OMICS A Journal of Integrative Biology
Volume 7, Number 2, 2003

A Computational Study of *Shewanella oneidensis* MR-1: Structural Prediction and Functional Inference of Hypothetical Proteins

CHRISTAL YOST,[1,2] LOREN HAUSER,[1,2] FRANK LARIMER,[1,2] DOROTHEA THOMPSON,[3] ALEXANDER BELIAEV,[3] JIZHONG ZHOU,[3] YING XU,[1,2,4] and DONG XU[1,2]

ABSTRACT

The genomes of many organisms have been sequenced in the last 5 years. Typically about 30% of predicted genes from a newly sequenced genome cannot be given functional assignments using sequence comparison methods. In these situations three-dimensional structural predictions combined with a suite of computational tools can suggest possible functions for these hypothetical proteins. Suggesting functions may allow better interpretation of experimental data (e.g., microarray data and mass spectroscopy data) and help experimentalists design new experiments. In this paper, we focus on three hypothetical proteins of *Shewanella oneidensis* MR-1 that are potentially related to iron transport/metabolism based on microarray experiments. The threading program PROSPECT was used for protein structural predictions and functional annotation, in conjunction with literature search and other computational tools. Computational tools were used to perform transmembrane domain predictions, coiled coil predictions, signal peptide predictions, sub-cellular localization predictions, motif prediction, and operon structure evaluations. Combined computational results from all tools were used to predict roles for the hypothetical proteins. This method, which uses a suite of computational tools that are freely available to academic users, can be used to annotate hypothetical proteins in general.

INTRODUCTION

As more and more genomes have been sequenced, it is becoming evident that computational predictions of protein structure and function play an essential role in understanding the genes coded in these genomes. Predicting the three dimensional structure of a protein from its amino acid sequence is not only possible, but also it has reached such an accuracy level that detailed functional information can be derived, as demonstrated in the CASP contests (Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction) (CASP, 1993, 1995, 1997, 2001). Sequence comparison tools such as PSI-BLAST (Altschul et al., 1997) using multiple sequence profiles and SAM (Karplus et al., 1998) using hidden Markov models (HMM) are able to identify very weak sequence homology signals. Protein

The [1]Life Sciences Division, [3]Environmental Sciences Division, and [4]Computer Sciences and Mathematics Division, Oak Ridge National Laboratory (ORNL), and [2]UT-ORNL Graduate School of Genome Science and Technology, Oak Ridge, Tennessee.

threading or protein fold recognition can identify even more remote homologs through sequence-structure comparison. In particular, the threading program PROSPECT (Xu and Xu, 2000) calculates the secondary structure matches and pair-wise interactions between residues in three dimensions; and it can detect homologs that are undetectable using sequence comparison techniques alone. Given the good prediction accuracy of structure and the structure-function relationship, it has been suggested that three-dimensional structures predicted by protein threading may be used to help predict the functions of proteins (Brenner, 2001; Skolnick et al., 2000; Xu et al., 2002). The technique of predicting function from sequence, through structural prediction, can be used both to study hypothetical proteins from a newly sequenced genome and to evaluate possible functions of known proteins being studied in experimental labs.

We have applied computational methods to predict the structures and functions of hypothetical genes from *Shewanella oneidensis* MR-1 (formerly *Shewanella putrefaciens* MR-1). *S. oneidensis* MR-1 is a gram negative, facultatively anaerobic, dissimilatory metal-reducing bacterium. It is found in soil, sediments, and ground water. During anaerobic growth, *S. oneidensis* MR-1 can use a number of terminal electron acceptors. These include iron, manganese, nitrate, nitrite, fumarate, thiosulfate, dimethylsulfoxide, trimethylamine-N-oxide, and elemental sulfur (Saffarini et al., 1994; Venkateswaran et al., 1999). *S. oneidensis* MR-1 is a candidate for bioremediation because it links carbon cycling (oxidation) to metal reduction, thus dissolving iron containing (insoluble) minerals while using toxic halogenated organic pollutants as a carbon source (Myers and Nealson, 1990; Nealson and Saffarini, 1994; Pessanha et al., 2001). Since the use of *S. oneidensis* MR-1 could produce unexpected environmental changes, it is important to understand the proteins and biochemical pathways involved in iron metabolism.

The genome of *S. oneidensis* MR-1 was recently sequenced by TIGR (http://www.tigr.org; Heidelberg et al., 2002). Computational tools were used to identify and functionally annotate the *Shewanella* genome. These annotations were available on-line as the *Shewanella* annotation database (Larimer et al., 2002). The *S. oneidensis* MR-1 genome appears to have approximately 4500 genes. About 60% of the genes in the *Shewanella* annotation database have functional annotations (Larimer et al., 2002), and the remaining 40% of the predicted genes had no sequence similarities to genes of known functions and were labeled as hypothetical genes. It would take a tremendous amount of time and effort for experimentalists to verify the functions of all the hypothetical genes identified in *S. oneidensis* MR-1. On the other hand, computational methods, other than sequence based annotation, may provide functional predictions (Brenner, 2001; Fetro et al., 2001; Xu et al.).

