Journal of Proteomics & Bioinformatics - Open Access

Research Article
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000150
JPB/Vol.3 Issue 9

Functional Insights from Computational Modeling of Orphan Proteins Expressed in a Microbial Community

Korin E. Wheeler[1], Adam Zemla[2], Yongqin Jiao[1], Daniela S. Aliaga Goltsman[3], Steven W.Singer[1], Jillian F. Banfield[3] and Michael P. Thelen[1]*

[1]Physical and Life Sciences Directorate, University of California, Berkeley
[2]Computations Directorate, Lawrence Livermore National Laboratory, University of California, Berkeley
[3]Department of Environmental Science, Policy and Management, University of California, Berkeley

Abstract

Environmental genomics and proteomics data are heavily populated with proteins that are not homologous to experimentally characterized proteins. We approached this problematic area by investigating a natural microbial community from a highly constrained niche in which critical roles are likely carried out by proteins of unknown function (ORFans). Based on several criteria, these proteins were not statistically similar to any protein sequences in the SwissProt database. We selected a target set of 545 ORFans and weakly annotated proteins expressed by the dominant bacterial member of the community, *Leptospirillum* Group II, and used an automated modeling system (AS2TS) incorporated with other computational tools to predict structures. This generated 484 models, 89% of the target set. Structure-based superfamilies, general functional categorizations, and specific gene ontology (GO) functions were predicted for 424, 386, and 117 ORFans, respectively. Structural predictions and classifications were integrated into a manually curated database, outlining *in silico* calculations and available proteomic data for each protein. This analysis facilitated the development of experimentally testable hypotheses for several enigmatic proteins, including confident predictions of copper transport proteins and cyclic diguanylate signaling proteins. As DNA sequencing of natural organisms rapidly expands, this computational structure-function approach can be applied to guide experimental testing of the structure and function of challenging ORFans.

Keywords: ORFan; Proteins of unknown function; Structural modeling; SCOP; Superfamily

Abbreviations: AMD: Acid Mine Drainage; AS2TS: Amino-acid Sequence to Tertiary Structure homology modeling system; GO: Gene Ontology database; CASP: Critical Assessment of Techniques for Protein Structure Prediction; CAT: categorically best models; CopA: copper-translocating P-type ATPase; PDB: Protein Data Bank; SCOP: Structural Classification of Proteins database

Introduction

Functional identification of proteins in a sequenced organism or natural community poses a critical challenge and has sparked great interest in high-throughput annotation approaches. Even for the well-studied *E. coli* species, 34% of the proteome consists of functional ORFans (Hu et al., 2009), with either insignificant sequence similarity to any known proteins, or only low confidence, broad generic annotations (Fischer and Eisenberg, 1999). Novel proteins identified from environmental genomic and proteomic studies of communities that include uncultivated organisms are especially important in understanding microbial biology and evolution. Although difficult to study experimentally, environmental samples provide great insight into biochemical contributions to biodiversity and distinctive adaptation mechanisms to niches within ecosystems. Novel proteins from environmental samples provide a window into the physiology and ecology of these diverse and complex communities. Nevertheless, analysis of large-scale metagenomic projects including surface seawaters, whale falls, soil, and acid mine drainage locations has indicated that 27-48% of genes sampled have no known function based on automated sequence similarity methods (Harrington et al., 2007). The novelty of these functionally unknown proteins makes them difficult to characterize, but underpins their key roles in distinctive aspects of adaptation and function in various ecosystems.

Our interest in microbial communities has led us to examine ORFan proteins that are expressed in a natural, extremophilic microbial community collected from an acid mine drainage (AMD) environment. The community grows as floating biofilms in hot, sulfuric acid rich solutions (pH ≤ 1) with high heavy metal concentrations (Tyson et al., 2004). Extensive proteogenomic analyses of this AMD community found that 42% of the proteome consists of proteins of unknown function, or expressed ORFans (Ram et al., 2005). We use the term *expressed ORFans* throughout, to indicate proteins that are identified by mass spectrometry (MS)-based proteomic analysis but have limited or no statistical similarity to annotated protein sequences. Based on previous studies, many of the expressed ORFans are present in high concentrations in these biofilm communities, indicating important functions in survival and community fitness (Ram et al., 2005).

Protein structure is a primary means of evolutionary selection. Thus, structure prediction is a powerful tool to assess function. Importantly, it is applicable well below sequence identity limits required by sequence alignment-based methods (Gough et al., 2001; Adams et al., 2007). Previous studies have explored the link between structural superfamilies and their functions, and show a strong tie between Structural Classification of Protein (SCOP) superfamilies and molecular functions (Adams et al., 2007; Malmström et al., 2007). Structural modeling has been performed previously on the genomic

***Corresponding author:** Michael P. Thelen, Physical and Life Sciences Directorate LLNL, L452, Livermore, CA, Tel: 925-422-6547; Fax: 925-422-2282; E-mail: mthelen@llnl.gov

#Current addresses: Department of Chemistry and Biochemistry, Santa Clara University, Santa Clara, CA 95053, SWS:Lawrence Berkeley National Laboratory, Mail Stop 90-R1116, Berkeley, CA 94720

Received September 02, 2010; **Accepted** September 20, 2010; **Published** September 22, 2010

Citation: Wheeler KE, Zemla A, Jiao Y, Aliaga Goltsman DS, Singer SW, et al. (2010) Functional Insights from Computational Modeling of Orphan Proteins Expressed in a Microbial Community. J Proteomics Bioinform 3: 266-274. doi:10.4172/jpb.1000150

Copyright: © 2010 Wheeler KE, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

scale with a few reports providing high-throughput functional insights (Huynen et al., 1998; Rychlewski et al., 1998; Sánchez and Sali, 1998; Bonneau et al., 2004; Zhang and Skolnick 2004). A recent study (Malmström et al. 2007) parsed proteins into domains and coupled large scale structure predictions with functional assignments by integration of SCOP superfamilies and gene ontology (GO) (Ashburner et al., 2000), providing insights into both structure and function on the domain level.

Structural modeling and analysis methods described here have been used to guide studies on individual ORFans expressed by the AMD microbial community. For example, this approach provided basic functional assessment of an isocitrate dehydrogenase (Goltsman et al., 2009) and facilitated the experimental design and testing of a highly expressed and novel cytochrome (Singer et al., 2008). Here, we expanded our approach to include over 500 expressed ORFan and weakly annotated proteins from the dominant bacterium of the AMD community, *Leptospirillum* Group II (Tyson et al., 2004), and to integrate structural predictions with expression data (Ram et al., 2005; Goltsman et al., 2009). For 422 (77%) of the proteins analyzed, no functional annotation was available through sequence alignment programs such as iterative PSI-BLAST.These ORFan proteins were not homologous to any proteins in the SwissProt database, as inferred by sequence identities below 30% and other conventional statistical measures of similarity. In our study we explored the structural predictions, structural relationships, and expression data to aid in development of experimentally testable hypotheses for the roles of specific proteins within this extremely acidic, metal rich environment.

