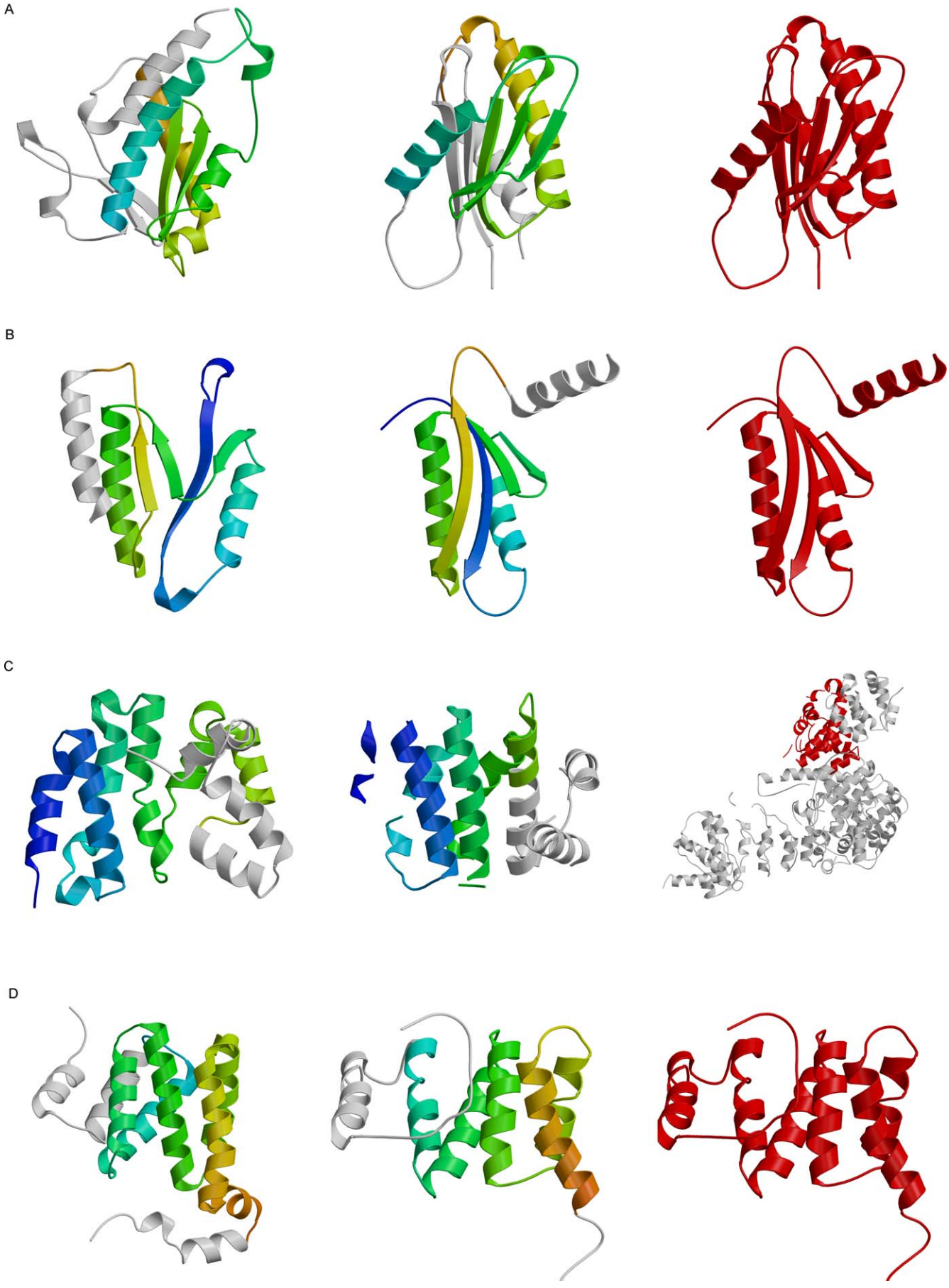
A
B
C
D

Figure 4. Predictions of Domains with Confident SCOP Superfamily Assignment Scores That Were Subsequently Solved

The predicted structure is shown in panel 1 (left), and the native structure or the structure of the homolog is shown in panel 2 (middle). The section matching the domain in the solved structure is colored red in panel 3 (right).

(A) TRS20/YBR254C is involved in the endoplasmic reticulum (ER) to Golgi transport and was predicted to belong to the SNARE-like (d.110.4) SCOP superfamily with a $P(SF|D,GO) = 0.98$ ($P_{MCM} = 0.69$). The *Z*-score between the predicted structure and the structure of a homolog with 33% sequence identity was 7.76.

(B) ECM15/YBL001C is involved in cell wall organization and biogenesis, and was predicted to belong to the MTH1187/YkoF-like (d.58.48) superfamily with $P(SF|D,GO) = 0.87$ ($P_{MCM} = 0.82$; MAMMOTH *Z*-score to native structure was 8.18).

(C) KAP104/YBR017C is involved in nuclear localization sequence binding and was predicted to belong to the ARM repeat (a.118.1) superfamily with $P(SF|D,GO) = 0.99$ ($P_{MCM} = 0.8$; MAMMOTH *Z*-score to homologous structure [31% sequence identity] was 4.41).