In this study, we focused on three hypothetical proteins that were identified to be related to the iron uptake metabolism in a fur mutant, using high-throughput experimental data (microarray data). Ferric uptake regulator protein, *fur*, has been identified primarily as a transcriptional regulator that suppresses or increases expression of proteins related to iron metabolism in *S. oneidensis* MR-1 (Hantke, 2001). When iron levels are high enough that diffusion provides the bacteria with sufficient amounts of iron, *fur* bound with Fe^{2+} negatively regulates genes related to iron uptake metabolism. In a *fur* mutant these genes are expected to be up regulated. F*ur* is also implicated in positive regulation of genes like super-oxide dismutase (SOD), which helps decrease the negative effects of free radicals that are present when iron levels are high. Since there is no expression of the *fur* protein in the mutant, genes like SOD, are expected to show lower expression levels (Braun, 2001; Dubrac et al., 2000; Escolar et al., 1999). A *fur* knockout strain (FUR1) of *S. oneidensis* MR-1 was generated by suicide plasmid integration into the gene, and the knockout was characterized using phenotype assays, DNA microarrays, and two-dimensional polyacrylamide gel electrophoresis (Thompson et al., 2001). While several of the differentially expressed genes from this *fur* mutant have known functions, three hypothetical genes (with unknown biochemical functions) with notable changes in expression pattern were identified. We applied protein structure prediction, in conjunction with other computational method to functionally annotate these three hypothetical proteins.

MATERIALS AND METHODS

We applied the same computational protocol to study each of the three hypothetical proteins, which are named gene 4840, gene 4719 and gene 3954. The protocol includes the following 12 steps.

Database search

Table 1 summarizes the information from database searches. All of the three proteins are annotated as "conserved hypothetical protein" in the current NCBI database.

Secondary structure prediction

The amino acid sequences of each of the three proteins were used to predict the secondary structures using PHD (Rost et al., 1993) and PSIPRED (McGuffin et al., 2000). The average accuracy of secondary structure prediction using PSIPRED is around 80%. The secondary structural predictions from PHD were used as inputs for PROSPECT.

Threading and modeling

The primary program used for threading was PROSPECT (Xu et al., 2000), PROtein Structure Prediction and Evaluation Computer Toolkit. PROSPECT recognizes similar structural folds between the query and templates. After all template candidates for the threading result had been returned by PROSPECT, a final evaluation based on compactness and neural network assessment of threading reliability was carried out (Xu et al., 2002). PROSPECT threading results return output templates that can be visualized using RasMol (Sayle et al., 1995) and CHIME (MDL, 1999). The results can be made available through a web based user interface generated by PROSPECT. Top templates chosen from PROSPECT were compared to top hits from other threading programs such as GenTHREADER (Jones, 1999) and the fold recognition program SAM-T98 (Karplus et al., 1998). If a fold is present in all three programs with high rank, then the confidence of the query protein adopting the fold becomes very high. Once the best threading template is chosen, the three dimensional structural prediction for the query protein can be generated using MODELLER (Sali et al., 1993).

Literature search

Given that the query proteins were all hypothetical, no information was available for them in the literature. However, information about the template proteins could be found in the literature. Once the threading template was chosen, literature and other database searches were used to provide information such as possible motifs (signature sequences), adjacent genes, subcellular location, and other information about the template protein. This information was used to direct the searches during the successive steps.

Sequence comparison using PSI-BLAST

Since the three proteins had been annotated as hypothetical proteins, BLAST (Altschul et al., 1997) found no new functional information. We therefore used PSI-BLAST (Altschul et al., 1997), which is more sensitive than BLAST, to search for remote homologs which may have very little sequence similarity. Five iterations were run for each query protein using PSI-BLAST. Literature searches were done on related PSI-BLAST hits. Results from PSI-BLAST searches sometimes include links to the SWISS-PROT database (Bairoch and Apweiler, 1999), and could be reviewed for possible functional annotations.

Table 1. Summary of the Three Hypothetical Proteins

ORF	*NCBI locus*	*TIGR locus*	*Length (a.a.)*	*NCBI annotation*
4840	NP_716997	SO1377	592	Conserved hypothetical protein
4719	NP_716815	SO1190	272	Conserved hypothetical protein
3954	NP_720235	SO4719	286	Conserved hypothetical protein

The ORF numbers that we used in this paper were based on the *Shewanella* annotation database (Larimer et al., 2002), and they are different from the TIGR Loci. The NCBI annotations of the proteins were all still hypothetical as of April 2003.

