Code Performance & Resource Costs

Structural modeling is performed using I-TASSER. The I-TASSER predictions themselves consist of four parts: LOMETS (template identification from multiple threading programs), cas (structure assembly simulation using replica exchange Monte Carlo), SPICKER (decoy conformation clustering) and FG-MD (full-atomic refinement). Over long experience, I-TASSER modeling has shown roughly linear increases in required computational time with protein size (at least over the typical range of protein sizes encountered in the organisms under consideration) across a wide range of computing platforms. Below we show calculations arising from benchmarks on SDSC Comet that allow us to estimate the required computational resources for the calculations described in our Research Objectives.

LOMETS is a meta-program that combines threading result of 15 different sub-programs, where 13 of the sub-programs are relatively fast, and taking 30 minutes to 1 hour for each one to complete. Two of the sub-programs, RRR6 and RAP2, are slow, requiring 15 GB of memory and 3 to 6 hours to complete. Therefore, the two slow sub-programs will be submitted as two independent jobs, while the other 13 sub-programs as well as the PSI-BLAST and PSSpred programs will run within one master job sequentially. LOMETS requires about $1.28x$ SUs, where x is the size of the protein in amino acid residues. A break-down of the run-times of three runs for template identification using LOMETS are given in Table 1. 5 cas jobs that correspond to REMC simulations will be run for each I-TASSER job. For each target, this step will require $(7.5 + 0.25x) \times 5$ SUs. Each simulation job will output approximately 100 MB of decoy structure files. SPICKER consumes $0.004x$ SUs and FG-MD refinement only takes a few seconds to run (and is thus considered negligible). Combining the SUs required across the steps enumerated above, for one protein target, in total, running all four steps in I-TASSER requires $1.28x + (7.5 + 0.25x) \times 5 + 0.004x = 37.5 + 2.534x$ Comet SUs.

Table 1: LOMETS run-time (in minutes) for different length of proteins, including a breakdown of required time for all run components.

Program (full name)	Target (length in amino acids)		
	S303739 (281)	S302297 (575)	S303731 (1177)
PSI-BLAST and PSSpred	4	15	19
SPX (SPARKS-X)	23	34	60
RRR6 (wdPPAS)	38	46	74
FF3 (FFAS3D)	6	12	30
MUS (MUSTER)	37	54	68
IIIj (HHsearch-1)	16	31	41
JJJb (Neff-PPAS)	34	58	66
RAP2 (RaptorX)	189	300	393
IIIe (HHsearch)	61	180	139
VVV (SP3)	17	32	40
RRR3 (FFAS)	5	9	12
WWW (cdPPAS)	33	47	49
pgen (pGenThreader)	28	61	104
BBB (PROSPECT2)	25	62	74
PRC (PRC)	28	61	109
HHP (HHsearch-2)	13	38	47
TOTAL	**557**	**1040**	**1315**

We have also performed empirical benchmarks of COFACTOR performance on Comet across a wide range of system sizes, and observed that it likewise shows linear scaling with protein size across the range of proteins pertinent to our study (Figure 1). Based on these benchmarks, we estimate the COFACTOR time requirements to be $0.03x$ SUs for a protein x amino acids long. Thus, the

combined costs for a protein of x amino acids, for both I-TASSER and COFACTOR together, is given by $37.5 + 2.564x$ Comet SUs.

The I-TASSER and COFACTOR benchmarks described above are used to estimate the time requirements provided in Section 3 of the accompanying research plan. Two additional notes are warranted: First, because all of the steps here can be trivially parallelized, requiring no communication between the $> 10,000$ distinct protein targets to be considered, we do not consider scaling with job size to be an issue – all proteins can be submitted independently as separate jobs. Second, while all time is requested on the SDSC Comet system, our code (written almost entirely in Fortran 95 and C++) should likewise be portable to, and run without issues on, the Stampede2 Skylake nodes, particularly given the absence of any communication requirements (which would normally require substantial porting efforts to optimize and benchmark). Thus, an equivalent amount of computational resources on Stampede2 could be used equally well.

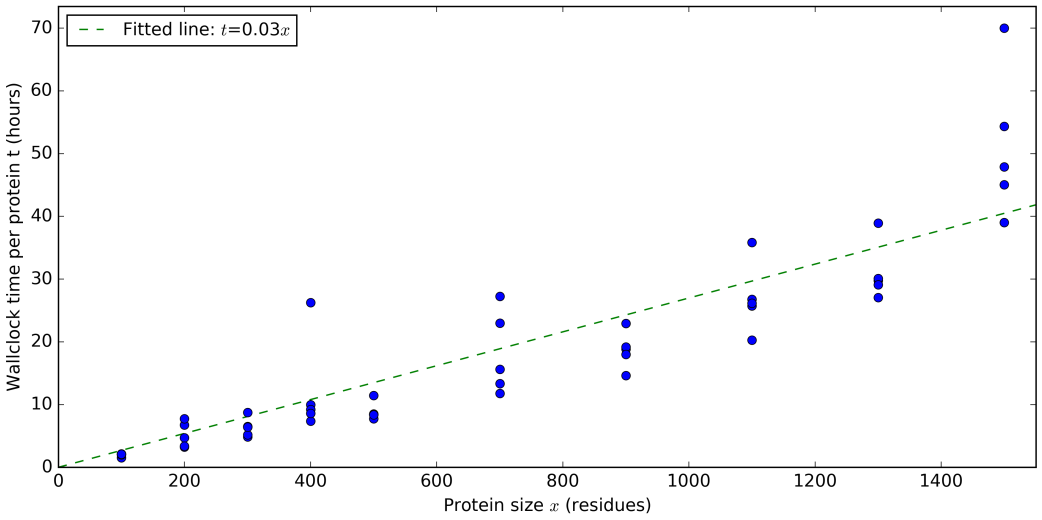

Figure 1: Runtime benchmark of COFACTOR on 50 proteins proteins performed on Comet. Five proteins are selected for each of the following ten protein lengths: 100, 200, 300, 400, 500, 700, 900, 1100, 1300, and 1500 residues. The running time (x-axis) of a protein linearly increases with the protein length (y-axis), with a Pearson Correlation Coefficient 0.96.