(D) FIS1/YIL065C was predicted to belong to the SCOP superfamily TPR-like (a.118.8) with a P_{MCM} of 0.98. The Mammoth *Z*-score between the predicted structure and the experimental structure was 13.34.

doi:10.1371/journal.pbio.0050076.g004

Twenty-seven domains from this project now have a homolog of known structure (see Materials and Methods for homology definition) and three of the 11 predictions with a P_{MCM} of 0.8 or higher are correct. Five of seven proteins from this set with a $P(SF|D,GO) \geq 0.8$ are correct; five of which (three correct) also have a P_{MCM} of 0.8. The small sample size represented by this dataset makes it difficult to assess accurate upper and lower bounds of the estimated error. We have also generated a much larger dataset (see Materials and Methods), as part of the Human Proteome Folding Project (HPF). This dataset, containing proteins from over 150 organisms, was derived without the use of fold recognition (and thus is not identical to the protocol for yeast), but provides valuable information as to the effect of domain prediction on our procedure. A total of 44% of the 207 predictions that were made for recently solved structures with P_{MCM} above 0.8 in this dataset were correct; 84% of the 51 predictions with $P(SF|D,GO)$ above 0.8 were correct; and 31 of these predictions had both a P_{MCM} and a $P(SF|D,GO)$ above 0.8, and 27 of these were correct. Over all three validation sets, more than 40% of the predictions with a P_{MCM} above 0.8, and more than 75% of the predictions with a $P(SF|D,GO)$ above 0.8, are correct, illustrating the value of data integration in this work.

Importantly, we were able to use these sets of recently solved proteins to better characterize the errors associated with different confidence Ginzu domain predictions. We found that a subset of the incorrect domain parses which significantly diminish the chances of correctly predicting fold and function are easily removed using a simple filter (described in Materials and Methods). This domain-prediction filter allows us to recover more-accurate predictions for multi-domain proteins. We were able to classify 50% of the amino acids from the 6,238 attempted ORFs to SCOP superfamilies which is significantly higher than the 35% coverage achieved by a sequence-based hidden Markov model approach [30].

Table 3. MAMMOTH Confidence Metric (MCM) Logistic Regression Model Parameters

Model	Mammoth *Z*-Score	CO	Convergence	Length Ratio	Intercept (C)
Alpha	0.66	0.13	0.09	4.08	4.53
Beta	0.66	0.09	0.35	6.72	1.60
Alpha/beta	0.67	0.05	0.03	5.16	4.10

doi:10.1371/journal.pbio.0050076.t003

Novel SCOP Superfamily Assignments

In this section, we discuss several protein complexes with components assigned to superfamilies by both GO-integration and MCM approaches. These predictions and the much larger set of predictions in the database accompanying this paper provide a basis for hypothesis generation and experimental testing, but it must be borne in mind that there is a significant probability that any single prediction is incorrect, as indicated by our estimates of error.

The mediator complex, a large complex containing 24 polypeptides [31], has been shown to be required for transcriptional activation in many eukaryotic organisms and play key roles in transmitting regulatory information to the pre-initiation complex. During transcriptional initiation, it interacts with the RNA polymerase II holoenzyme and the promoter region. The role of the mediator complex in transcriptional regulation, and the complete makeup of this complex and its dynamic composition throughout different cell and developmental states (in response to specific regulators) are active areas of research. To date, several studies have explored the overall makeup of the complex by probing protein–protein interactions [31] and by electron microscopy of purified mediator complex, but to our knowledge, this complex has eluded higher resolution methods such as crystallographic analysis. Although there exists an extensive body of work on the overall function of this complex, the roles, positions, and structures of most of the individual polypeptide components remain undetermined.

We find confident superfamily predictions for several proteins within this complex that were not structurally annotated prior to this work. Table 4 outlines these predictions, as well as their sources and confidence estimates. Several proteins in the Mediator head domain are predicted to contain DNA-binding domains. In addition, multiple head domain proteins are predicted to be long helical bundles, potentially serving as scaffolds. ROX3 contains two predicted domains, see Figure 5A. The first domain is predicted to belong to the Homeodomain-like superfamily ($P_{MCM} = 0.43$; GO-term: transcription from RNA polymerase II promoter; $P(SF|D,GO) = 0.83$); implying DNA binding. MED4 (Figure 5B), in the middle region of the mediator complex, is also a two-domain protein with a N-terminal homeodomain-like superfamily assignment ($P_{MCM} = 0.65$; GO-term: transcription from RNA polymerase II promoter; $P(SF|D,GO) = 0.98$). Several superfamily predictions for this complex (such as hits to superfamilies like spectrin-like and Rossmann folds) are difficult to interpret unambiguously due to the large number of functions compatible with each of these superfamilies, but are not incompatible with DNA-binding functions. Gal11 (Figure 5C) contains a diverse mix of predicted domain