Materials and Methods

Expressed ORFan protein dataset

Leptospirillum environmental Group II ORFan protein sequences were chosen from metagenomic datasets (Tyson et al., 2004; Goltsman et al., 2009) that fit two criteria: 1) Sequence-based approaches gave little or no indication of protein function; and 2) Proteomic datasets from AMD community studies indicated relatively high expression. Automatic annotations of *Leptospirillum* Group II were run as described previously (Ram et al., 2005) and these were manually curated (Goltsman et al. 2009). In prior studies of these AMD biofilm communities (Tyson et al., 2004; Ram et al., 2005; Goltsman et al., 2009), the term “protein of unknown function” was used when a hypothetical gene product (<30% sequence identity) was identified as an expressed protein. “Probable” was added to functional descriptions for predicted proteins with a sequence identity between 30% and 70% (irrespective of the alignment length) to homologous proteins in the SwissProt database, but which lacked certain functional elements or domains. For these cases, BLAST matches in the NCBI non-redundant (nr) protein sequence database (http://blast.ncbi.nlm.nih.gov/) were also considered. Using all of these criteria, a total of 545 proteins were designated as expressed ORFan proteins from *Leptospirillum* Group II (Ram et al., 2005; Goltsman et al., 2009). Of these, 317 (58%) are unique proteins of unknown function, 110 (20%) are conserved proteins of unknown function, and 118 (22%) are weakly annotated proteins, previously described with a probable function (Goltsman et al., 2009). Signal peptides, which most likely lack any relevance to the overall structure and function of proteins in their designated cellular locations, were predicted using SignalP 3.0 (Bendtsen et al., 2004) and truncated from the full length protein sequences, where appropriate. Based on the sequence without the signal peptide, each protein’s molecular weight and isoelectric point (pI) were calculated using Compute pI/Mw (Gasteiger et al., 2005). Protein expression based on MSproteomic data was estimated using the normalized MS spectral counts (Zybailov et al., 2006) as previously reported (Goltsman et al., 2009).

Whole protein structural modeling

Comparative structural modeling techniques were chosen due to their high reliability and low computational demands (Moult et al., 2007). For the best results in identification of structural templates for modeling, several different techniques were combined (Ginalski et al., 2005) with AS2TS, as previously described (Zemla et al., 2005). In addition, AS2TS iteratively generated local libraries to support multiple sequence alignments and created local databases of intermediate models to aid in structural template selection. These steps were repeated for each protein until no new libraries or intermediate models were generated. In the case of long sequences, multiple runs were performed using fragmentation of the query sequence into ≤700 residue segments. Structural alignments between all templates identified and preliminary models were calculated with LGA (Zemla, 2003), and secondary structure predictions were calculated with PSIPRED (Jones, 1999). All of these results were used for the final selection of structural templates and to further guide the process of 3D model construction. Regions of insertiondeletion or uncertain sequence-structure alignments were built as loops using LGA by grafting in suitable fragments from related structures in the Protein Data Bank (PDB). Finally, models were completed using SCWRL (Bower et al., 1997) to predict coordinates for missing side chain atoms.

After protein models were created, they were classified by standard grouping criteria (Table 1). It is expected that above 45% sequence identity the model is as close to the correct structure as to the template (Baker and Sali, 2001); thus, we placed these models in the best, or ‘A’, category (our criteria for similarity to the templates from PDB or to intermediate AS2TS models: sequence identity >45% and alignment overlap >75%). Category ‘B’ (sequence identity >20% and alignment overlap >75%) models overlap with the twilight zone of 20-35% sequence similarity and the required structural completeness of the model. In our classification, category C1 (sequence identity >15% and alignment overlap >50%) models gave an overall structure that would either be roughly correct or contain only single domains of the whole protein. In the final two categories, C2 models retained very little similarity to the template structure, resulting in only a small fraction of the overall protein modeled. The C3 proteins retained so little similarity to structures in the PDB (or to intermediate AS2TS models) that no structural model could be confidently constructed. From multiple possible models, we considered the top seven models constructed for each protein based upon the quality of alignment with the identified structural templates, according to the best: e-value; sequence identity; sequence coverage (alignment overlap); alignment compactness (minimal number of gaps); alignment overlap at the Nterminus; and alignment overlap at the C-terminus. The final model, which we indicated as the categorically best (CAT) model, ranked highest in each of the following three categories: evalue, sequence identity, and sequence coverage (Table S1). The quality of all automatically created structural models was evaluated using the Procheck package (Laskowski et al., 1993). For Category A models an average percent of residues in disallowed regions was only 0.54% with a median of 0.15%; for Category B models, 0.95% and 0.85%; for Category C1, 1.18% and 1.00%; and for Category C2, 1.52% and 1.30%. More detailed evaluation of the local quality of the created structures was not included for the function annotation approach described here. Further improvements of evaluation procedures and possible

Research Article
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000150
JPB/Vol.3 Issue 9

refinements of automatically created models were not critical for the current data processing since we mostly concentrated on the accuracy of calculated alignments, PDB template identification, template selection, and structure comparison-based assignments of the created models to proper SCOP folds and Superfamilies. In particular, the results from the analysis of calculated multiple structure alignments enhance our confidence in the identified critical residues and Superfamily assignments. For each protein, we performed structural comparisons between the models created and the identified structural templates. Results are available through our protein model website at http://proteinmodel.org/AS2TS/research/M_Thelen/FUN_545/. Examples of analysis and comparison plots are provided here (Figure 4B) with similar summary results (comparison plots) provided on the web for each modeled protein.

Structural and functional assessment

The CAT models created were compared to the structural domains from the SCOP (Murzin et al., 1995) database (release 1.73, Sept. 2007) using LGA, and clustered to ASTRAL_95 (Brenner et al., 2000; Chandonia et al., 2004). Clustering was based on structural alignments performed by LGA (distance cutoff set at 4 Å). Positive matches to SCOP domains were constrained by the following criteria: (1) LGA_S >35%, used as a scoring function to evaluate the overall level of structure similarity (local and global), calculated relative to the modeled protein; (2) LGA_M >50%, used to avoid matches to only short fragments from SCOP domains, so the model should cover a larger portion of the domain and with a structure similarity score of at least 50% relative to the SCOP domain; and, (3) tight local superposition of C-alphas, where at least 10 residues from continuous segments were within a local RMSD cut-off <0.5 Å (Zemla et al., 2007). Each domain hit that passed our structure similarity criteria, up to a total of ten, was scored (Table S2).

General and specific functions were assigned to proteins annotated by SCOP Superfamily using the SUPERFAMILY database (Vogel et al., 2004; Vogel and Chothia, 2006). When available, specific GO functions (Ashburner et al., 2000) were added as provided by the SUPERFAMILY2GO database (Gough et al., 2001), which compiled abstracts from InterPro (Hunter et al., 2009) to correlate SCOP superfamilies with GO functions.

Results and Discussion

Structural modeling

It has been demonstrated that protein modeling by comparison is the most reliable method for structural predictions (Moult et al., 2009; Venclovas et al., 2003). Therefore, in this study we applied the AS2TS modeling system (Zemla et al., 2005). AS2TS is primarily focused on the modeling at the domain level; however, we utilized a set of all identified alternative templates, which may cover different domains, enabling prediction of whole protein structure and providing data for insights into multidomain protein function. AS2TS accesses a set of tools for structure similarity assessment (http://proteinmodel.org) that facilitates structural predictions, refines the models created, and aids functional prediction (Cosman et al., 2008; Zemla and Zhou, 2008; Anisimov et al., 2010; Chakicherla et al., 2009). For each modeled protein these structure comparison and analysis tools can be applied to the set of identified templates, providing possible insights into evolutionary relationships based upon structure. For identification of SCOP superfamilies, we avoided domain parsing applications used in previous studies (Bonneau et al., 2004; Malmström et al., 2007) and simplified the approach by identifying SCOP superfamilies using structural features within the best models constructed by AS2TS. This straightforward approach reduced computational time and avoided the introduction of additional errors from parsing techniques (Holland et al., 2006).