References

[1] UniProt Consortium, et al. (2014) Activities at the universal protein resource (uniprot). Nucleic acids research 42:D191–D198.

[2] Human Microbiome Project Consortium and others (2012) Structure, function and diversity of the healthy human microbiome. Nature 486:207–214.

[3] Friedberg I (2006) Automated protein function prediction the genomic challenge. Briefings in bioinformatics 7:225–242.

[4] Tian W, Skolnick J (2003) How well is enzyme function conserved as a function of pairwise sequence identity? Journal of molecular biology 333:863–882.

[5] Rost B (2002) Enzyme function less conserved than anticipated. Journal of molecular biology 318:595–608.

[6] Schnoes AM, Brown SD, Dodevski I, Babbitt PC (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 5:e1000605.

[7] Schnoes AM, Ream DC, Thorman AW, Babbitt PC, Friedberg I (2013) Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. PLoS computational biology 9:e1003063.

[8] Lee D, Redfern O, Orengo C (2007) Predicting protein function from sequence and structure. Nature Reviews Molecular Cell Biology 8:995–1005.

[9] Skolnick J, Brylinski M (2009) Findsite: a combined evolution/structure-based approach to protein function prediction. Briefings in bioinformatics :bbp017.

[10] Schmidt T, Haas J, Cassarino TG, Schwede T (2011) Assessment of ligand-binding residue predictions in casp9. Proteins: Structure, Function, and Bioinformatics 79:126–136.

[11] Oh M, Joo K, Lee J (2009) Protein-binding site prediction based on three-dimensional protein modeling. Proteins: Structure, Function, and Bioinformatics 77:152–156.

[12] Lee S, Hinz A, Bauerle E, Angermeyer A, Juhaszova K, et al. (2009) Targeting a bacterial stress response to enhance antibiotic action. Proc Natl Acad Sci U S A 106:14570–14575.

[13] Zhang C, Freddolino PL, Zhang Y (2017) Cofactor: improved protein function prediction by combining structure, sequence and protein–protein interaction information. Nucleic Acids Research .

[14] Zhang C, Zheng W, Freddolino PL, Zhang Y (2018) Metago: Predicting gene ontology of non-homologous proteins through low-resolution protein structure prediction and protein-protein network mapping. Computation Resources for Molecular Biology 430:2256–2265.

[15] Jiang Y, Oron TR, Clark WT, Bankapur AR, D Andrea D, et al. (2016) An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome biology 17:184.

[16] Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, et al. (2013) A large-scale evaluation of computational protein function prediction. Nature Methods 10:221.

[17] Gong Q, Ning W, Tian W (2016) Gofdr: a sequence alignment based method for predicting protein functions. Methods 93:3–14.

[18] Skotnicov P, Sobotka R, Shepherd M, H jek J, Hrouzek P, et al. (2018) The cyanobacterial protoporphyrinogen oxidase hemj is a new b-type heme protein functionally coupled with coproporphyrinogen iii oxidase. Journal of Biological Chemistry 293:12394–12404.

[19] Palevsky N, Shemer B, Connolly JP, Belkin S (2016) The highly conserved escherichia coli transcription factor yhaj regulates aromatic compound degradation. Frontiers in microbiology 7.

[20] Keseler IM, Bonavides-Martínez C, Collado-Vides J, Gama-Castro S, Gunsalus RP, et al. (2009) Ecocyc: a comprehensive view of *Escherichia coli* biology. Nucleic Acids Res 37:D464–D470.

[21] van Dijl J, Hecker M (2013). Bacillus subtilis: from soil bacterium to super-secreting cell factory.

[22] Collier J, Shapiro L (2007) Spatial complexity and control of a bacterial cell cycle. Protein technologies / Systems biology 18:333–340.

[23] Le TBK, Imakaev MV, Mirny LA, Laub MT (2013) High-resolution mapping of the spatial organization of a bacterial chromosome. Science (New York, NY) 342:731–734.

[24] Bibb M (2013) Understanding and manipulating antibiotic production in actinomycetes. Biochem Soc Trans 41:1355.

[25] Roy A, Kucukural A, Zhang Y (2010) I-tasser: a unified platform for automated protein structure and function prediction. Nature protocols 5:725–738.

[26] Stuhmer W, Conti F, Suzuki H, Wang XD, Noda M, et al. (1989) Structural parts involved in activation and inactivation of the sodium channel. Nature 339:597–603.

[27] Folander K, Douglass J, Swanson R (1994) Confirmation of the assignment of the gene encoding kv1.3, a voltage-gated potassium channel (kcna3) to the proximal short arm of human chromosome 1. Genomics 23:295–6.

[28] DeCoursey TE, Chandy KG, Gupta S, Cahalan MD (1984) Voltage-gated k+ channels in human t lymphocytes: a role in mitogenesis? Nature 307:465–8.