Table 4. Predicted Domains and Structures for Components of the Mediator Complex

ORF	Domain Number	Type	Span	Length	Comment
ROX3/YBL093C	1	MSA	1–83	83	$P_{GI} = 0.830$ to Homeodomain-like (a.4.1)
SIN4/YNL236W	2	Fold recognition	458–641, 689–773	269	Matched PDB: 1qoyA
SIN4/YNL236W	3	Fold recognition	642–688	47	Matched PDB: 1qoyA
SRB6/YBR253W	1	Unassigned	1–121	121	$P_{MCM} = 0.838$ to t-snare proteins (a.47.2)
SSN3/YPL042C	1	PSI-BLAST	1–416	416	Matched PDB: 1opkA_
SSN3/YPL042C	2	PSI-BLAST	417–472	56	Matched PDB: 1b38A_
MED2/YDL005C	2	Fold recognition	133–206	74	Matched PDB: 1l8mA
MED2/YDL005C	3	Fold recognition	207–286	80	Matched PDB: 1l8mA
MED2/YDL005C	4	Fold recognition	287–356	70	Matched PDB: 1l8mA
MED2/YDL005C	5	Fold recognition	357–431	75	Matched PDB: 1l8mA
SSN8/YNL025C	1	PSI-BLAST	1–178, 280–323	222	Matched PDB: 1jkw__
SSN8/YNL025C	2	PSI-BLAST	179–279	101	Matched PDB: 1jkw__
SRB7/YDR308C	1	PSI-BLAST	1–140	140	Matched PDB: 1i84S_
MED4/YOR174W	1	Unassigned	1–64	64	$P_{GI} = 0.978$ to Homeodomain-like (a.4.1)
GAL11/YOL051W	1	PSI-BLAST	1–296	296	Matched PDB: 1i84S_
GAL11/YOL051W	2	Fold recognition	297–507	211	Matched PDB: 1m2vB
GAL11/YOL051W	5	Fold recognition	682–916	235	Matched PDB: 1k83A

doi:10.1371/journal.pbio.0050076.t004

structures: for the first domain, we find a PSI-BLAST match to the motor domain of myosin, and the second domain shows a strong fold recognition hit to a structural domain from sec24 (a component of the secretion system). Rosetta models for the third domain match the spectrin-like superfamily, and the models of the fourth domain match the DNA-binding lambda-repressor–like fold (with a competing hit to the tRNA-binding fold). The fifth domain shows a match to the RNA pol II–like fold (RPB1) by confident fold recognition. Although these domain and structure predictions are insufficient by themselves to localize specific molecular function to components of the mediator complex, it is encouraging that we can make some headway in localizing specific superfamilies and functions to components of this large complex.

TIF35 (Figure 5D) is a subunit of the translation initiation factor complex, a complex essential for translation [32]. We predict three separate domains exist within this protein. The middle and C-terminal domains of this protein are both strongly predicted to belong to the RNA-binding domain, RDB superfamily (d.58.7; identified by PSI-BLAST). The N-terminal 65 amino acids lack sequence-detectable homologs of known structure, but the Rosetta-generated models and GO-selected structure prediction for this protein shows a strong match to the Translation proteins SH3-like domain ($P_{MCM} = 0.44$; GO-term: translation initiation factor activity; $P(SF|D,GO) = 0.92$).

The mitochondrial ribosome, or the mitoribosome, shares a number of protein components with bacterial ribosomes, but it is believed that the mitoribosomes have comparatively more proteins than their bacterial counterparts; many of the proteins associated with the mitoribosome have no detectable sequence similarity to other mitochondrial proteins [33]. We have predicted the structure for two components known to be associated with the mitoribosome [34,35]. MRPL37 (Figure 5E) is predicted to belong to the Ribosomal protein L6 superfamily ($P_{MCM} = 0.31$; GO-term structural constituent of ribosome; $P(SF|D,GO) = 0.86$), a superfamily involved in RNA binding. We predict that MRPL44 (Figure 5F) belongs to the dsRNA-binding domain–like superfamily ($P_{MCM} = 0.78$; GO-term: structural constituent of ribosome; $P(SF|D,GO) = 0.86$). Overall, the structure predictions for these mitoribosome proteins suggest that they belong to superfamilies compatible with known, although highly diverged, components of both the bacterial and eukaryotic ribosome.

INH1 (Figure 5G) is an ATPase inhibitor predicted to be a member of the ARM repeat superfamily ($P_{MCM} = 0.73$; GO-term: ATP synthesis couple protein transport; $P(SF|D,GO) = 0.82$). Inh1 dimerizes and binds to the F_1 complex of the ATPase, thereby inhibiting its function [36,37]. Many members of the ARM repeat superfamily are involved in protein and peptide binding, which is consistent with both the dimerzation and the binding to the ATPase.

Data Access

All data are accessible via the Yeast Resource Center (YRC) public data repository [38] at http://rd.plos.org/10.1371__journal.pbio.0050076__01. The data will also be made available in other formats upon request.

Discussion

Comprehensive generation of three-dimensional structures with resolution or reliability of those determined by X-ray crystallography or nuclear magnetic resonance (NMR) is currently beyond the capabilities of any protein structure prediction method; these methods can, however, play an important role in generating structural annotations for whole genomes due to the much lower investment of resources required per protein domain. In this work, we have shown that it is possible to: (1) generate protein structure models on a genome-wide scale, (2) automate the assessment of the structure prediction quality, (3) convert the results into pre-existing encodings of structure in the form of SCOP superfamily classifications, and (4) augment the model-based assignment of SCOP superfamily by integrating with pre-existing function, process, and component information encoded in the GO database.