Domain parsing

When there is no confident template for the entire query protein sequence, it is advantageous to divide the protein into domains and submit those partial sequences to the threading program. The automated web tool ProDom (Corpet et al., 2000) was used to identify possible domain divisions within the protein sequences.

Motif search

We searched for motifs (signature sequences) in the query amino acid sequences using the web-based tool MOTIF (www.motif.genome.ad.jp), which uses the databases of ProSite (Atwood et al., 1999), BLOCKS (Henikoff, 1991), ProDom (Corpet et al., 2000), Pfam (Bateman et al., 2002), and PRINTS (Attwood et al., 2002). Motifs often reveal themselves in functional domains of proteins. Motifs typically consist of perfectly spaced highly conserved amino acids, surrounded by a group of amino acids that have similar physical characteristics (hydrophobicity, size, and acidity) to those of other similar related proteins. It is usually assumed that these more highly conserved groups of amino acids remain constant over evolutionary time because they are involved in maintenance of the function or structure of the protein. Domains belonging to a particular family generally contain similar signature sequences and functional attributes. GeneFIND (Wu et al., 1996) was used to search for sequence motifs and family classification. GeneFIND provides an on-line interface that allows direct submission of the query sequence and returns a gene family classification of query sequences. GeneFIND uses the ProClass (Atwood et al., 1999) database with ProSite predictions.

Trans-membrane domain search

SOSUI (Hirokawa et al., 1998) allows the prediction of hydrophobic regions of the protein that may either be trans-membrane domains (TMD) or signal sequence residues. The results from this tool include an average hydrophobicity, a prediction of the number of TMDs and their locations, a hydrophobicity profile, and helical wheel diagrams for predicted trans-membrane segments.

Signal sequence prediction

SignalP (Henrik et al., 1997) was used to predict the presence, location and probable cleavage site of signal sequences that might be present in the query. Once hydrophobic helical regions are identified in the query it is important to attempt to determine the likelihood as to whether those regions act as a signal sequence, or are inserted into the membrane.

Subcellular localization prediction

PSORT (Nakai, 2000) is a computer program for the prediction of protein localization sites in cells based on the protein sequence.

Genomic sequence analysis

The final step in the protocol was genomic sequence analysis using the *Shewanella* annotation database (Larimer et al., 2002) with an Artemis interface (Rutherford et al., 2000), TransTerm (Ermolaeva et al., 2000) and the Genome Channel (http://compbio.ornl.gov/channel/). The orientation and proximity of genes surrounding each hypothetical was visualized using Artemis. When one promoter region begins transcription and more than one gene is transcribed on the same mRNA transcript, those genes are said to be on the same operon. Operon structure was studied in order that gene function could be better understood. Genes that are transcribed on the same operon could be involved in the same pathway or they could be components of the same complex. Often when genes are in an operon, open reading frames either overlap or are less than 50 base pairs apart. If the gene of interest was less than 100 bases from adjacent genes, the possibility of the genes being on the same operon was also considered.

When adjacent genes appeared to be in an operon with the gene of interest, affiliated genes were annotated as well. This can be advantageous since knowing the function of some of the genes in an operon can

be informative about the functions of surrounding genes. Promoter regions were inspected for TATA boxes and *fur*-binding sites. Trailing intergenic regions were inspected for rho-independent hard transcription termination sequences. The program TransTerm was used to identify the location of all rho-independent termination sequences for the entire genome. If following a gene there is an interrupted GC-rich region followed by a run of T's, this is indicative of a rho-independent hard transcription stop site. The pallindromic GC rich region would form a stem, the RNA polymerase would be kicked off. The presence of these transcription stop sites suggests the end of an operon.

Other tools

A number of other tools may be used, such as COILS (Lupas et al., 1991) for coiled-coil prediction. More information is available at http://compbio.ornl.gov/structure/resource/.

RESULTS AND CONCLUSIONS

Gene 4840

Gene 4840 was down-regulated 1.6-fold in the *fur* mutant, FUR1. The results of SOSUI suggest that the protein has one trans-membrane helix from residue 9 to 31. When there is only one hydrophobic region and it is near the N terminus it should be evaluated to determine if it could be a signal sequence. SignalP results suggest there is a signal peptide between residues 1 and 33 with the most likely cleavage site between residues 33 and 34; however, the probability of cleavage is low. This indicates that the query 4840 may have a signal that is inserted into the cytoplasmic membrane.