[29] Cordero-Morales JF, Jogini V, Chakrapani S, Perozo E (2011) A multipoint hydrogen-bond network underlying kcsa c-type inactivation. Biophys J 100:2387–93.

[30] Herrou J, Crosson S (2011) Function, structure and mechanism of bacterial photosensory lov proteins. Nature reviews

microbiology 9:713–723.
[31] Möglich A, Moffat K (2010) Engineered photoreceptors as novel optogenetic tools. Photochemical & Photobiological Sciences 9:1286–1300.
[32] Pudasaini A, Kaley K, Zoltowski BD (2015) Lov-based optogenetic devices: light-driven modules to impart photoregulated control of cellular signaling. Frontiers in Molecular Biosciences 2.

High-throughput combined structure/function prediction for microbial proteomes

1 Scientific background

Modern progress in genome sequencing has lead to an explosion in the number of known protein sequences; currently 55 million such sequences are available in the Uniprot database [1], but fewer than 500,000 of these proteins have been experimentally annotated. Understanding the diverse biology of organisms ranging from humans to environmental microbes will require us to effectively utilize our now massive knowledge in sequence space and translate it to knowledge regarding protein structure and function. This transition requires, at a coarse level, the ability to predict and annotate protein functions computationally (as it is not practically feasible to do so experimentally), and at a finer level, the ability to combine appropriate experiments with biophysical simulations to determine the detailed mechanisms and properties of those proteins. The sequence-function gap poses a particularly pressing problem in the field of microbiology, due to the sheer number of different species to be considered; the human microbiome for a single healthy volunteer, for example, may contain $> 10^4$ operational taxonomic units [2]. The diversity of proteins at work in such a system is completely beyond the reach of experimental characterization, either now or in the near future, and thus the bulk of information that we have on non-model organisms will remain limited to computational inferences. While annotations based on sequence homology have long proved useful, they are fundamentally limited by the need for high homology and functional divergence of proteins with similar sequences (reviewed by Friedberg [3]), leading to high mis-annotation rates in public databases such as UniProtKB [4–7].

Molecular modeling has already proven to be a crucial tool in both coarse-grain functional annotation and fine-grain mechanistic investigations. For the purposes of annotation, most current annotation predictions are based purely on sequence homology [8], which has been shown to become unreliable for lower homology structures [4, 5, 7]. Because structure-function relationships persist and correlate with functional information even for proteins of very low sequence homology (*e.g.*, [9]), it is now becoming possible to perform purely structure based annotation predictions [10], even when using predicted rather than experimentally obtained protein structures [11]. For more detailed investigations, molecular dynamics simulations have emerged as a "computational microscope" [12] providing insight into experimentally inaccessible aspects of biophysical processes. The ongoing development and application of molecular modeling for both high-throughput and high-detail investigation of biomolecular function is crucial to expanding our ability to understand and manipulate biomolecular systems.

Working in collaboration with the laboratory of Prof. Yang Zhang (U. Michigan), we have recently developed a high-performance workflow for unified prediction of the structure and function of poorly annotated proteins, COFACTOR/MetaGO [13, 14]. The COFACTOR/MetaGO workflow is schematized in Fig. 1, and involves prediction of a three-dimensional structure for the protein of interest using either homology of *ab initio* modeling, followed by consensus annotations using information from sequence, structure, and protein-protein interaction databases to provide predicted annotations for the protein of interest. As described in detail below (Section 2.1), large-scale benchmarks demonstrate that the hybrid annotation approach used in COFACTOR/MetaGO provides very strong performance relative to purely sequence-based methods, especially for protein targets without any close reliably-annotated homologs. In addition, our pipeline provides predictions regarding potential ligands/substrates, ligand binding sites, and structural homologs that assist in interpretation and testing of the predictions.

In pilot applications, we have applied our new methods to perform whole proteome structure/function predictions for the important model organism *Escherichia coli* K12. In addition to providing high quality annotations that will be of use to any researchers encountering those genes in their work (many of which we have already experimentally validated), global analysis of our newly completed whole-proteome annotation revealed several classes of genes that appear systematically under-annotating in