We were able to assign SCOP superfamilies to 7,094 of the

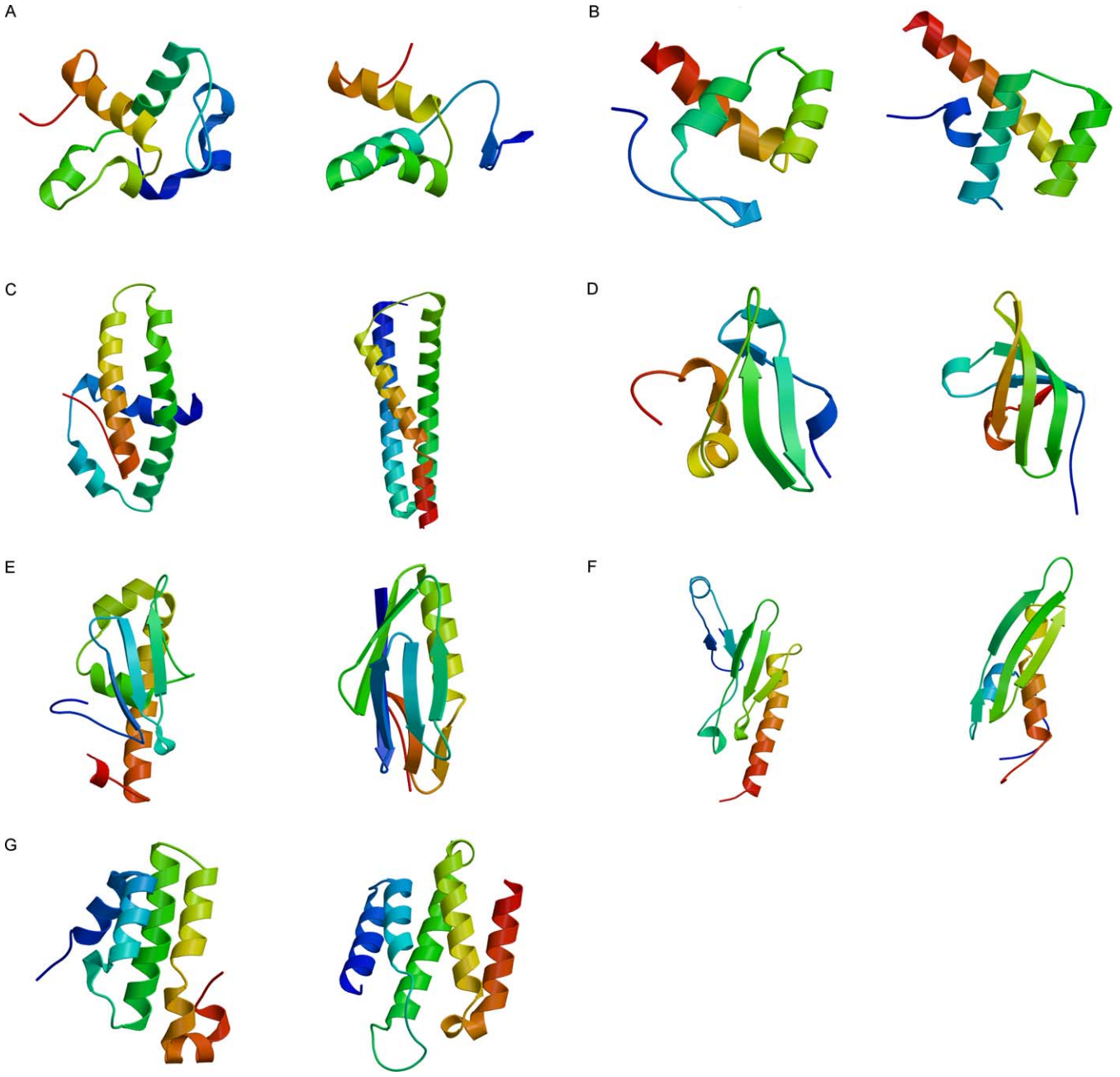

Figure 5. Structure Predictions with High *P*(SF|D,GO)

The predicted structure with the highest *P*(SF|D,GO) (left) is shown for (A) ROX3, (B) MED4, (C) GAL11, (D) TIF35, (E) MRPL37, (F) MRPL44, and (G) INH1 displayed next to the matching SCOP representative (right).

doi:10.1371/journal.pbio.0050076.g005

14,934 predicted domains in yeast using PSI-BLAST and fold recognition methodology. A total of 4,006 of the remaining 7,840 domains were short enough (less than 150 amino acids) for de novo structure prediction. Of these, 668 were omitted because they contained at least one predicted transmembrane helix. Low-resolution structure models were built for the remaining domains using Rosetta; of these, 404 were assigned to superfamilies with confidence using MCM, and an additional 177 were assigned with confidence after integrating with GO process, component, and function annotations.