The best template found in the threading results using PROSPECT is 1dg3a. The alignment for the template 1dg3a encompassed almost the entire length of the query protein, from residue 50 to 490. The alignment yields a compact protein structure with good reliability for prediction accuracy (90% confidence level), and the aligned secondary structures on the template are in good agreement with the predicted secondary structures on the query protein. 1dg3a was also returned as a favourable template by the programs GenTHREADER and SAM. The consensus in the three computational tools further increases the confidence level of the prediction. Hence, 1dg3a was selected as the final template for hypothetical protein 4840, and a three dimensional model (Fig. 1) was derived using MODELLER. The N-terminal trans-membrane segment, which is expected as an anchor on the membrane, was removed before threading.

From the literature search, 1dg3a is an interferon-induced guanylate-binding protein, and it is a member of the dynamin family of proteins (Danino et al., 2001). The dynamin family proteins are known as mechanoenzymes. The proteins of this family are large GTPases. They are involved in fundamental cellular processes. Dynamin proteins contain five domains. These domains include a GTPase, two self-assembly re-

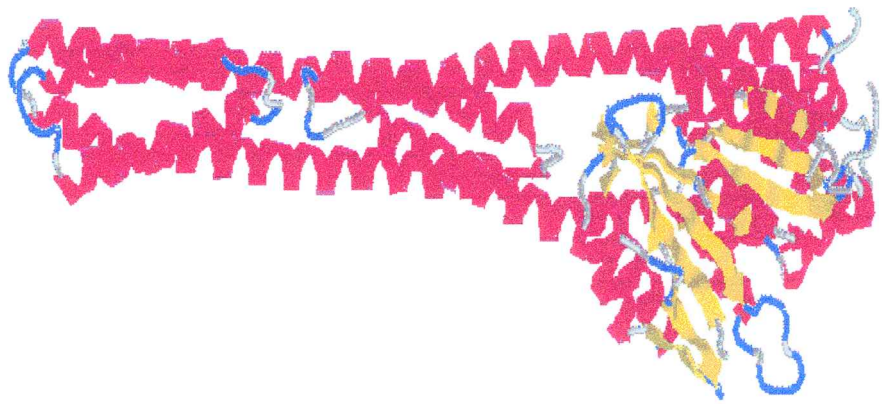

FIG. 1. Three-dimensional model for hypothetical protein 4840.

gions, a pleckstrin homology domain for membrane binding, a coiled-coil (GTPase effector) domain, and an arginine, proline rich domain. Known functions of 1dg3a (from PDB) include GTP binding, chaperone and cell signalling. 1dg3a contains a Bag-1 nucleotide exchange factor (Rothman et al., 1990), which increases interaction between the ATPase and the chaperone coiled-coil domains. The Bag domain contains a highly conserved sequence of residues (Sondermann et al., 2001). We identified a Bag conserved sequence pattern (Fig. 2) between the ATP/GTPase and coiled-coil domains on the query protein (hypothetical protein 4840) as well. This strongly increases the confidence of our structure prediction, and further suggests that interaction between the two domains ATP/GTPase and coiled-coil may occur in the hypothetical protein 4840 as well.

Using GeneFIND, a section of the query protein sequence was identified as a ProSite (*PS01036*) signature pattern related to ATP binding in DNAK. DNAK is the primary heat shock protein found in *E. coli*, which binds ATP (ProSite: PDOC00269). Functions of these proteins include controlling folding of other proteins (chaperones), assembly of complexes, and translocation of polypeptides across membranes. Heat Shock proteins contain an ATPase domain similar to the template protein 1dg3a.

The results from PSORT search for subcellular localization prediction include the statement that the protein "seems to have an un-cleavable N-term signal sequence." According to the PSORT Discriminant Score, the sequence is a signal which most likely remains and may act as a single trans-membrane anchor. The subcellular localization of this protein is predicted by PSORT to be the cytoplasmic side of the bacterial inner membrane. These results are also consistent with the possibility that the protein is either in the dynamin family or that it is a chaperone.

The secondary structural prediction showed a high percentage of helix from residue 200 to 480. The large unbroken stretches of helix indicate the possibility of a coiled coil. Based on this observation the web-based tool COILS was run. Results predicted the coiled-coil region from amino acid 200 to amino acid 480 with high confidence (Fig. 3).

The *Shewanella* annotation database and Artemis were used for genomic sequence analysis to assess whether the gene 4840 was possibly on the same operon as other identified genes. The gene identified in the *Shewanella* annotation database as 4840 was separated from the gene upstream 4839 by only 30 residues. This close proximity makes it unlikely that there is an individual promoter for the gene 4840. The intergenic space upstream of gene 4839 is large enough that there is no likely association with upstream genes. Three hundred base pairs of the intergenic region upstream of hypothetical gene 4839 were inspected for possible *fur* binding sites. None were found. The intergenic space downstream was evaluated for transcription stop sites. Using TransTerm it was verified that the region contains a sequence predicted to be a rho-independent transcription stop site. This suggests that the genes downstream of 4840 are not a part of the same operon. It appears that genes 4840 and 4839 are on the same operon. The hypothetical gene 4839 is a probable oxidoreductase. This makes no suggestion for the possible function of 4840. Gene 4840 could assist folding and insertion of the protein product of 4839 or associated proteins.