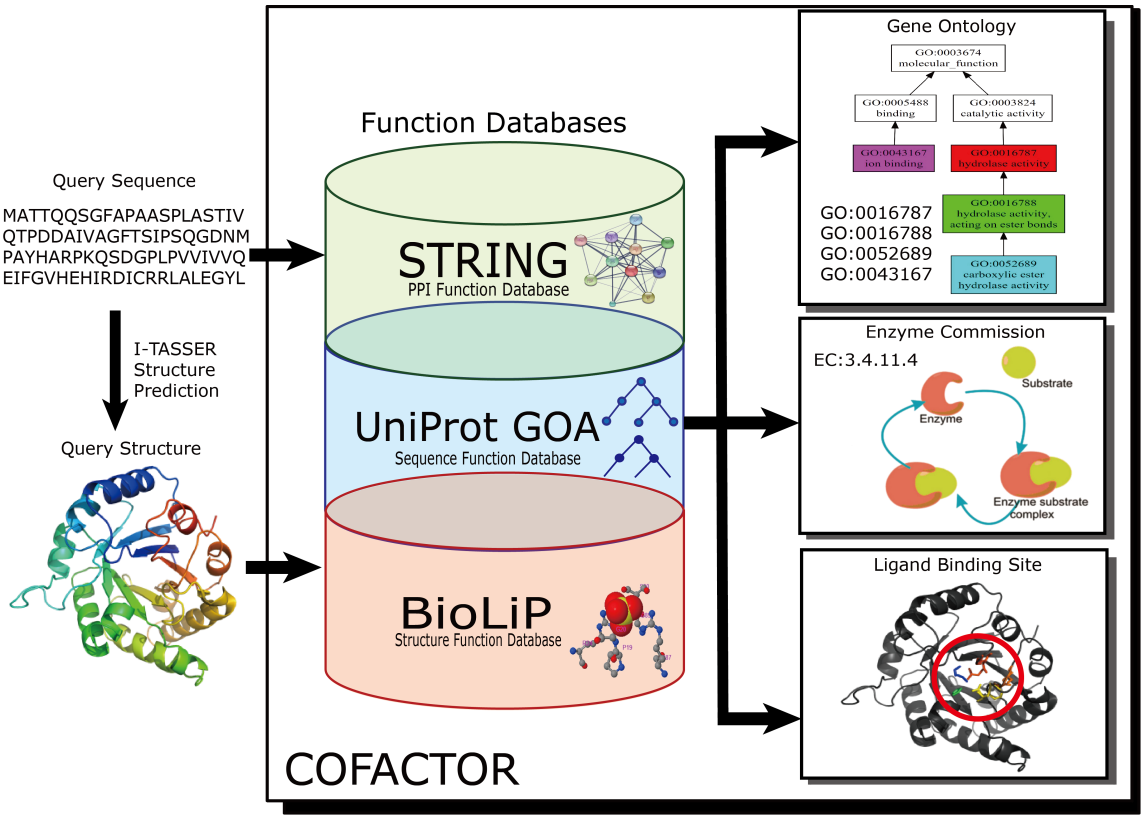

Figure 1: Flowchart for integrated prediction of protein structure, Gene Ontology, Enzyme Commission, and ligand binding.

existing data sets (detailed in Section 2.2). Here we propose the application of similar massive-scale structure/function predictions to the poorly annotated proteins of three additional widely-used model microbes: *Bacillus subtilis*, *Caulobacter crescentus*, and *Streptomyces coelicolor*. Taken together completion of these annotations will dramatically expand our understanding of the biology of the target organisms, providing a massive resource for researchers in the field to interpret any findings (e.g., from genetic screens) that implicate poorly annotated proteins, and yielding a new pool of protein functions which will likely reveal industrially useful enzymes. Our ongoing work on these topics is supported by NIH grants R35GM128637 and R01AI13467801.

2 Research Objectives

Our objectives are to obtain structure/function annotations for all poorly annotated proteins in the proteomes of the model organisms *Bacillus subtilis*, *Caulobacter crescentus*, and *Streptomyces coelicolor*, following the workflow shown in Figure 1. Here we outline the preliminary data that we have obtained showing the efficacy of our methods (Section 2.1) and outline our findings from a whole proteome structure/function prediction for *E. coli* K12 (Section 2.2), followed by a description of our proposed work in Section 2.3.

2.1 Preliminary data: Performance of COFACTOR on annotation tasks

To provide a useful assessment of the performance of COFACTOR/MetaGO in annotating previously cryptic proteins, we performed a large-scale benchmark simulation in which we attempted to re-annotate ∼1,200 *E. coli* proteins of known function, while masking all possible annotation templates within a progressively larger radius from our database (thus simulating the case of truly new targets with progressively poorer templates available for annotation transfer). As shown in Figure 2A, our hybrid pipeline shows stronger performance than both standard methods (BLAST/PSI-BLAST) and a top performer in the CAFA2 competition [15]. The advantages of COFACTOR become particularly striking as the homology of available templates decreases, driven by the robustness of the structure and PPI pipeline components. The detailed predictions provided by our pipeline for ten targets from this testing set are shown in Table 1, demonstrating that the predictions show an extremely high success rate even for proteins with no close homologs of known structure or function. Further details are provided in [13, 14].

COFACTOR (prior to the MetaGO enhancements) was also used in the community-wide CAFA3 competition (`https://biofunctionprediction.org/cafa/`) as the "ZhangFreddolinoLab" method, with official results expected to be released publicly late in 2018. As official rankings for each method have been passed along to that method's creators, however, we can state that COFACTOR was the top ranked method in several categories, including overall prediction of Biological Process GO terms across all organisms for several scoring methods, as well as top rankings for several other organism/GO aspect combinations. As the sheer variety of different scoring methods used by CAFA (as in [15, 16]) prevents simple quotation of a single overall performance metric, we instead show the distribution of COFACTOR's percentile rankings across all target subdivisions and scoring metrics in Fig. 2B, demonstrating that it was a consistently high performer across targets and evaluations.

2.2 Preliminary data: Experimental validation of COFACTOR/MetaGO predictions

As an initial large-scale application of COFACTOR/MetaGO, we recently completed whole-proteome structure/function predictions for the entire *E. coli* K12 protome [Rahimpour, Zhang, Zhang, and Freddolino; in preparation]. We were thus able to obtain a global perspective on entire classes of proteins

Table 1: Results of preliminary testing of COFACTOR annotations for genes with previously known functions. [a]: sequence identity to the closest known structure; [b]: validation of ligand prediction (Y-correct; N-not correct); [c]: validation of function prediction (Y-correct; N-not correct); [d]: technically correct, but misses the crucial ppGpp binding site)