MS proteomics analysis indicates that 545 ORFans or weakly annotated proteins are expressed by the dominant organism, *Leptospirillum* Group II (Ram et al., 2005). Using AS2TS in conjunction with other molecular structure tools, structural models (complete or fragmented) were predicted for 484 (89%). Models were grouped into categories A, B, C1 and C2 (Table 1) according to quality and confidence. A total of 125 models (23%) were high confidence, with sequence coverage greater than 75% and sequence identity greater than 20% when compared to templates (A or B quality, Figure 1). For the majority of the highest quality proteins modelled (73% of category A proteins), the best PSI-BLAST search result was a match to another protein of unknown function from different genus (Table S1a). This emphasized that an approach comparing structural models was capable of providing information beyond what is available through basic sequence comparison tools.

The 210 lower quality models in the C1 quality category did not meet a sufficient level of sequence identity or coverage (Sánchez and Sali, 1998), but may still provide insights into structure that could guide experimental approaches. Even when single structural templates and alignments did not cover the entire query protein sequence, PDB structure searches were often able to identify alternative templates that could be combined to enable more complete modelling and, in many cases, provide some insights for functional hypotheses. Similarity in predicted structures derived from multiple templates imparts additional confidence in a compiled structure and in eventual functional hypotheses.

In an analysis of modeling efforts, we looked for biases in model quality based upon physicochemical properties of the polypeptides or a bias towards certain kinds of structural templates. *Leptospirillum* Group II, an acidophilic bacterium living at pH ~1, has an average pI for the entire proteome of approximately one pH unit higher than common neutrophilic microbes (Ram et al., 2005). The high average pI is a result of a change in the proportion of charged amino acids, making correct functional annotations more difficult when based on sequence similarities alone (Figure S1). Although many proteins of *Leptospirillum* Group II have unusually high calculated pI values (Ram et al., 2005), we found that the quality of structural modeling was independent of pI (data not shown). Not surprisingly, however, molecular weight was inversely correlated with model quality: Larger proteins were generally more difficult to model over the entire sequence and resulted in models of lower confidence (Table 2).

Many of the high quality models relied upon modeling templates from structural genomics projects, indicating the significant role of these projects in diversifying the available protein structures to enhance homology modeling. To emphasize the utility of structural genomics, 241 expressed ORFans (44%) were modeled using at least one template from a structural genomics project. Those templates were particularly useful in generating high quality models, 68% of which fell within category A or B. There were also 21 proteins within

Category	Sequence Identity	Coverage
A	>45%	>75%
B	>20%	>75%
C1	>15%	>50%
C2	very low or no homology	
Not Modeled (C3)	no homology	

Table 1: Model quality assessment criteria.

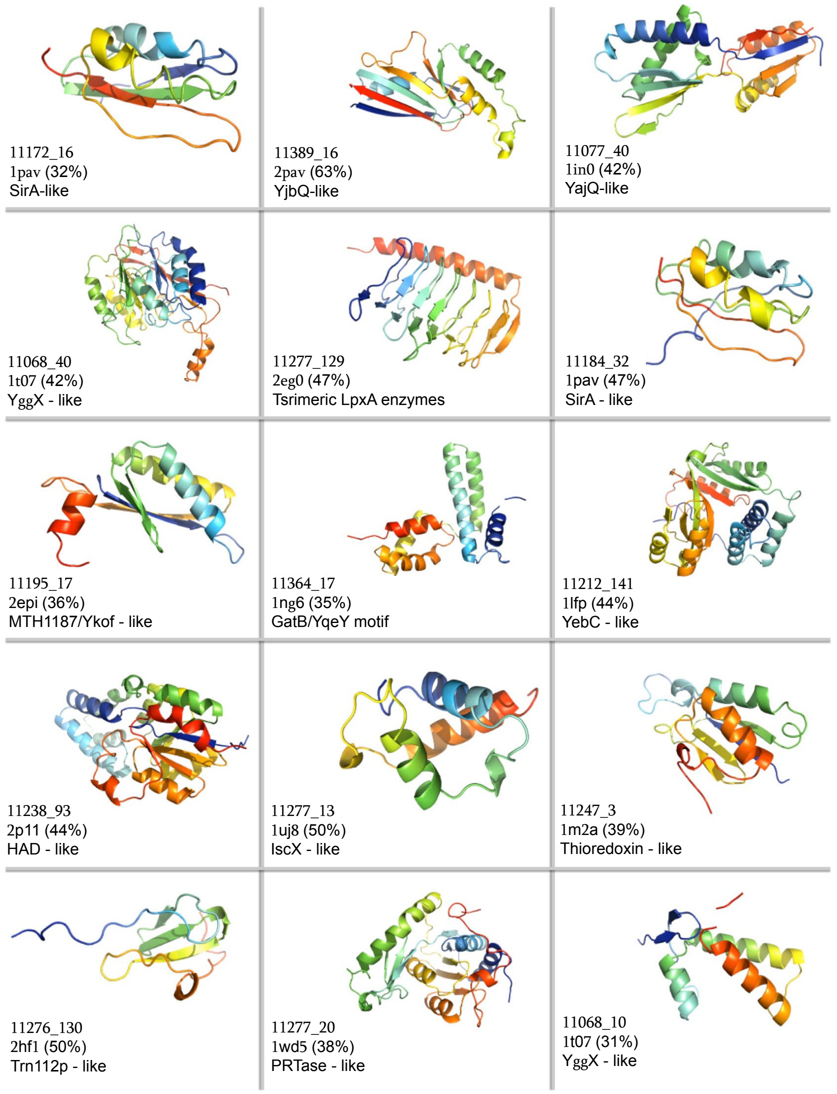


Figure 1: Category A quality models with top SCOP Superfamily assignments. For each protein model: cartoon plots colored blue to red from the N- to C- termini, followed by gene identification number, information about close structural templates from PDB, and SCOP Superfamily assignment (text from top to bottom).

		Molecular weight (kDa)		Structural Genomics Template	
Model category	Total proteins	Maximum / Minimum	Median	In the top 7 models	In the CAT model
A	16	37 / 6.7	17	15 (94%)	10 (62%)
B	109	59 / 6.8	18	70 (64%)	50 (46%)
C1	210	107 / 7.7	25	93 (44%)	30 (14%)
C2	149	81 / 7.9	30	62 (45%)	31 (21%)
C3	67	62 / 2.9	18	-	-

Table 2: Analysis of structural models by category. Distribution of model quality, molecular weight and structural genomics templates utilized are categorized for each created model.

our dataset for which the construction of structural models relied solely upon structural genomics templates.

SCOP Superfamily and functional assessment

To assess general functional categories and verify the quality of structural modeling, SCOP Superfamily domains were matched to 78% of the 484 expressed ORFans modeled by searching for structural domains within each of the seven models produced (Table S1). Top Superfamily assignments for the Category A models are shown in Figure 1. The SUPERFAMILY database was used to obtain a distribution profile of the general functional assignments to each SCOP Superfamily (Figure 2), and indicated that most of them predict metabolic functions. More detailed functional predictions were obtained for 24% of the modeled proteins with specific GO functions (Table S1).

