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.

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

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 diguanosylate 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.

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...
specific proteins within this extremely acidic, metal rich environment. 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 or structural annotations. 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 345 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 MS proteomic 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 (Ginalska 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 insertion/deletion 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 N-terminus; 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: e-value, 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, the 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
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://proteinform.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://proteinform.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 pl for the entire proteome of approximately one pH unit higher than common neutrophilic microbes (Ram et al., 2005). The high average pl 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 pl values (Ram et al., 2005), we found that the quality of structural modeling was independent of pl (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.
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
Biochemistry of the AMD community.

Their sequence-based annotation as “conserved” or “unique proteins of unknown or simply because the SCOP database is not as current as the PDB. Superfamily, either because of the low quality of the models created, was inadequate to provide a domain-based function; in large part, annotations (Table 3). In the final 15%, the structure based approach extracted from structural modeling and SCOP Superfamily (Goltsman et al., 2009) and compared to functional information small set of weakly annotated proteins were included in our dataset functional insights, sequence-based functional assignments for a identified SCOP Superfamilies, 75% were C1 or C2 quality. These low higher propensity for false positives. Of the models with five or more matches to a SCOP superfamily were obtained for 35%, or 167 protein functions

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 sulfoxide 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 quinone 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) 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 AMD environment.

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 sulfoxide isomerases to aid in protein folding (Pott and Dahl, 1998).
Figure 3: Domain fusion model created for proteins 11276_155 and 11276_154, compared with \textit{T. thermophilus} isocitrate dehydrogenase. (A) Structure of \textit{T. thermophilus} isocitrate dehydrogenase (PDB 2d1c) is colored by its correspondence to the \textit{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 copper-translocating 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 created. (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 not aligned, red; not aligned and terminal residues not aligned, grey.
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 through 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), and Geobacter *sulfurreducens* (PDB: 3ezu), and *Geobacter sulfurreducens* (PDB: 3ezu) (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 *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.
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 C(terminal) 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 Synecococcus (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 (PDB 1w25) and Pseudomonas aeruginosa (PDB 3BRE) with a diguanilate cyclase SCOP Superfamily domain. Although prokaryotes generally do not use cGMP for signalling, c-diGMP has been shown to regulate cyclase SCOP Superfamily domain. 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 10961_61, a C1 quality model, aligned well to the SCOP superfamily domain. Although prokaryotes generally do not use cGMP for signalling, c-diGMP has been shown to regulate cyclase SCOP Superfamily domain. However, in prokaryotes, 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 in a number of bacterial species (Chan et al., 2004).

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://proteinform platform. J Proteomics Bioinform 15: 759-762.
17. Fontecave M, Elkind 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 acylase amide of Bacillus subtilis 168 trpC2. J Gen Microbiol 137: 1987-96.

19. Galvagnion C, Smith MT, Broom A, Vassall KA, Meglej 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. Ginalska K, Grishin NV, Godzik A, Rychlewski L (2005) Practical lessons from protein structure prediction. Nuc 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 chemoaerotrophic iron-oxidizing Leptospirillum rubarum (Group II) and Leptospirillum ferrooxidotrophum (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. Hanada 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. (2004) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Nucleic Acids Res 32: 2339-2340.

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, Diaz-Mejia JJ, Butland G, et al. (2009) Global Protein Identi fi... Community genomic and proteomic analyses of chemoautotrophic iron-oxidizing microbial community. Appl Environ Microbiol 75: 43924-43931.

28. Walker (ed): The Proteomics Protocols Handbook, Humana Press.

29. Venclovas C, Zemla A, Fidelis K, Moult J (2003) Assessment of progress over the CASP experiments. Proteins 53: W111-5.

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

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

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

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

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

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

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

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

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

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

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

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