Gene name	Annotation	Top COFACTOR ligands	Top COFACTOR GO terms	ID[a]	COL[b]	COF[c]
Proteins with known structures (or structures of a homolog from a different organism)						
b0945/*pyrD*	Dihydroorotate dehydrogenase	Flavin mononucleotide, 5-iodoorotate	Dihydroorotate oxidase activity, FMN binding,	1.00	Y	Y
b0523/*purE*	N5-CAIR ribonucleotide mutase	Nitro-AIR (substrate analog)	AIR carboxylase activity, N5-CAIR mutase activity	0.99	Y	Y
b3860/*dsbA*	Protein disulfide oxidoreductase	Peptides	Protein disulfide oxidoreductase activity	1.00	Y	Y
b0145/*dksA*	ppGpp-binding, RNA polymerase- binding TF	Zn2+, bacterioruberin	Zinc ion binding	1.00	Y/N[d]	Y/N[d]
b1530/*marR*	Antibiotic-binding transcription factor	Nucleic acid oligomers, ferulic acid (crystallographic ligand)	DNA binding, transcription factor activity	1.00	Y	Y
b0464/*acrR*	Transcription factor binding polycyclic aromatic compounds	Triclosan, chloramphenicol, rhodamine 6G, nucleic acid oligomer	Protein binding, DNA binding, transcription factor activity	1.00	Y	Y
Proteins with no structures of close homologs						
b2494/*bepA*	Metalloprotease, chaperone	Peptides, zinc	Exopeptidase, protein binding	0.11	Y	Y
b3632/*waaQ*	LPS core heptosyltransferase	ADP-heptose, a nucleotide diphosphate-sugar, UDP-glucose	Nucleotide binding, hexosyl transferase activity, LPS heptosyltransferase activity	0.26	Y	Y
b1049/*opgH*	Synthesis of periplasmic glucans	ß-D-glucose, Mn^{2+}, UDP, Mg^{2+}	Ion binding, metal cluster binding	0.14	Y	Y
b3828/*metR*	Homocysteine-binding transcription factor	Muconic acid, benzoic acid, thiocyanate	DNA binding, transcription factor activity	0.24	N	Y

that are systematically over- and under-represented among the poorly annotated set (defined henceforth as the set of proteins with a UniProt Annotation Score of 1 or 2, out of 5). As seen for Biological Process GO terms in Fig. 3, we are able to identify GO terms belonging to both categories. We find that the poorly annotated gene set tends to be depleted for genes involved in core metabolic processes (*e.g.*, ribosomal proteins, carbon and nitrogen metabolism, transporters), but that several other types of proteins are significantly enriched, including prophage components, transposases, pathogenesis proteins, and enzyme activators [Rahimpour, Zhang, Zhang, and Freddolino; in preparation].

The detailed structural and functional predictions provided by our workflow allowed us to design compact and efficient experiments to test several of the new annotations. For example, we predicted (and subsequently confirmed) that the cryptic *E. coli* protein YbhP functions as an endonuclease (Fig. 4), and identified the family of small molecule co-effectors bound by the poorly characterized transcription factor YhaJ (Fig. 5). Both these cases and several other successful experimental follow-ups are described in two forthcoming manuscripts [Rahimpour and Freddolino, submitted; Rahimpour, Zhang, Zhang, and Freddolino, in preparation]. Other research groups have likewise already begun to make use of these predictions, for example, to identify the heme binding site of an unusual protoporphyrinogen oxidase [18].

As of this writing, we are in the final stages of preparing an online database to provide all of the structure/function predictions described above in a freely available online database; we currently refer to the database as the *E. coli* Protein Information Collection (EPIC). We will also coordinate with maintainers of other *E. coli* genomic resources (*e.g.*, EcoCyc [20]) to incorporate our annotations to whatever extent possible. All annotations derived from the work proposed below will similar be disseminated to the community both through peer-reviewed publications, a central database on our institutional webpage, and through coordination with the maintainers of appropriate organism-specific

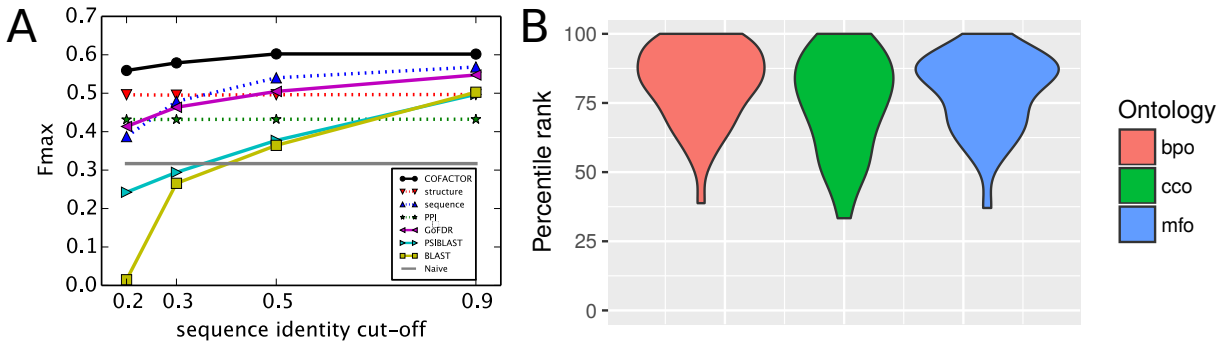

Figure 2: COFACTOR performance in internal and external assessments. (**A**) Performance of COFACTOR in an internal re-annotation task on a set of ∼1,200 well-annotated proteins from the *E. coli* proteome. Dashed lines show the perfomance of the COFACTOR component pipelines. The y-axis indicates the Fmax performance metrics used in CAFA competitions [15, 16]; along the x axis, the value indicates a sequence identity threshold such that all entries in functional databases with higher than the specified sequence identity to the target protein are masked (thus, moving from right to left simulates the case of attempting to annotate proteins with only progressively more distance functional templates available; *n.b.* for this benchmark all *structure* predictions were performed masking all potential templates above 30% identity, but subsequent annotations still used the database thresholding specified on the x axis). GoFDR is from [17]; BLAST and PSIBLAST are baseline methods used in CAFA. Adapted from [13]. (**B**) Distribution of percentile ranks of COFACTOR across all target collections and scoring metrics in the CAFA3 competition (assuming no ties), based on final rankings provided via private communication from the organizers. "bpo", "cco", and "mfo" refer to the biological process, cellular component, and molecular function ontologies, respectively.