In summary, the hypothetical protein 4840 can be divided into four domains based on the findings from computational searches (Fig. 4). All the evidence suggests that the hypothetical protein may have homology to the Dynamin family of proteins and or Heat shock proteins. This protein may be involved in protein folding (chaperone), signalling, GTP/ATP binding, and or protein translocation in the iron transport/metabolism pathways.

FIG. 2. The sequence of the hypothetical protein 4840 that falls between the ATPase/GTPase domain and the coiled coil domain has been superimposed on the alignment of the Bag proteins (Sondermann et al., 2001), indicating the conserved signature is present in the hypothetical protein 4840.

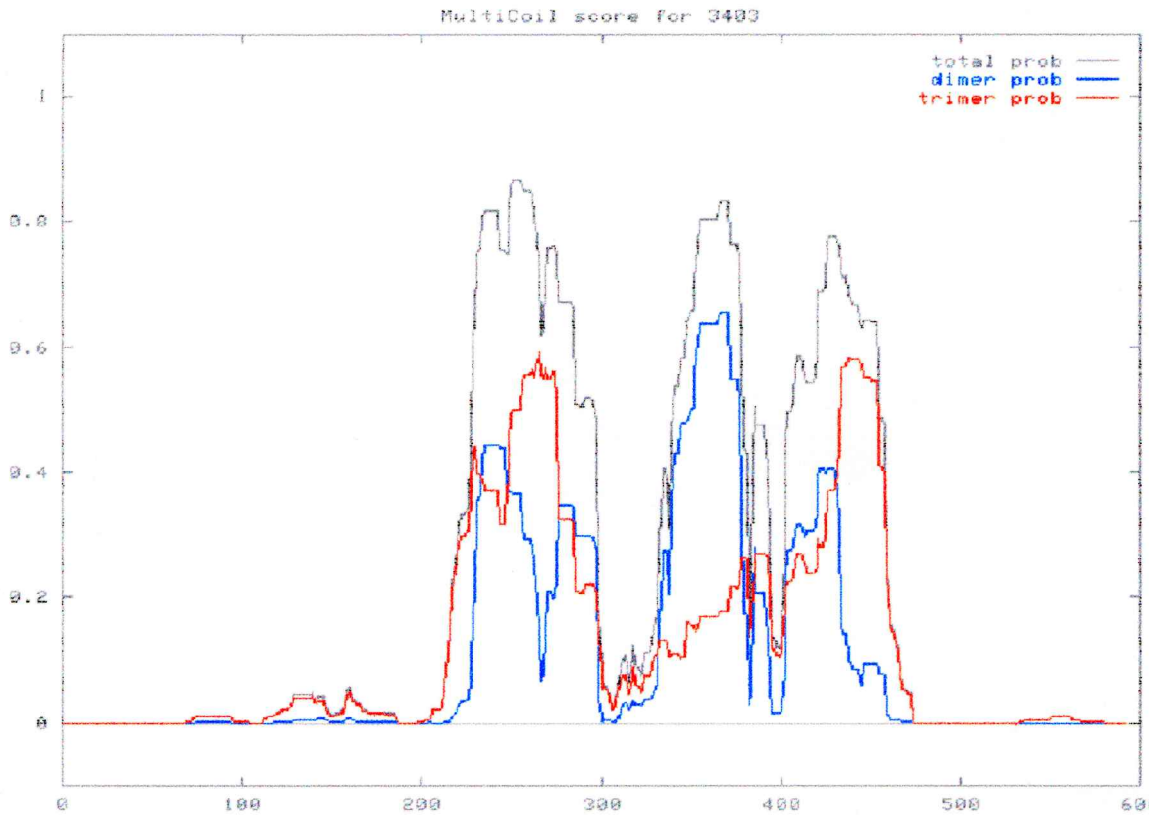

FIG. 3. The graphical representation from the web tool COILS (Lupas et al., 1991) showing the coiled coil region from residue 200 to about 480. The horizontal axis indicates the residue number; the vertical axis shows the probability of coiled coil.

Gene 4719

Gene 4719 was up-regulated approximately three fold in the *fur* mutant. SOSUI predicted the protein to be a soluble protein. The hydrophobicity plot indicates that the N terminal signal sequence is very hydrophobic and may be a signal. SignalP predicts the signal peptide to be between residues 4 and 23. The signal is predicted to be cleaved between residues 23 and 24. The Motif search from BLOCKS contains an amidation site, which is found among proteins in the periplasmic space attached to the cell wall peptidoglycan by an amide bond.