Because of the abundance of small proteins (<40kDa) in our dataset, the majority were found to be single domain proteins associated with only one SCOP Superfamily (Table S1). Strong

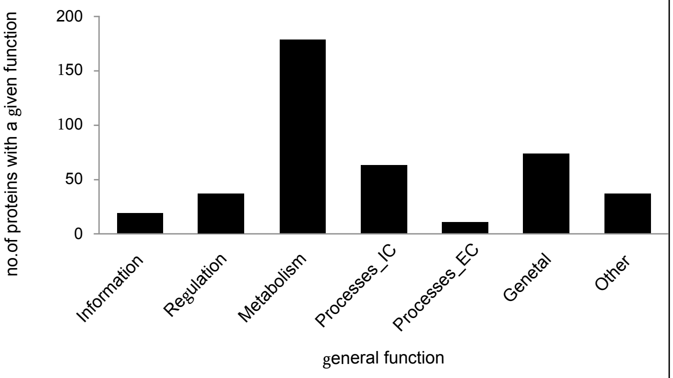


Figure 2: Overview of the SUPERFAMILY general functions for the primary SCOP superfamilies identified within the structural models.

A. Weakly annotated proteins	No. of proteins
SCOP Superfamily functions in agreement with sequence-based annotations	88
Functional predictions with discrepancies	12
Models with no SCOP Superfamily match	15
Proteins not modeled	3
Total	118

B. Expressed ORFans	conserved	unique
New functional predictions based upon assigned SCOP Superfamilies	89	231
Modeled proteins with no SCOP Superfamily	9	55
Proteins not modeled	12	31
Total	110	317

Table 3: Summary of SCOP Superfamily assignments and predicted general functions. **(A.)** Weakly annotated proteins are compared to previously published sequence-based annotations and **(B.)** expressed ORFans are separated by their sequence-based annotation as "conserved" or "unique proteins of unknown function." Functional insights into conserved ORFans have implications beyond the biochemistry of the AMD community.

matches to a SCOP superfamily were obtained for 35%, or 167 protein models, with a structural alignment (LGA_S) score greater than 75%. Additionally, we found an inverse correlation between the quality of model and the number of different SCOP Superfamilies identified for each protein, which suggested that lower quality models have a higher propensity for false positives. Of the models with five or more identified SCOP Superfamilies, 75% were C1 or C2 quality. These low quality models are often fragmented and align well to multiple SCOP Superfamilies.

To assess the accuracy of structural modeling in providing functional insights, sequence-based functional assignments for a small set of weakly annotated proteins were included in our dataset (Goltsman et al., 2009) and compared to functional information extracted from structural modeling and SCOP Superfamily assessment. A total of 86 SCOP Superfamily identifications confirmed previous low-confidence, or 'probable', sequence-based functional annotations (Table 3). In the final 15%, the structure based approach was inadequate to provide a domain-based function; in large part, this was due to the inability to cluster models to any known SCOP Superfamily, either because of the low quality of the models created, or simply because the SCOP database is not as current as the PDB.

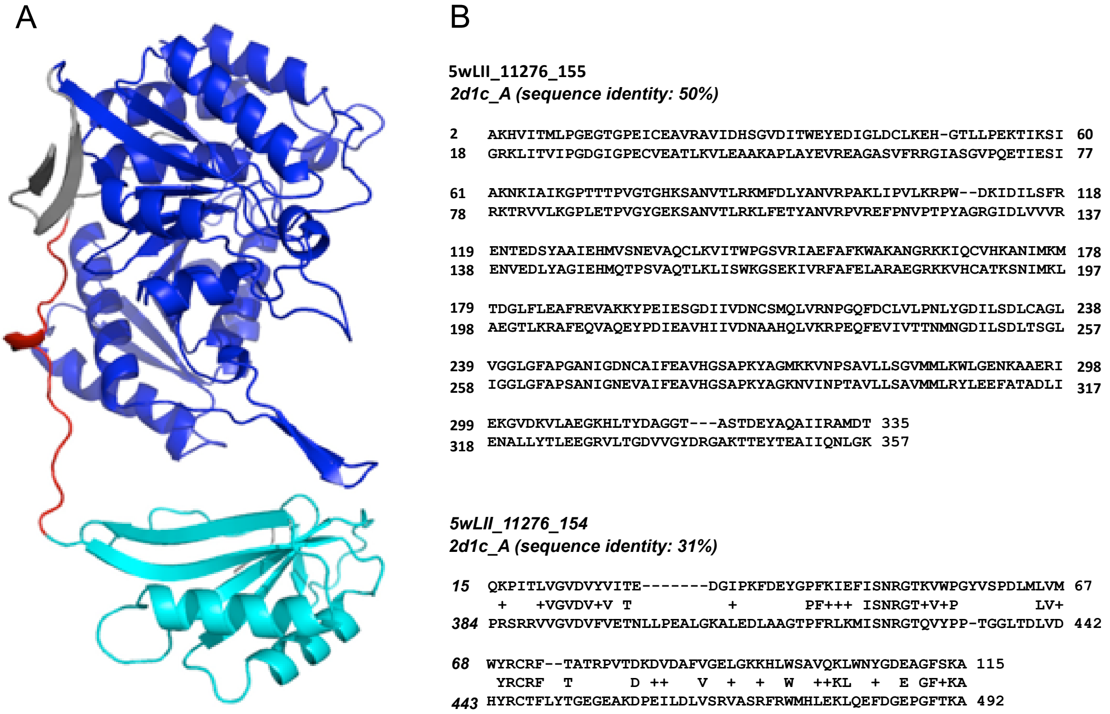
Predicted protein functions related to AMD

By structural prediction and Superfamily assignment, functional predictions were considered in the context of their relationship to potential adaptations to the AMD environment and microbial community life style. For example, harsh conditions may necessitate the prevalence of ORFan proteins that have predicted DNA binding and repair functions, including five restriction endonuclease-like SCOP superfamilies and four lambda repressor-like DNA-binding domains (Table S1a). The best PSI-BLAST match to all but one of these nine proteins was to another hypothetical protein, and five of these nine are conserved ORFans. Experimental validation of these proteins would therefore provide annotation across several genera. Moreover, we hypothesize that three of the proteins identified here with thioredoxin-like SCOP superfamily domains may be involved in sulfur metabolism, including genes 11389_17, 11233_42 and 11238_7. Sulfur metabolism is expected to be important in the AMD community as both defense against sulfur-containing radicals, and as disulfide isomerases to aid in protein folding (Pott and Dahl, 1998).

Energy metabolism functions are also considered crucial under the AMD conditions. Five of the ORFan proteins modeled were matched to DsrEFH-like domains, a group of energy-related SCOP domains found in DsrEFH-like proteins (Table S1a). DsrEFH-like domains, although poorly characterized, have been experimentally linked to sulfur metabolism for energy generation (Pott and Dahl, 1998; Galvagnion et al., 2009). Along with a previously identified siroheme-like enzyme, a rhodanese-like protein and sulfide quinine reductase (Goltsman et al. 2009), DsrEFH-like proteins are thought to be involved in sulfur oxidation, which may be important to energy metabolism given the abundance of sulfur present as pyrite (FeS2) in the AMD environment.