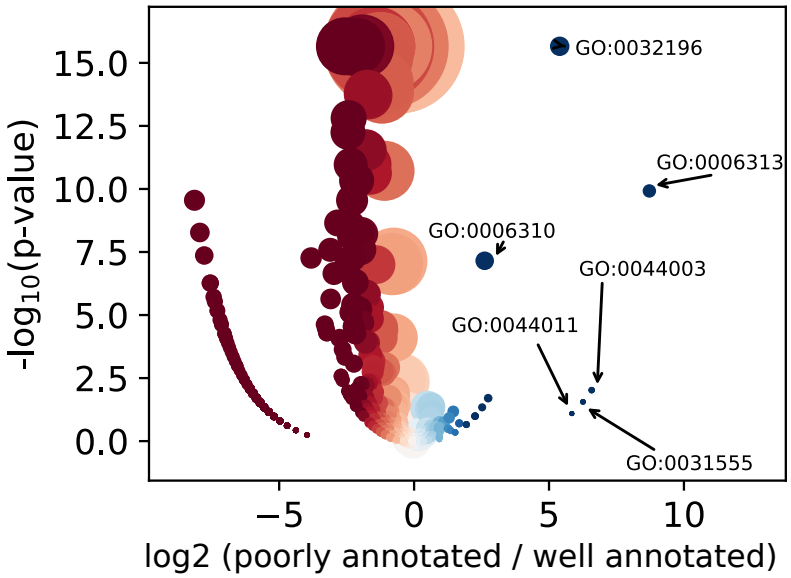

Figure 3: Volcano plot showing over- or under-representation of GO terms among the poorly annotated set of proteins in *E. coli*, based on whole proteome structure/function predictions. The x axis shows the $\log_2$ ratio of abundance of a GO term in the poorly annotated vs. well annotated fraction of the genome for high-confidence predictions; the y axis shows $\log_{10}$ p-values using a binomial test, and the size of each point is proportional to the number of proteins in that term. The results shown are for Biological Process GO terms only, similar patterns are apparent for the other GO aspects.

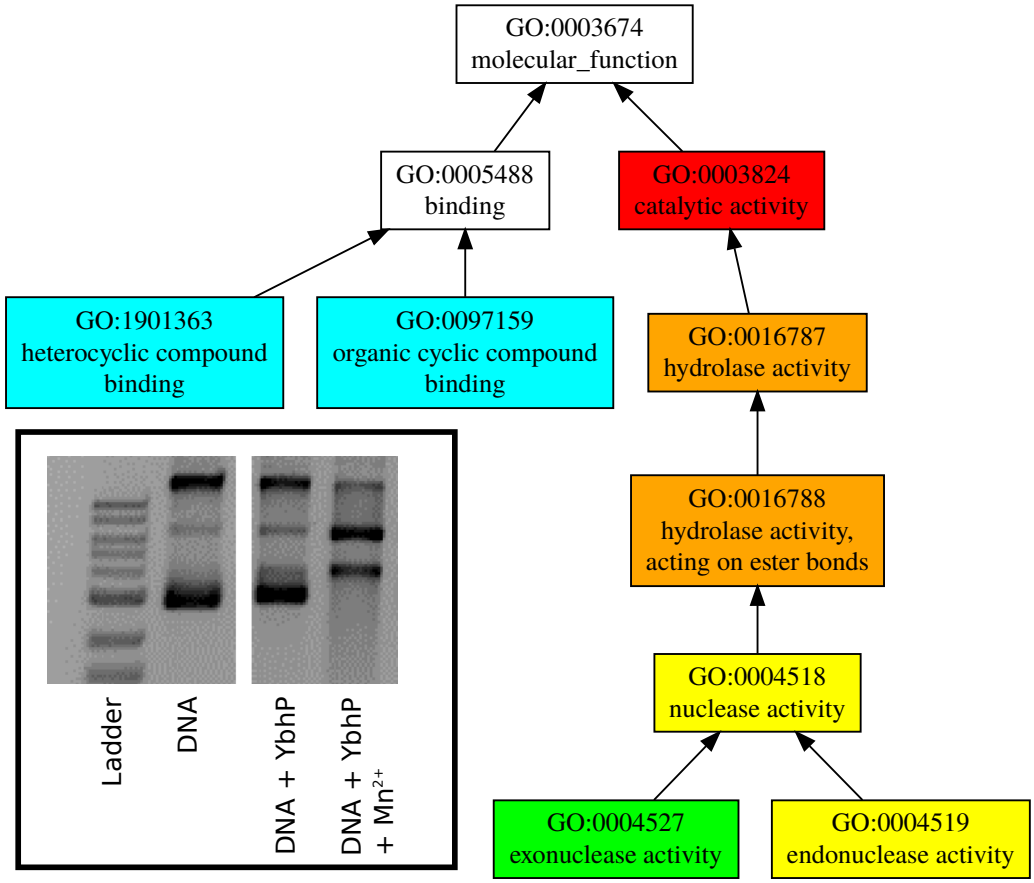

Figure 4: Example COFACTOR output for the poorly annotated *E. coli* gene *ybhP*. The GO hierarchy for molecular function (MF) GO terms is shown, with increasing confidence as the color of terms scales from blue to red (white terms have confidence scores of less than 0.5). Inset: Follow-up experiment demonstrating manganese-dependent endonuclease activity of purified YbhP in a plasmid supercoiling relaxation assay; irrelevant gel lanes are omitted. [Rahimpour, Zhang, Zhang, and Freddolino; manuscript in preparation]