A significant challenge in carrying out this work was the magnitude of the computation required for generating de novo structure predictions for large numbers of domains. Robust and fast methodology, efficient data storage, analysis tools, and data organization were required. Our use of distributed computing (http://wcgrid.org), innovative database architecture [39,40], and fully automatic methods were essential for this full-genome annotation. Yeast is particularly interesting because it is the focus of a vast global research effort. Future work will include an ongoing effort to scale this procedure to over 150 completely sequenced genomes as well as to employ recently developed higher resolution structure

Table 5. Summary of Benchmark Results

Protein Class	Subcategory	Total	Correct in Top 1	Correct in Top 5	Correct in Set	Correct SF Prediction in Top30 with $P_{MCM} > 0.8$
All		979	294	490	536	415
Alpha		323	149	214	233	158
Beta		233	16	47	53	48
Alpha-beta		423	129	229	250	209
1–100		449	135	228	251	205
101–120		203	67	105	116	96
121–150		209	59	100	108	70
151–200		118	33	57	61	44
1–100	Alpha	172	69	105	115	81
	Beta	105	11	27	30	40
	Alpha-beta	172	55	96	106	84
101–120	Alpha	60	36	43	46	38
	Beta	48	1	9	11	6
	Alpha-beta	95	30	53	59	52
121–150	Alpha	49	26	41	44	24
	Beta	60	3	9	10	2
	Alpha-beta	100	30	50	54	44
151–200	Alpha	42	18	25	28	15
	Beta	20	1	2	2	0
	Alpha-beta	56	14	30	31	29

The number of proteins for which the top cluster center or top five cluster centers had a MaxSub *e*-value of 10^{-6} or better to the native structure is reported in the columns "Correct in Top 1" and "Correct in Top 5," respectively. The number of proteins for which the native superfamily had a P_{MCM} of 0.8 or better is shown in the column to the right. The rows of the table give the statistics for domains in different secondary structure classes and length ranges.

doi:10.1371/journal.pbio.0050076.t005

prediction methods [41] that produce more-accurate and reliable models, but require significantly greater computational resources per protein domain.

The information content in the predicted structures may be further leveraged by integration with other data such as global quantitative measurements of mRNA, protein expression levels, DNA–protein, and protein–protein interactions. Such datasets are available for yeast and several other organisms as part of ongoing functional genomics efforts, and integration of these data types with the predicted structures should contribute to the annotation of protein functions.

Materials and Methods

Benchmark set; folding representatives from SCOP. Two representative domains from each SCOP [12–14] superfamily were folded using the Rosetta de novo method [1,2,28]. Superfamilies without members shorter than 200 amino acids were excluded, as were proteins for which Rosetta failed to produce predictions within a reasonable time. One thousand models were generated for each domain. This resulted in structure predictions for 998 domains for which the structures have been experimentally determined. The predicted structures were clustered by root mean square deviation (RMSD), and the centers of the top 30 clusters were compared to a domain database generated from ASTRAL 1.67 (reduced to 40% sequence identity) [42,43] using a modified version of MAMMOTH [21] that calculates the contact order of the aligned regions of the predicted structure and the ASTRAL domain. An overview of the statistics is presented in Table 5, and a detailed description of the results in Table S4.

The MAMMOTH Confidence Metric. The MCM estimates the probability that the MAMMOTH match between predicted structure and the ASTRAL domain (see previous section) has identified the correct superfamily and is based on the closeness of match (MAMMOTH *Z*-score), the length of the two proteins involved, L_{Astral} and $L_{\mathrm{predicted}}$, the CO of the region of the predicted structure that was superimposable on the experimental structure, and the degree to which Rosetta converged during the generation of the set of predicted conformations (converg below; estimated during the clustering step). The general formula for the confidence functions is given in Equation 2, and the weights of the parameters (*a, b, c, d,* and the constant C) for the three models described in the following paragraph are presented in Table 3.

$$\log\left(\frac{P_{\mathrm{MCM}}}{1-P_{\mathrm{MCM}}}\right) = a\cdot Z\text{-score} + b\cdot \mathrm{CO} + c\cdot \mathrm{converg} + d\cdot\left|\log\left(\frac{L_{\mathrm{Astral}}}{L_{\mathrm{predicted}}}\right)\right| + C \quad (2)$$

This model is similar to that used in previous studies [11], with two improvements. First, we have fit three separate logistic regression models, one for all alpha proteins, one for all beta proteins, and one for alpha and beta proteins; the size of the benchmark set and the fact that we are fitting a small number of parameters allows for this trifurcation of the benchmark set. Second, we compute the CO [29] over the matched region. This penalizes the scenario in which small numbers of long secondary structure elements (usually helices) are aligned; the CO term as well as the length ratio corrects for the overly confident score we would otherwise calculate based on convergence and MAMMOTH *Z*-score alone. We used 5-fold cross-validation to fit each of the three secondary structure class–specific confidence functions. For selecting between the three models for a query protein, we use secondary structure content predicted by PsiPred [44]. The alpha model is used for proteins with over 15% predicted alpha-helical content and under 15% beta-sheet content. The beta model is used for protein with more than 15% predicted beta strand and less than 15% alpha helical. The alpha/beta model was used for all other domains.