The hypothetical protein 4719 shows good alignment with residues 33 to 286 of the best template 3pvaa in PROSPECT. The PROSPECT results returned a compact structure with high confidence level (>80%). The structural neighbours of the template 3pvaa were among the top hits by GenTHREADER and SAM, indicating a higher confidence of the prediction. The three dimensional structure of the hypothetical protein 4719 was generated using Modeller (Fig. 5).

FIG. 4. The computational searches identified four possible domains within the sequence of the query 4840, including N terminal TMD, an ATPase/GTPase domain, a Bag domain, and a coiled-coil domain.

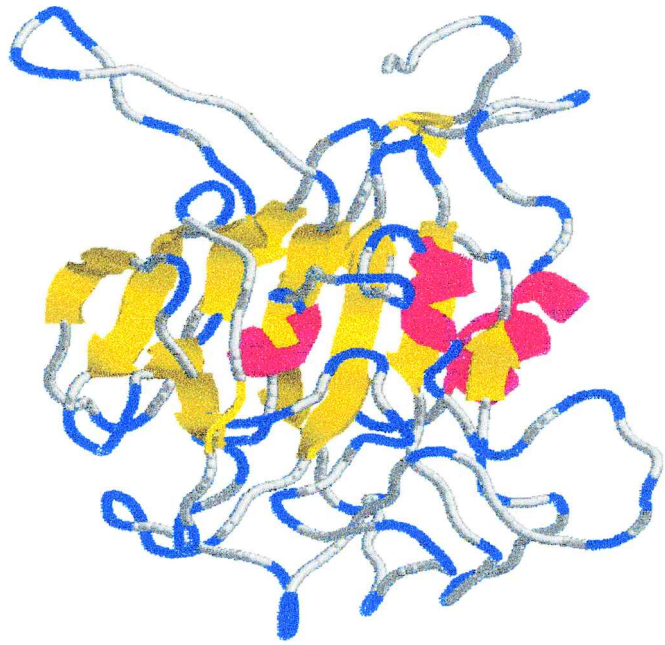

FIG. 5. The three-dimensional structure generated for protein 4719.

The sub-cellular location of the template penicillin acylase is periplasmic. It is consistent with the PSORT prediction of the protein 4719, which is also localized to the periplasmic space. Penicillin acylase can hydrolyse the antibiotic penicillin. Penicillin acylase belongs to the family of N-terminal-nucleophile hydrolases (Ntn hydrolases). The enzymes in this family all have a distinctive fold, but with no significant sequence consensus. The Ntn hydrolases fold to form an oxyanion hole containing an N terminal catalytic nucleophillic residue, which can differ (S, C, T). Beside the obvious function of breaking the antibiotic penicillin, the penicillin acylase may perform the function of breaking down aromatic compounds to use as a carbon source (Brannigan et al., 2002).

The *Shewanella* annotation database and Artemis were used to assess whether the gene 4719 was possibly on the same operon as other identified genes. The gene identified in the *Shewanella* files database as 4719 maps 24bp downstream of gene 4718. Gene 4718 maps 19bp downstream of gene 4717. These three genes appear to be on the same operon (Fig. 6). There is a very large intergenic region above gene 4717 which would indicate the presence of a promoter region. There is a 150 bp region downstream of gene 4719. TransTerm identified a rho independent termination sequence in this region which suggests that the downstream genes are not a part of the operon. 300 base pairs of the sequence upstream of gene 4717 were inspected for possible *fur* binding sites. A probable *fur*-binding site was identified at 118 bp upstream of the ATG transcriptional start site for gene 4717. This possible *fur*-binding site in *S. oneidensis* MR-1 is shown below together with the known one in *E. coli*.:

E. coli consensus	= GAT	AAT	GAT	AAT	CAT	TATC	(TATC)
S. oneidensis 4719	= GAT	AAT	AAA	TAT	CAT	TTAC	TATC

If this is in fact a *fur*-binding site it would indicate that these genes are co-regulated by the *fur* homolog of *S. oneidensis* MR-1 and are quite likely related to both each other in iron metabolism. Genes 4717 and 4718 on the same operon may also be related in some manner, although their functions are unknown (hypothetical genes as well). Gene 4717 has three TMDs and is located in the bacterial inner membrane. Gene 4718 is a soluble protein and is located in the cytoplasm associated with the bacterial inner membrane.