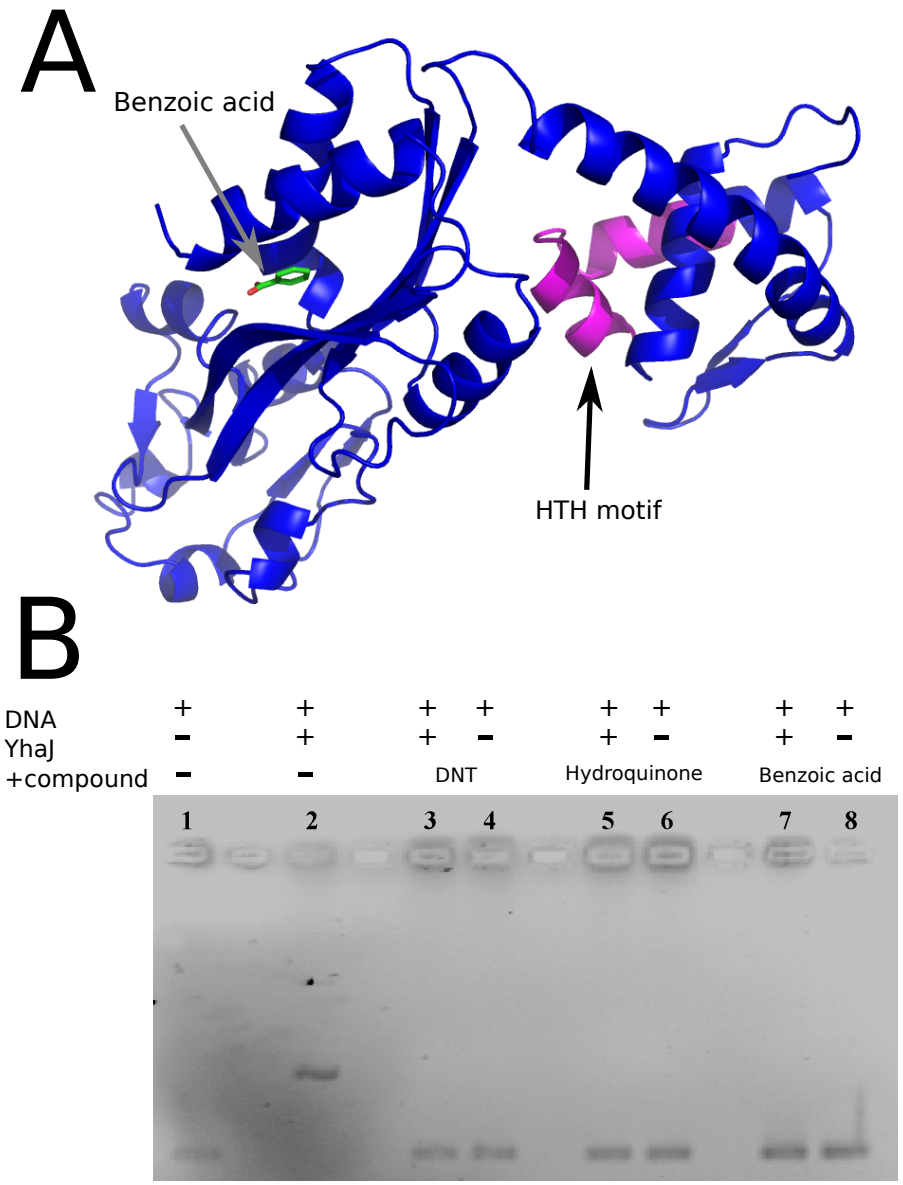

Figure 5: Prediction and testing of response of the poorly annotated *E. coli* transcription factor YhaJ to small organic molecules. (**A**) Predicted structure and ligand binding site of YhaJ, highlighting a predicted binding site for benzoic acid and DNA-binding helix-turn-helix (HTH) motif. (**B**) Gel shifts showing response of YhaJ to several small organic compounds including 2,4-dinitrotoluene (DNT) while binding the *yhaJ* promoter. Consistently, and in parallel with our experiments, [19] recently showed that YhaJ mediates transcriptional responses to DNT, although direct effects on YhaJ binding by the effector were not shown.

community databases.

2.3 Research Plan

In order to substantially expand the breadth of high-quality whole-proteome structure/function predictions available to the microbiology community, we will apply the complete COFACTOR/MetaGO workflow to the poorly annotated proteins of three important model microbes:

Bacillus subtilis strain 168 – *B. subtilis* serves as an important and widely used model organism for chromosomal biology and differentiation, as well as a major strain used in industrial enzyme production [21]. The *B. subtilis* genome includes 4,307 protein coding genes, of which 2,639 are poorly annotated (*i.e.*, UniProt annotation score of 1 or 2 out of 5 [1]) and 1,245 are especially poorly annotated (UniProt annotation score of 1).