Estimating superfamily probabilities, given the structure predictions. Given a set of predicted structures D for a given protein, we estimate the probability the protein belongs to superfamily, SF, $P(\mathrm{SF}|\mathrm{D})$ as follows. Each superfamily is initially assigned a probability corresponding to the maximum P_{MCM} value for that superfamily over the top five P_{MCM} values for all predicted conformations for the query protein; probabilities less than 0.2 are set to zero. If the sum of the raw probabilities is greater than 0.8, they are scaled linearly so that the sum is 0.8. Because of the uncertainties of de novo structure prediction, these scaled probabilities, $P_{\mathrm{scaled}}(\mathrm{SF}|\mathrm{D})$, are then linearly combined with the background superfamily distribution, $P(\mathrm{SF})$ (Equation 3):

$$P(\mathrm{SF}|\vec{\mathrm{D}}) = P_{\mathrm{scaled}}(\mathrm{SF}|\vec{\mathrm{D}}) + \left(1 - \sum_{\mathrm{SF}} P_{\mathrm{scaled}}(\mathrm{SF}|\vec{\mathrm{D}})\right)\cdot P(\mathrm{SF}) \quad (3)$$

The final distributions, $P(\mathrm{SF}|\mathrm{D})$, are guaranteed to have non-zero probabilities for every superfamily, and to sum to 1. The background distribution $P(\mathrm{SF})$ ensures that (1) we do not disregard useful functional information at the integration with GO stage and (2) that we do not over interpret the confidence values derived from the benchmark training set.

Integration of function. We obtain $P(\mathrm{SF}|\mathrm{D},\mathrm{GO})$ of a superfamily, SF, given both protein structure prediction, D, and GO annotations, GO, using Bayes' rule and the assumption that $P(\mathrm{GO},\mathrm{D}|\mathrm{SF}) \sim P(\mathrm{GO}|\mathrm{SF})^{*}P(\mathrm{D}|\mathrm{SF})$:

$$P\left(\mathrm{SF}|\vec{\mathrm{D}},\mathrm{GO}\right) = \frac{P(\mathrm{SF})\cdot P(\mathrm{GO}|\ \mathrm{SF})\cdot P(\vec{\mathrm{D}}|\ \mathrm{SF})}{P(\mathrm{GO})\cdot P(\vec{\mathrm{D}})} \quad (4)$$

We obtain $P(\mathrm{D}|\mathrm{SF})$ via Equation 5:

$$P(\vec{\mathrm{D}}|\mathrm{SF}) = \frac{P(\mathrm{SF}|\vec{\mathrm{D}})P(\vec{\mathrm{D}})}{P(\mathrm{SF})} \quad (5)$$

After substituting Equation 5 into Equation 4, both $P(\mathrm{SF})$ and $P(\mathrm{D})$ cancel. $P(\mathrm{SF}|\mathrm{D})$ is computed as described in the previous section, and $P(\mathrm{GO}|\mathrm{SF})$, $P(\mathrm{GO})$, and $P(\mathrm{SF})$ are computed from proteins in the PDB that are annotated with GO function, component, or process and also classified in SCOP. To deal with cases in which there is a single function annotation for a given superfamily, we allow for the possibility that the uniqueness of this mapping is due to undersampling of superfamily space (as represented by the PDB) or function space (as represented by GO) by adding pseudo counts distributed according to the background superfamily distribution, $P_{\mathrm{astral95}}(\mathrm{SF})$, computed from ASTRAL 1.67 culled so that no sequences are more then 95% identical.

$$P(\mathrm{GO}|\mathrm{SF}) = \frac{N(\mathrm{GO},\mathrm{SF}) + M\cdot P_{\mathrm{astral95}}(\mathrm{SF})}{N(\mathrm{GO}) + M} \quad (6)$$

The parameter M (a regularization parameter controlling the relative contribution of our pseudo-counts) was estimated by carrying out function assignment given the superfamily over the benchmark set: we chose M to minimize the classification error estimated using 10-fold cross-validation. The overall procedure was relatively insensitive to the value of M ranging from one to ten with an optimal value of four. The $P(\mathrm{SF}|\mathrm{GO})$ are too diffuse for confident superfamily prediction from GO annotations alone, hence the integration with the structure prediction data is critical for accurate superfamily predictions.

Equation 6 relies on the assumption that the functional annotations are independent and mutually exclusive, which is not the case. (GO is a directed acyclic graph [DAG], with an implicit conditional dependence of lower nodes on parent nodes.) Nodes can have multiple parents, thus the probability of the child nodes of a more general term are not guaranteed to sum to the probability of the parent term. To circumvent this problem, we assigned the combined probability for each superfamily by taking the maximum probability for that superfamily given the predicted structures and all functions, i.e., Equation 7:

$$P(\mathrm{SF}|\vec{\mathrm{D}},\mathrm{GO}) = \max(P(\mathrm{SF}|\vec{\mathrm{D}},\mathrm{GO}_1),P(\mathrm{SF}|\vec{\mathrm{D}},\mathrm{GO}_2),\ldots,P(\mathrm{SF}|\vec{\mathrm{D}},\mathrm{GO}_N)) \quad (7)$$

Finally, the sum of $P(\mathrm{SF}|\mathrm{D},\mathrm{GO})$ for any given protein domain is normalized to sum to one; thus confident assignments are not made when there are strong matches to more than one superfamily.

Datasets for evaluation. The performance of the MCM and the GO integration was evaluated on two independent datasets. The first dataset, from HPF project, consists of 768 predicted domains that now have a homolog with a known structure that is classified in SCOP 1.69. The homologs were identified by blasting predicted domains against all sequences from ASTRAL 1.69 and selecting those with a PSI-BLAST *e*-value less than 1×10^{-3}. We also require that the shorter of the two sequences is more than 80% of the length of the longer one, and that 60% or more of the predicted domain is aligned with the ASTRAL domain. These domains are part of an ongoing project in which we predict structures for over 150 genomes; although domains with any homology to known structures are excluded, a number of structures have been solved and classified in SCOP during the 18 mo the project has been running. The scope of this separate project prohibited us from carrying out fold recognition calculations on these domains, and since domains that can be assigned using fold recognition methods will on average have higher MAMMOTH structural similarities to known structures than domains that cannot be assigned, results from this dataset represent an upper bound on performance on the dataset in this paper.