Other energy related proteins include eight *c*-type cytochromes and nine thioredoxin-like proteins. These were structurally modeled, and six have been identified here as new cytochromes and thioredoxin-like proteins based upon their predicted SCOP Superfamily domains (Table S1a). These proteins may be involved in Fe(II) oxidation, an energy source and process that contributes to the highly acidic mine drainage (Tyson et al., 2004; Ram et al., 2005). Two small proteins have been modeled to contain possible monoheme cytochrome domains (genes 11077_47 and 11077_6). Both are assigned GO terms for iron ion binding, electron carrier activity, and heme binding. The gene encoding one of the proteins (11077_6) is within an operon consisting of eight genes, two of which were previously annotated to encode probable cytochrome oxidases and in close proximity to a mono-heme subunit of cytochrome C oxidase and a probable iron-sulfur protein (Goltsman et al., 2009). Interestingly, the best structural template for protein 11077_47 was a p-cresol methylhydroxylase (PDB 1wve). Based on reports that p-cresol methylhydroxylase degrades the toxic phenol p-cresol in the protocatechuate metabolic pathway of other bacteria (Cunane et al., 2000), the protein encoded by 11077_47 could be involved in the degradation of aromatic compounds.

Cell wall proteins and stress-induced proteins are also important for microbial survival in the AMD environment. One protein predicted to have a PGBD-like SCOP Superfamily domain and a peptidoglycan-binding motif was gene 11276_107. This Superfamily has been shown to function in catalyzing the hydrolysis of the link between N-acetylmuramoyl residues and Lamino acid residues in certain bacterial cell-wall glycopeptides, essential to cell adhesion and bacterial cell wall biosynthesis (Foster, 1991). A stress inducible YceI protein, gene 11391_14, was predicted by structural modeling and was determined to be highly expressed in the proteomics dataset (Ram et al., 2005). In *E. coli* this is an alkaline pH induced periplasmic protein and is conserved in many bacteria and archaea (Stancik et



5wLII_11276_155
2d1c_A (sequence identity: 50%)

```
2    AKHVITMLPGEGTGPEICEAVRAVIDHSGVDITWEYEDIGLDCLKEH-GTLLPEKTIKSI  60
18   GRKLITVIPGDGIGPECVEATLKVLEAAKAPLAYEVREAGASVFRRGIASGVPQETIESI  77

61   AKNKIAIKGPTTTPVGTGHKSANVTLRKMFDLYANVRPAKLIPVLKRPW--DKIDILSFR  118
78   RKTRVVLKGPLETPVGYGEKSANVTLRKLFETYANVRPVREFPNVPTPYAGRGIDLVVVR  137

119  ENTEDSYAAIEHMVSNEVAQCLKVITWPGSVRIAEFAFKWAKANGRKKIQCVHKANIMKM  178
138  ENVEDLYAGIEHMQTPSVAQTLKLISWKGSEKIVRFAFELARAEGRKKVHCATKSNIMKL  197

179  TDGLFLEAFREVAKKYPEIESGDIIVDNCSMQLVRNPGQFDCLVLPNLYGDILSDLCAGL  238
198  AEGTLKRAFEQVAQEYPDIEAVHIIVDNAAHQLVKRPEQFEVIVTTNMNGDILSDLTSGL  257

239  VGGLGFAPGANIGDNCAIFEAVHGSAPKYAGMKKVNPSAVLLSGVMMLKWLGENKAAERI  298
258  IGGLGFAPSANIGNEVAIFEAVHGSAPKYAGKNVINPTAVLLSAVMMLRYLEEFATADLI  317

299  EKGVDKVLAEGKHLTYDAGGT---ASTDEYAQAIIRAMDT 335
318  ENALLYTLEEGRVLTGDVVGYDRGAKTTEYTEAIIQNLGK 357
```

5wLII_11276_154
2d1c_A (sequence identity: 31%)

```
15  QKPITLVGVDVYVITE-------DGIPKFDEYGPFKIEFISNRGTKVWPGYVSPDLMLVM 67
     +  +VGVDV+V  T          +       PF+++ ISNRGT+V+P        LV+
384 PRSRRVVGVDVFVETNLLPEALGKALEDLAAGTPFRLKMISNRGTQVYPP-TGGLTDLVD 442

68  WYRCRF--TATRPVTDKDVDAFVGELGKKHLWSAVQKLWNYGDEAGFSKA 115
     YRCRF  T      D ++   V  +  +  W  ++KL  +  E GF+KA
443 HYRCTFLYTGEGEAKDPEILDLVSRVASRFRWMHLEKLQEFDGEPGFTKA 492
```

Figure 3: Domain fusion model created for proteins 11276_155 and 11276_154, compared with *T. thermophilus* isocitrate dehydrogenase. (A) Structure of *T. thermophilus* isocitrate dehydrogenase (PDB 2d1c) is colored by its correspondence to the *Leptospirillum* Group II isocitrate dehydrogenase proteins: 11276_155, royal blue; 11276_154, cyan; N-terminal and linker regions with no corresponding residues in 11276_155, grey; and, linker region with very poor alignment to N-terminal region of 11276_154, red. (B) Sequence alignments calculated between PDB template 2d1c chain A and proteins: 11276_155 (top), and 11276_154 (bottom).