The *Shewanella* annotation database was used to search for homologs for the three related proteins. The three proteins each were related with high sequence similarity to *Xylella fastidosa* proteins XF0597, XF0596

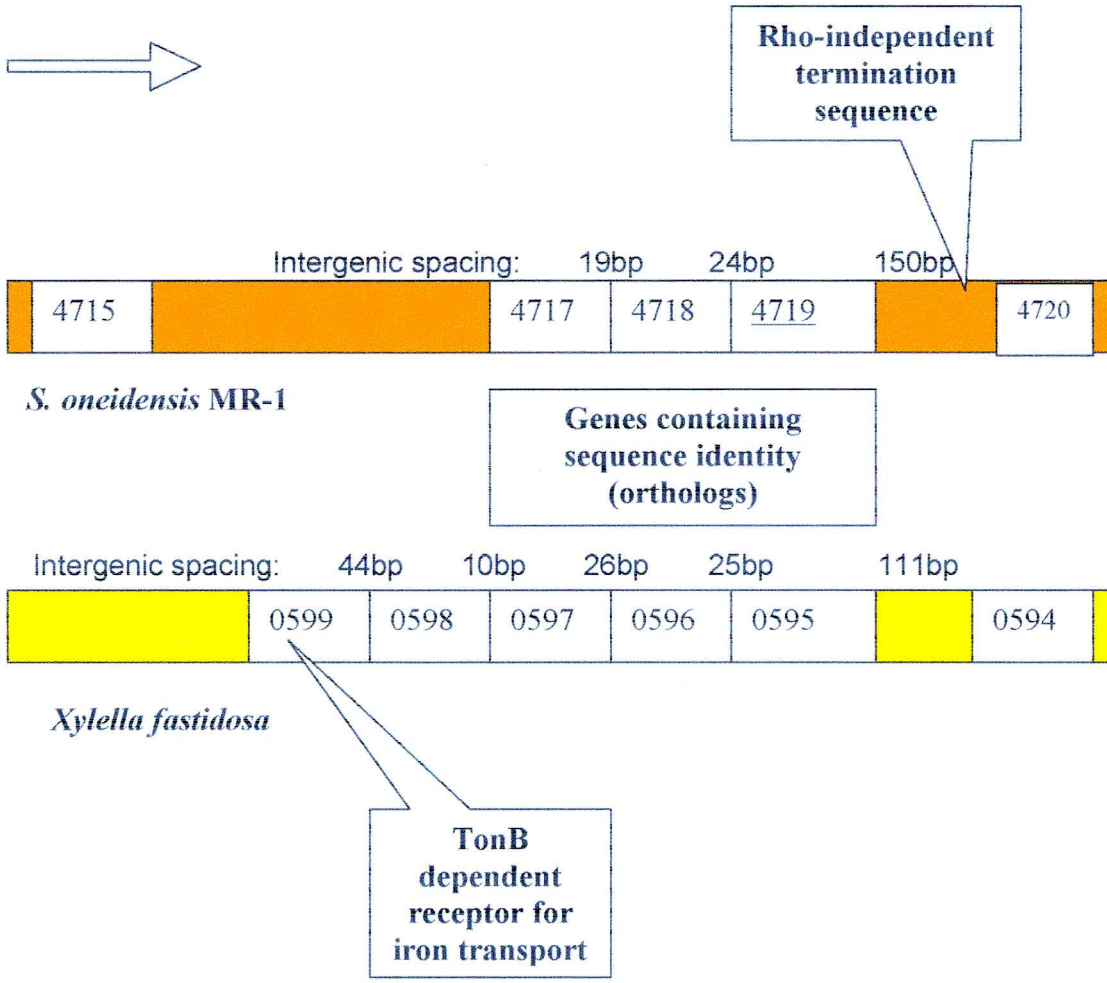

FIG. 6. The operon structure of the *Shewanella* genes 4719, 4718, and 4717 aligned with the operon structure of *Xylella fastidosa*.

and XF0595. The operon structure of these proteins was inspected using the Genome Channel. These three *Xylella fastidosa* proteins were found on the genome in the same order and orientation as the *S. oneidensis* MR-1 proteins. The *Xylella fastidosa* genes are presented on a possible operon with an annotated gene (Fig. 6). The upstream genes xf0598 and xf0599 in *Xylella fastidosa* appeared to be on the same operon as the three homologs to the *S. oneidensis* MR-1 genes, given the short intergenic regions. The functional annotation of gene xf0599 is TonB dependent receptor for iron transport, which implies it may be part of a TonB system for iron transport.