The second dataset was generated the same way the HPF set was generated, but limited to yeast domains. The proteins from which these domains are derived have been subjected to fold recognition and hence give a better estimate of the true performance. This dataset is, however, too small for statistically significant conclusions to be made.

Domain filter. Based on inspection of the results on the HPF dataset, domains from predicted two-domain proteins are excluded if both the domains are predicted using less-confident methods (MSA, unassigned, or Pfam domains), or if the domain under consideration is an MSA domain regardless of the neighboring domain type. A large fraction of these proteins have single domains, and correct superfamily matches are quite unlikely when models are only generated from domain fragments.

Data production. The generation of structure predictions was divided into three completely automated steps: pre-processing, production (the running of Rosetta), and post-processing (clustering, superfamily assignment, and function integration). The pre-processing protocol includes domain prediction, prediction of secondary structure, disordered regions, trans-membrane helices [45], and signal peptides [46], and the local structure fragments and other files necessary for running Rosetta. This step was conducted in-house on two 64-CPU Linux clusters. The production step, generating 10,000 structure predictions, was completed in collaboration with IBM running Rosetta on the World Community Grid as part of a larger effort, and is estimated to have used 12 million CPU hours, or 1,350 CPU years. The post-processing step was performed in-house (using the same hardware as the pre-processing step), and included clustering and superfamily assignment by MCM and GO integration. The resulting dataset is complex, and is stored, queried, organized, and analyzed using an open-source software package, 2DDB [39,40] of our own construction.

Supporting Information

Table S1. Complete Listing of Domain Predictions for All ORFs in Yeast

All 14,934 domains predicted from the 6,238 sequences are presented in detail.

Found at doi:10.1371/journal.pbio.0050076.st001 (4.7 MB PDF).

Table S2. Protein Structure Predictions with $P_{\mathrm{MCM}} \geq 0.8$

The most confident predictions using the MCM are listed.

Found at doi:10.1371/journal.pbio.0050076.st002 (133 KB PDF).

Table S3. Protein Structure Prediction with $P(\mathrm{SF}|\mathrm{D},\mathrm{GO}) \geq 0.8$

The most confident predictions using the GO-integration strategy are listed.

Found at doi:10.1371/journal.pbio.0050076.st003 (77 KB PDF).

Table S4. Benchmark Results

Best predicted structure—the best predicted structure by RMS among the 1,000 created; Top5 is the cluster center from the five largest clusters; Best Match—the best domain match from all 30 cluster centers.

Found at doi:10.1371/journal.pbio.0050076.st004 (476 KB PDF).

Acknowledgments

We thank IBM (Viktors Berstis, Bill Bovermann, Rick Alther, and Robin Willner) for dedicated access to the World Community Grid (http://www.wcgrid.org) and for porting Rosetta to the grid-client. We also thank Phil Bradley and Bill Noble for helpful discussions.

Author contributions. LM, TND, RB, and DB conceived and designed the experiments. LM performed the experiments. LM, RB, and DB analyzed the data. LM, MR, CEMS, and DC contributed reagents/materials/analysis tools. LM, RB, and DB wrote the paper.

Funding. This work was funded by the National Center for Research Resources of the National Institutes of Health by a grant to TND entitled "Comprehensive Biology: Exploiting the Yeast Genome," P41 RR11823, the Howard Hughes Medical Institute, and the U.S. Department of Defense USAMRAA W81XWH-04-1-0307.