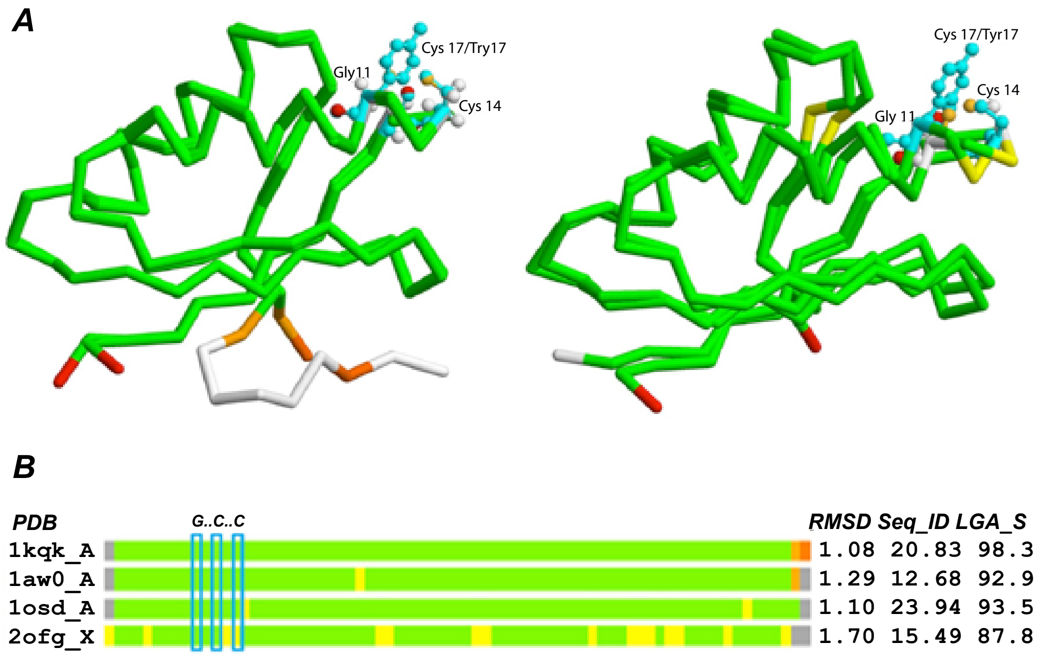
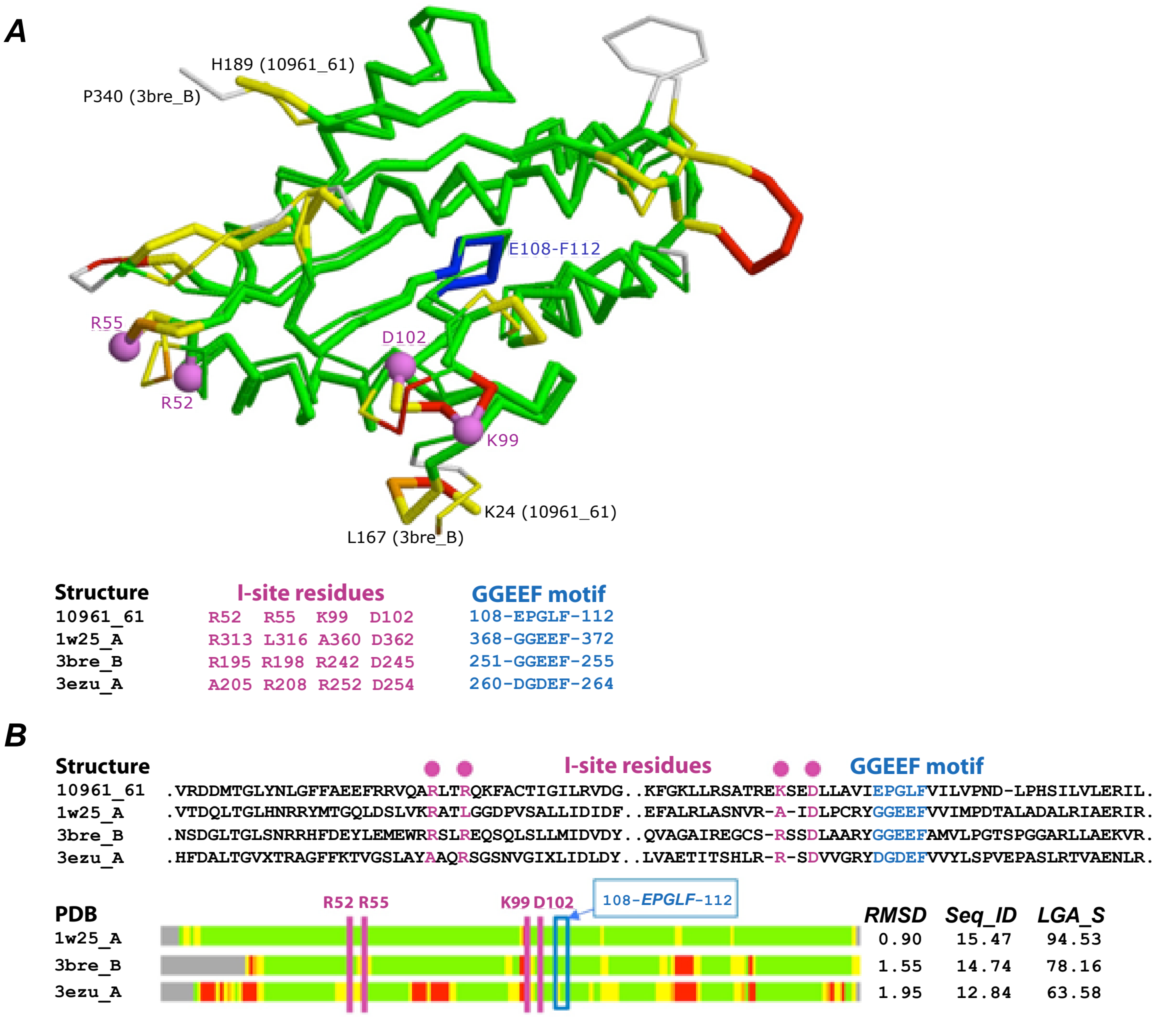


Figure 4: Structural superposition of the model for protein 11238_88 and coppertranslocating P-type ATPases (CopA). (A.) Left, structural comparison of the model with CopA from *Bacillus subtilis*, PDB: 1kqk; 1st bar in (B). Right, structural comparison of the model with MerP from *Cupriavidus metallidurans*, PDB: 1osd; 3rd bar in (B). Positions of the key amino acids: G, C, and Y are shown in cyan rectangles as observed in PDB templates and the model constructed. (B.) Bar representation of structural deviations between four CopA proteins using created model of 11238_88 as a frame of reference. The location of the GxxCxxC motif is highlighted in cyan. The colors of the bars indicate the distance deviation between superimposed corresponding residues using the following color scheme: deviation <2Å, green; <4Å, yellow; <6Å, orange; <8Å, brown; >8Å or not aligned, red; not aligned and terminal residues not aligned, grey.

Research Article
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000150
JPB/Vol.3 Issue 9



Structure	I-site residues	GGEEF motif
10961_61	R52 R55 K99 D102	108-EPGLF-112
1w25_A	R313 L316 A360 D362	368-GGEEF-372
3bre_B	R195 R198 R242 D245	251-GGEEF-255
3ezu_A	A205 R208 R252 D254	260-DGDEF-264

```
Structure
10961_61  .VRDDMTGLYNLGFFAEEFRRVQARLTRQKFACTIGILRVDG..KFGKLLRSATREKSEDLLAVIEPGLFVILVPND-LPHSILVLERIL.
1w25_A    .VTDQLTGLHNRRYMTGQLDSLVKRATLGGDPVSALLIDIDF..EFALRLASNVR-A-IDLPCRYGGEEFVVIMPDTALADALRIAERIR.
3bre_B    .NSDGLTGLSNRRHFDEYLEMEWRRSLREQSQLSLLMIDVDY..QVAGAIREGCS-RSSDLAARYGGEEFAMVLPGTSPGGARLLAEKVR.
3ezu_A    .HFDALTGVXTRAGFFKTVGSLAYAAQRSGSNVGIXLIDLDY..LVAETITSHLR-R-SDVVGRYDGDEFVVYLSPVEPASLRTVAENLR.
```

PDB	RMSD	Seq_ID	LGA_S
1w25_A	0.90	15.47	94.53
3bre_B	1.55	14.74	78.16
3ezu_A	1.95	12.84	63.58

Figure 5: Structural superpositions between protein 10961_61 and diguanylate cyclases. (A.) The structures of the model constructed for protein 10961_61 (backbone thickened) and diguanylate cyclase from *Pseudomonas aeruginosa* pao1 (backbone thinned, PDB: 3bre chain B) are superimposed and colored by the distance deviation of the corresponding C-alpha atoms (2nd bar in (B)). The 108-EPGLF-112 sequence from protein 10961_61, which corresponds to the GGEEF sequence motif of the active site from 3bre is colored in blue. Positions that correspond to selected residues from the allosteric inhibitory site (I-site) in 3bre (De et al. 2008) are indicated by violet spheres. (B.) Structure-based sequence alignment (top; fragment: 29-VRDD...ERIL-131), and bar representation of deviations in structural alignment (bottom) of protein 10961_61 with diguanylate cyclases from *Caulobacter vibrioides* (PDB: 1w25), *Pseudomonas aeruginosa pao1* (PDB: 3bre), and *Geobacter sulfurreducens* (PDB: 3ezu). Distance deviations are calculated using model of 10961_61 as a frame of reference. Distance deviations between superimposed corresponding residues are indicated using the same color scheme as in Figure 4B.

al., 2002). Compared with YceI proteins from *E. coli* and *Thermus thermophilus* (Handa et al., 2005), the *Leptospirillum* Group II YceI has a calculated pI of 9.9, over four pH units higher than its known counterparts. This large shift in pI is due to a higher number of arginines and lysines and is a common feature in other proteins of *Leptospirillum* Group II, noted above as a likely result of selection within the extremely acidic environment.