Caulobacter crescentus CB15 – *C. crescentus* is a heavily used bacterial model for development, motility, and chromosomal biology (*e.g.*, [22, 23]. The *C. crescentus* CB15 genome includes 3,725 annotated protein coding genes, of which 3,423 are poorly annotated and 2,915 have a particularly poor annotation score of 1.

Streptomyces coelicolor M145 – *S. coelicolor* is a model member of the *Actinomycetales* order, which collectively have been the origin half of currently used antibiotics and numerous classes of chemotherapeutic, immunomodulatory, and antifugal medications [24]. The *S. coelicolor* M145 genome contains 8,124 protein-coding genes, of which 7,630 are poorly annotated and 7,047 have a particularly poor annotation score of 1.

For each of the organisms noted above, we will apply I-TASSER [25] to obtain a predicted structure (omitting only the small number of cases for which crystal structures are already available), followed by application of our COFACTOR/MetaGO annotation pipeline to obtain predicted annotations and small molecule binding partners [13, 14]. While it would be fruitful to perform combined structure/function modeling on the entire proteomes, or at least all proteins with UniProt annotation scores of 2 or less, simply due to immense computational resources required, in the plan below we will restrict ourselves only to the particularly poorly-annotated proteins of each organism that have UniProt annotation scores of 1.

Experimental validation of a selected subset of cases, plus analysis of the overall trends present in the prediction sets and concordence with existing annotations, will be performed using local resources (both financial and computational), and is funded by the above-noted grants. In addition, as with the EPIC database noted above, we will subsequently make all predictions available to the community for both individual and bulk download, hosted on resources provided by the University of Michigan. Obtaining complete functional annotations for several additional bacterial species to stand alongside our already-obtained predictions for *E. coli* will enable us to assess the accuracy of current computational predictions in databases such as UniProt, which were typically derived using earlier sequence-based methods, and may have error rates of 80% or more for some protein families [6]. The datasets that we establish will also provide an immensely valuable resource for any researchers working on the targeted organisms, for example, enabling immediate access to structural information and functional insight to scientists studying the developmental role of any poorly annotated *B. subtilis* protein, or determining the biosynthetic capabilities enabled by a particular pathway in *S. coelicolor*.

3 Resource Usage Plan and Resource Justification

All structure predictions and annotation calculations described in Section 2.3 will be performed using I-TASSER and COFACTOR, which have already been ported to and optimized on SDSC Comet. I-TASSER [25] involves force field based structural refinement using a combined physics-based and knowledge-based potential; all of the required calculations are performed in parallel to generate a large

ensemble of 13,000 models, from which many conformations (typically 13,000) from which up to 5 final models are selected by structure clustering. Because the calculations are embarrassingly parallel, consisting of a huge number of single-threaded calculations, no particular effort is necessary to achieve good scalability, only appropriate run management to ensure that all cores of allocated nodes are kept busy throughout the calculation.

Hybrid structure/sequence/PPI based function prediction will be carried out using COFACTOR [13] adapted to use the refined MetaGO combining rules [14]. The COFACTOR algorithm draws the functional insight mainly from global and local structure alignment of I-TASSER models to structure templates in the BioLiP database. For each query protein, this structure alignment process typically takes a few hours for one single thread job, with time complexity linearly correlates with the number residues in the query protein.

We provide scaling and performance information based on benchmarks run on SDSC Comet in the attached Code Performance & Resource Costs document. Based on those benchmarks, our resource needs are as follows (note that because we find that our computational costs are linear with protein size, it suffices to calculate costs based on the average protein length observed in each organism and the number of proteins). Data storage estimates are given based on the fixed set of output files generated by I-TASSER and COFACTOR, and are likewise linear with system size, at roughly 0.16 MB per amino acid.

Bacillus subtilis: For 1,245 target proteins with an average size of 170 amino acid residues, we request $(1245 * (37.5 + 2.564 * 170)) = 589,358$ Comet SUs. In addition, we request $(1245 * 170 * 0.00016) = 33.9$ GB of Oasis storage for the data produced.

Caulobacter crescentus: For 2,915 target proteins with an average size of 315 amino acids, we request $(2915 * (37.5 + 2.564 * 315)) = 2,463,641$ Comet SUs. In addition, we request $(2915 * 315 * 0.00016) = 146.9$ GB of Oasis storage for the data produced.

Streptomyces coelicolor: For 7,047 target proteins with an average size of 318 amino acids, we request $(7047 * (37.5 + 2.564 * 318)) = 6,010,048$ Comet SUs. In addition, we request $(7047 * 318 * 0.00016) = 358.6$ GB of Oasis storage for the data produced.

Total request: Summing the above requests for the three organisms under consideration, we request a total of **9,063,047 Comet SUs** and **539.4 GB of Oasis storage**.

3.1 Access to other computing resources

The PI has access to a high-memory 32-core server used by his research group for data analysis and small simulations, and additional computing time on the University of Michigan's large Flux cluster on a pay-per-use basis. These local computing resources will be relied upon for data analysis and postprocessing, relying on XSEDE resources only for production calculations (structural modeling and annotation calculations). However, the pay-per-use model of the Flux cluster would make performing the production runs described here prohibitively expensive. The PI also has access to modest allocation on the PSC anton2 machine, but that allocation is earmarked for a completely different project (not described here). In addition, as described in the attached Progress Report, many of these local computing resources will, to the extent available, be allocated to complete analysis and secondary calculations related to our XSEDE simulations performed over the last allocation period, whereas our present request for XSEDE resources is based solely on the protein structure/function predictions described above.