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

References

1. Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 268: 209–225.
2. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, et al. (1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 34: 82–95.
3. Bonneau R, Strauss CE, Baker D (2001) Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. Proteins 43: 1–11.
4. Rohl CA, Strauss CE, Misura KM, Baker D (2004) Protein structure prediction using Rosetta. Methods Enzymol 383: 66–93.
5. Bradley P, Chivian D, Meiler J, Misura KM, Rohl CA, et al. (2003) Rosetta predictions in CASP5: Successes, failures, and prospects for complete automation. Proteins 53: 457–468.
6. Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, et al. (2005) Free modeling with Rosetta in CASP6. Proteins 61: 128–134.
7. Chivian D, Kim DE, Malmstrom L, Bradley P, Robertson T, et al. (2003) Automated prediction of CASP-5 structures using the Robetta server. Proteins 53: 524–533.
8. Lesk AM, Lo Conte L, Hubbard TJ (2001) Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, and interresidue contacts. Proteins (Suppl 5): 98–118.
9. Hazbun TR, Malmstrom L, Anderson S, Graczyk BJ, Fox B, et al. (2003) Assigning function to yeast proteins by integration of technologies. Mol Cell 12: 1353–1365.
10. Bonneau R, Baliga NS, Deutsch EW, Shannon P, Hood L (2004) Comprehensive de novo structure prediction in a systems-biology context for the archaea *Halobacterium* sp. NRC-1. Genome Biol 5: R52.
11. Bonneau R, Strauss CE, Rohl CA, Chivian D, Bradley P, et al. (2002) De novo prediction of three-dimensional structures for major protein families. J Mol Biol 322: 65–78.
12. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2002) SCOP database in 2002: Refinements accommodate structural genomics. Nucleic Acids Res 30: 264–267.
13. Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, et al. (2000) SCOP: A structural classification of proteins database. Nucleic Acids Res 28: 257–259.
14. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536–540.
15. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, et al. (1997) CATH—A hierarchic classification of protein domain structures. Structure 5: 1093–1108.
16. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27: 29–34.
17. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.
18. Hegyi H, Gerstein M (1999) The relationship between protein structure and function: A comprehensive survey with application to the yeast genome. J Mol Biol 288: 147–164.
19. Kim DE, Chivian D, Malmstrom L, Baker D (2005) Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. Proteins 61: 193–200.
20. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
21. Ortiz AR, Strauss CE, Olmea O (2002) MAMMOTH (matching molecular models obtained from theory): An automated method for model comparison. Protein Sci 11: 2606–2621.
22. Camon E, Magrane M, Barrell D, Binns D, Fleischmann W, et al. (2003) The Gene Ontology Annotation (GOA) project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro. Genome Res 13: 662–672.
23. Hirschman JE, Balakrishnan R, Christie KR, Costanzo MC, Dwight SS, et al. (2006) Genome Snapshot: A new resource at the *Saccharomyces* Genome Database (SGD) presenting an overview of the *Saccharomyces cerevisiae* genome. Nucleic Acids Res 34: D442–445.
24. Ginalski K, Elofsson A, Fischer D, Rychlewski L (2003) 3D-Jury: A simple approach to improve protein structure predictions. Bioinformatics 19: 1015–1018.
25. Ginalski K, Rychlewski L (2003) Detection of reliable and unexpected protein fold predictions using 3D-Jury. Nucleic Acids Res 31: 3291–3292.
26. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, et al. (2004) The Pfam protein families database. Nucleic Acids Res 32: D138–D141.
27. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, et al. (2002) The Pfam protein families database. Nucleic Acids Res 30: 276–280.
28. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, et al. (2001) Rosetta in CASP4: progress in ab initio protein structure prediction. Proteins (Suppl 5): 119–126.
29. Bonneau R, Ruczinski I, Tsai J, Baker D (2002) Contact order and ab initio protein structure prediction. Protein Sci 11: 1937–1944.
30. Gough J, Karplus K, Hughey R, Chothia C (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 313: 903–919.
31. Guglielmi B, van Berkum NL, Klapholz B, Bijma T, Boube M, et al. (2004) A high resolution protein interaction map of the yeast Mediator complex. Nucleic Acids Res 32: 5379–5391.
32. Phan L, Zhang X, Asano K, Anderson J, Vornlocher HP, et al. (1998) Identification of a translation initiation factor 3 (eIF3) core complex, conserved in yeast and mammals, that interacts with eIF5. Mol Cell Biol 18: 4935–4946.
33. Graack HR, Wittmann-Liebold B (1998) Mitochondrial ribosomal proteins (MRPs) of yeast. Biochem J 329: 433–448.
34. Kitakawa M, Graack HR, Grohmann L, Goldschmidt-Reisin S, Herfurth E, et al. (1997) Identification and characterization of the genes for mitochondrial ribosomal proteins of *Saccharomyces cerevisiae*. Eur J Biochem 245: 449–456.
35. Grohmann L, Graack HR, Kruft V, Choli T, Goldschmidt-Reisin S, et al. (1991) Extended N-terminal sequencing of proteins of the large ribosomal subunit from yeast mitochondria. FEBS Lett 284: 51–56.
36. Dienhart M, Pfeiffer K, Schagger H, Stuart RA (2002) Formation of the yeast F1F0-ATP synthase dimeric complex does not require the ATPase inhibitor protein, Inh1. J Biol Chem 277: 39289–39295.
37. Devenish RJ, Prescott M, Roucou X, Nagley P (2000) Insights into ATP synthase assembly and function through the molecular genetic manipulation of subunits of the yeast mitochondrial enzyme complex. Biochim Biophys Acta 1458: 428–442.
38. Riffle M, Malmstrom L, Davis TN (2005) The Yeast Resource Center Public Data Repository. Nucleic Acids Res 33: D378–D382.
39. Malmstrom L, Malmstrom J, Marko-Varga G, Westergren-Thorsson G (2002) Proteomic 2DE database for spot selection, automated annotation, and data analysis. J Proteome Res 1: 135–138.
40. Malmstrom L, Marko-Varga G, Westergren-Thosson G, Laurell T, Malmstrom J (2006) 2DDB—A bioinformatics solution for analysis of quantitative proteomics data. BMC Bioinformatics 7: 158.
41. Bradley P, Misura KM, Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. Science 309: 1868–1871.
42. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, et al. (2004) The ASTRAL Compendium in 2004. Nucleic Acids Res 32: D189–D192.
43. Brenner SE, Koehl P, Levitt M (2000) The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res 28: 254–256.
44. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292: 195–202.
45. Sonnhammer EL, von Heijne G, Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol 6: 175–182.
46. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340: 783–795.