Insights into enzymes involved in central metabolism were also provided thruough structural predictions. As reported previously, genomic analysis indicates that *Leptospirillum* Group II has an incomplete TCA cycle, also known as a TCA horseshoe (Goltsman et al., 2009) that requires two adjacent isocitrate dehydrogenase genes, 11276_154 and 11276_155. Structure prediction showed that these genes are an example of a domain fusion protein. The analysis reported here indicated that both proteins were modeled upon different structural domains within the same template, *T. thermophilus* isocitrate dehydrogenase (PDB ID 2D1C, Figure 3A) (Lokanath and Kunishima, 2005). The predicted structure of 11276_155 overlaps the first ~350 amino acids at the N-terminus (top alignment in Figure 3B), while 11276_154 overlaps the final ~110 amino acids at the C-terminus (bottom alignment in Figure 3B). Twenty amino acid residues close to the N-terminal region of 11276_154 (red in Figure 3A) are not modeled as it did not align well to any region of the structural template. Interestingly, 11276_155 was found to contain the binding sites for both nicotinamide adenine dinucleotide and citric acid, while no clear functional role can yet be defined for 11276_154 (Miyazaki et al., 1994; Ohzeki et al., 1995; Steen et al., 2001). Nevertheless, structural data provided a suggestion of evolutionary linkage between 11276_154, 11276_155, and isocitrate dehydrogenases from other organisms (see Figure S1 for a phylogenetic tree).

Novel function predictions available by structure modeling and analysis

Part of our approach involved generating a network of local libraries of multiple sequence alignments and a database of intermediate structural models. Because of this, in several cases structure-based homology detection resulted in protein fold predictions and functional insights for proteins for which sequence analysis methods alone, such as PSI-BLAST (5 iterations), showed especially weak alignments (E values ≥ 0.1).

One such example is for the protein encoded by gene 11238_88. The best template for this structure was a copper translocating P-type ATPase (CopA) (Boal and Rosenzweig 2009) from *Bacillus subtilis.* Alignment of the modeled structure for gene 11238_88 and the N-terminal region of the CopA protein from *B. subtilis* resulted in a high quality category B model (Figure 4). The Cu(I) binding region, with a N-terminal conserved sequence GxxCxxC motif, is well conserved in all CopA proteins (Boal and Rosenzweig 2009). Alignment of CopA proteins with gene 11238_88 suggested a slightly modified motif of GxxCxxY, which would result in copper ligation via a cysteine and tyrosine. Experimental testing is necessary to confirm copper ligation. Although it is not a favored residue for copper ligation, tyrosine can be the ligand in some previously identified proteins, such as amine oxidase, galactose oxidase, and when copper is (mis)incorporated into the iron transport protein transferrin (Fontecave and Eklund, 1995). Close homologs to CopA were found in *Enterococcus hirae*, *Helicobacter pylori*, *E. coli* and*Synechococcus* (Figure 4). Also, CopA can catalyze copper extrusion in *E. coli* (Rensing et al., 2000). Based on these several lines of evidence, we predicted that 11238_88 has a copper export function, which would be an important if not essential function in the AMD environment where copper and other heavy metals are abundant.

The predicted structure for the protein encoded by gene 10961_61, a C1 quality model, aligned well to the SCOP superfamily of diguanylate cyclases (Figure 5). Indeed, the best structural templates are signalling proteins from *Caulobacter vibrioides* (PDB 1w25) and *Pseudomonas aeruginosa* (PDB 3BRE) with a diguanylate cyclase SCOP Superfamily domain. Although prokaryotes generally do not use cGMP for signalling, c-diGMP has been shown to regulate cell surface-associated traits and community behavior such as biofilm formation in a number of bacterial species (Chan et al., 2004). Further experiments are necessary to verify the role of 10961_61 in biofilm formation.

Conclusions

In this study a collection of 545 ORFan proteins produced by an extreme niche-adapted microbial community were selected for *in silico* structural analysis. These proteins represented a dataset for which sequence analysis tools provided low confidence or no insights for functional annotation. Homology modeling was performed, resulting in high confidence structural models for 125 proteins. The structural models were compared to known functional domains to provide additional confidence in the models and potential SCOP Superfamily classification. General hypotheses for function were assigned via the SUPERFAMILY database based upon SCOP Superfamily classifications, and potential GO functions were assessed for a small subset. This analysis, in combination with previously published proteomic data and physicochemical characterizations, provided a database from which hypotheses were drawn about the roles of these unusual proteins within the extremophilic microbial community. This approach will be useful for future experimental structural elucidation and experimentally derived functional assessment.

Additional data files

A comprehensive spreadsheet containing integrated data on each of the 545 expressed ORFan proteins is given in Table S1. Lists of proteins highlighted here, along with associated *in silico* data, are extracted from Table S1 and presented in Table S1a, including data on all category A models, and proteins with the following SCOP Superfamily domains: lambda repressor-like DNA binding domains, restriction endonuclease-like domains, c-type cytochromes, and thioredoxin-like domains. SCOP superfamilies for each model, along with designated functions, can be found in Table S2. Additionally, detailed results from AS2TS homology modeling are available at: http://proteinmodel.org/AS2TS/research/M_Thelen/FUN_545/.

Acknowledgements

Funding was provided by the U. S. Department of Energy, Office of Science, from the Genome Sciences Program, and work at LLNL was performed under the auspices of the DoE, contract DE-AC52-07NA27344. We thank Ryan Mueller at UC Berkeley for insightful discussions on the various protein and protein structures within this dataset and their possible biological implications.

References

1. Adams MA, Suits MD, Zheng J, Jia Z (2007) Piecing together the structure-function puzzle: experiences in structure-based functional annotation of hypothetical proteins. Proteomics 7: 2920-2932.
2. Anisimov AP, Dentovskaya SV, Panfertsev EA, Svetoch TE, Kopylov PKh, et al. (2010) Amino acid and structural variability of Yersinia pestis LcrV protein. Infect Genet Evol 10: 137-45.
3. 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.
4. Baker D, Sali A (2001) Protein Structure Prediction and Structural Genomics. Science 294: 93-96.
5. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved Prediction of Signal Peptides: SignalP 3.0. JMol Biol 340: 783-795.
6. Boal AK, Rosenzweig AC (2009) Structural Biology of Copper Trafficking. Chem Rev 109: 4760-4779.
7. 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.
8. Bower MJ, Cohen FE, Dunbrack RL Jr (1997) Prediction of protein side-chain rotamers from a backbonedependent rotamer library: a new homology modeling tool. J Mol Biol 267: 1268-1282.
9. Brenner SE, Koehl P, Levitt M (2000) The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res 28: 254-256.
10. Chakicherla A, Ecale Zhou CL, Dang ML, Rodriguez V, Hansen JN, et al. (2009) SpaK/SpaR two-component system characterized by a structure-driven domain-fusion method and in vitro phosphorylation studies. PLoS Comput Biol 5: e1000401.
11. Chan C, Paul R, Samoray D, Amiot NC, Giese B, et al. (2004) Structural basis of activity and allosteric control of diguanylate cyclase. Proc Natl Acad Sci USA 101: 17084-17089.
12. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, et al. (2004) The ASTRAL Compendium in 2004. Nucl Acids Res 32: D189-92.
13. Cosman M, Pesavento JB, Zemla A, Beernink PT, Balhorn R, et al. (2008) Identification of a thermo-regulated glutamine-binding protein from Yersinia pestis. Protein Pept Lett 15: 887-894.
14. Cunane LM, Chen ZW, Shamala N, Mathews FS, Cronin CN, et al. (2000) Structures of the flavocytochrome p-cresol methylhydroxylase and its enzyme-substrate complex: gated substrate entry and proton relays support the proposed catalytic mechanism. J Mol Biol 295: 357-374.
15. De N, Pirruccello M, Krasteva PV, Bae N, Raghavan RV, et al. (2008) Phosphorylation-independent regulation of the diguanylate cyclase WspR. PLoS Biol 6: e67.
16. Fischer D, Eisenberg D (1999) Finding families for genomic ORFans. Bioinformatics 15 : 759-762.

Research Article
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000150
JPB/Vol.3 Issue 9

17. Fontecave M, Eklund H (1995) Copper amine oxidase: a novel use for a tyrosine. Structure 3: 1127-1129.

18. Foster SJ (1991) Cloning, expression, sequence analysis and biochemical characterization of an autolytic amidase of Bacillus subtilis 168 trpC2. J Gen Microbiol 137: 1987-98.

19. Galvagnion C, Smith MT, Broom A, Vassall KA, Meglei G, et al. (2009). Folding and Association of Thermophilic Dimeric and Trimeric DsrEFH Proteins: Tm0979 and Mth1491. Biochemistry 48: 2891-906.

20. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, et al. (2005) Protein Identification and Analysis Tools on the ExPASy Server. (In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press.

21. Ginalski K, Grishin NV, Godzik A, Rychlewski L (2005) Practical lessons from protein structure prediction. Nucl Acids Res 33: 1874-1891.

22. Goltsman DS, Denef VJ, Singer SW, VerBerkmoes NC, Lefsrud M, et al. (2009) Community genomic and proteomic analyses of chemoautotrophic iron-oxidizing Leptospirillum rubarum (Group II) and Leptospirillum ferrodiazotrophum (Group III) bacteria in acid mine drainage biofilms. Appl Environ Microbiol 75: 4599-4615.

23. 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.

24. Handa N, Terada T, Doi-Katayama Y, Hirota H, Tame JR, et al. (2005) Crystal structure of a novel polyisoprenoid-binding protein from Thermus thermophilus HB8. Protein Sci 14: 1004-1010.

25. Harrington ED, Singh AH, Doerks T, Letunic I, von Mering C, et al. (2007) Quantitative assessment of protein function prediction from metagenomics shotgun sequences. Proc Natl Acad Sci USA 104: 13913-13918.

26. Holland TA, Veretnik S, Shindyalov IN, Bourne PE (2006) Partitioning Protein Structures into Domains: Why Is it so Difficult? J Mol Biol 361: 562-590.

27. Hu P, Janga SC, Babu M, Díaz-Mejía JJ, Butland G, et al. (2009) Global Functional Atlas of Escherichia coli Encompassing Previously Uncharacterized Proteins. PLoS Biol 7 : e96.

28. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, et al. (2009) InterPro: the integrative protein signature database. Nucl Acids Res 37: D211-5.

29. Huynen M, Doerks T, Eisenhaber F, Orengo C, Sunyaev S, et al. (1998) Homology-based fold predictions for Mycoplasma genitalium proteins. J Mol Biol 280: 323-326.

30. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292: 195-202.

31. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK - a program to check the stereochemical quality of protein structures. J App Cryst 26: 283-291.

32. Lokanath NK, Kunishima N (2005) Crystal Structure Of TT0538 protein from Thermus thermophilus HB8.To Be Published.

33. Malmström L, Riffle M, Strauss CE, 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: e76.

34. Miyazaki K, Yaoi T, Oshima T (1994) Expression, purification, and substrate specificity of isocitrate dehydrogenase from Thermus thermophilus HB8. Eur J Biochem 221: 899-903.

35. Moult J, Fidelis K, Kryshtafovych A, Rost B, Tramontano A (2009) Critical assessment of methods of protein structure prediction - Round VIII. Proteins 77: 1-4.

36. Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, et al. (2007) Critical assessment of methods of protein structure prediction-Round VII. Proteins 69: 3-9.

37. 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: 536-40.

38. Ohzeki M, Yaoi T, Moriyama H, Oshima T, Tanaka N (1995) Crystallization and Preliminary X-Ray Studies of Isocitrate Dehydrogenase from Thermus thermophilus HB8. J Biochem 118: 679-80.

39. Pott AS, Dahl C (1998) Sirohaem sulfite reductase and other proteins encoded by genes at the dsr locus of Chromatium vinosum are involved in the oxidation of intracellular sulfur. Microbiology 144: 1881-94.

40. Ram RJ, Verberkmoes NC, Thelen MP, Tyson GW, Baker BJ, et al. (2005) Community proteomics of a natural microbial biofilm. Science 308: 1915-20.

41. Rensing C, Fan B, Sharma R, Mitra B, Rosen BP (2000) CopA: An Escherichia coli Cu(I)-translocating P-type ATPase. Proc Natl Acad Sci USA 97: 652-656.

42. Rychlewski L, Zhang B, Godzik A (1998) Fold and function predictions for Mycoplasma genitalium proteins. Fold Des 3: 229-38.

43. Sánchez R, Sali A (1998) Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. Proc Natl Acad Sci USA 95: 13597- 602.

44. Singer SW, Chan CS, Zemla A, VerBerkmoes NC, Hwang M, et al. (2008) Characterization of cytochrome 579, an unusual cytochrome isolated from an iron-oxidizing microbial community. Appl Environ Microbiol 74: 4454-4462.

45. Stancik LM, Stancik DM, Schmidt B, Barnhart DM, Yoncheva YN, et al. (2002). pH-Dependent Expression of Periplasmic Proteins and Amino Acid Catabolism in Escherichia coli. J Bacteriol 184: 4246-4258.

46. Steen IH, Madern D, Karlström M, Lien T, Ladenstein R, et al. (2001) Comparison of Isocitrate Dehydrogenase from Three Hyperthermophiles Reveals Differences in Thermostability, Cofactor Specificity, Oligomeric State, and Phylogenetic Affiliation. J Biol Chem 276: 43924-43931.

47. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428: 37- 43.

48. Venclovas C, Zemla A, Fidelis K, Moult J (2003) Assessment of progress over the CASP experiments. Proteins 53: 585-595.

49. Vogel C, Berzuini C, Bashton M, Gough J, Teichmann SA (2004) Supra-domains: evolutionary units larger than single protein domains. J Mol Biol 336: 809-23.

50. Vogel C, Chothia C (2006) Protein family expansions and biological complexity. PLoS Comput Bio l 2: e48.

51. Zemla A (2003) LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res 31: 3370-3374.

52. Zemla A, Zhou CE, Slezak T, Kuczmarski T, Rama D, et al. (2005) AS2TS system for protein structure modeling and analysis. Nucleic Acids Res 33: W111-5.

53. Zemla A, Geisbrecht B, Smith J, Lam M, Kirkpatrick B, et al. (2007) STRALCP structure alignment-based clustering of proteins. Nucleic Acids Res 35: e150.

54. Zemla AT, Ecale Zhou CL (2008) Structural re-alignment in an immunogenic surface region of ricin A chain. Bioinform Biol Insights 2: 5-13.

55. Zhang Y, Skolnick J (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci USA 101: 7594-7599.

56. Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, et al. (2006) Statistical Analysis of Membrane Proteome Expression Changes in Saccharomyces cerevisiae. J Proteome Res 5: 2339-47.