The TonB system is involved in import across the outer membrane (Enz et al., 2000). TonB is found in the periplasmic space. It is a large protein that spans from the cytoplasm to the outer membrane. Outer membrane receptors need energy to transport siderophores or other iron carriers such as ferrichrome across the membrane, but the source of the cells energy is the cytoplasmic membrane. TonB spans the periplasmic space and it transfers energy, via conformational change, to outer membrane receptors. The energy for this conformational change in TonB comes from proton translocation in the cytoplasm. This transport system also transports becteriophages, Vitamin B_{12}, colicins, and antibiotics across the outer membrane. This system is therefore involved in iron uptake and it may also be involved in extrusion of drugs (like antibiotics) and organic solvents via an unknown mechanism (Godroy et al., 2001). This is significant since the hypothetical protein 4719 has sequence similarity to a penicillin acylase. The outer membrane receptors

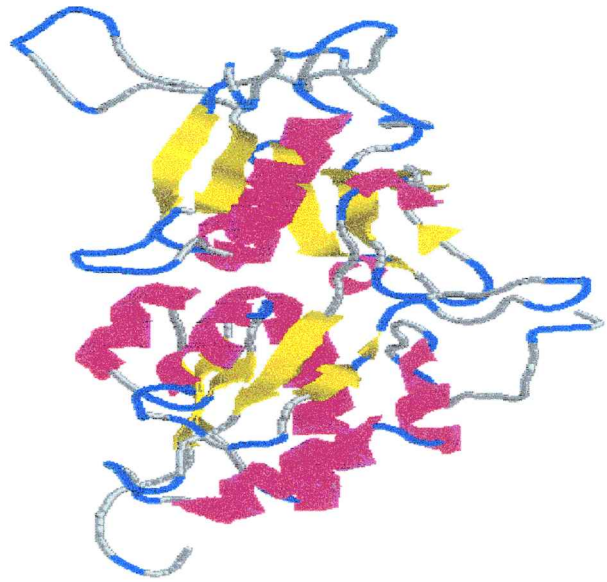

FIG. 7. The three-dimensional structure generated for the protein encoded by gene 3954.

like FhuA that interact with TonB and transport siderophores with bound iron into the cell also bind and transport vitamin B_{12} and antibiotics. There are many ways to deal with antibiotics once they are taken up. One way is to pump them out and another way to deal with them is to modify them. A penicillin acylase type protein may provide this modification within the periplasmic space.

Based on the results of the evaluation of gene 4719, the functional annotation suggested by the threading template is supported. Results suggest that the gene may be a member of the Ntn hydrolase family. The gene product of 4719 is predicted to be located in the periplasmic space and may have a function related to TonB and iron uptake metabolism.

Gene 3954

Gene 3954 of *S. oneidensis* MR-1 was up-regulated approximately threefold in the *fur* mutant. The protein was predicted to be soluble by SOSUI. A cleavable N-terminal signal sequence from residue 5 to 27 was predicted by SignalP. The subcellular localization of gene 3954 given by PSORT was periplasmic.

The hypothetical protein 3954 shows alignment with residues 10–270 of the best PROSPECT template 1atg with a compact structure. The PROSPECT template 1atg was also returned by the threading program GenTHREADER and SAM-T98. The certainty for GenTHREADER was 1.0, with probability of >99%. These results in combination were enough to allow the acceptance of the template 1atg. The three dimensional structure of the hypothetical protein 3954 was created using Modeller (Fig. 7).

Based on our literature search, 1atg is a periplasmic molybdate-binding protein (ModA). The periplasmic cell membrane receptor for molybdenum is a part of the ABC type molybdate transport system in bacteria.[1] From the FSSP database neighbours list close neighbours of 1atg are involved in binding and transport of inorganic ions. Receptor proteins for ion transport (such as iron or molybdenum) via the ABC system, such as AfuC and ModA, are in the same functional category in COGs as 1atg, "inorganic ion transport and metabolism." One class of ABC transporters, which are ATP binding cassette transporters, consist of

[1]Interestingly, when threading was carried out in 2001, no significant hit with known structure or function was found. The PSI_BLAST search was repeated in 2002. In the 2002 PSI_BLAST the search returned over a hundred hits better than threshold. These hits included known homologs to ModA. This result supports the original functional annotation done 1 year prior.

an outer membrane receptor such as FhuA which binds and transports nutrients such as vitamin B_{12}, sources of iron such as heme and siderophores, and antibiotics such as albomycin (an antibiotic of fungal origin with structural homology to ferrichrome), across the outer membrane using energy provided by the TonB system. Once in the periplasm these imported substances are transported across the inner membrane by a three-component system such as FhuCDB. The best characterized ABC transporter for iron uptake is the Fhu system in *E. coli* (Dassa et al., 2001). The FhuD protein is soluble and it is in the periplasmic space and associated with the cytoplasmic membrane, the FhuB protein is an integral membrane protein found in the cytoplasmic membrane and associated with FhuB and FhuC. FhuC is provides energy via hydrolysis of ATP and is localized to the inner side of the cytoplasmic membrane. The ATPase domain is highly conserved in all ABC transporters in all species; the amino acid sequence contains Walker A, and B motifs and a linker peptide signature motif (Dassa et al., 2001). Therefore, if an ATPase domain is identified, the other components of the ABC transporter are typically also identified. This is the case with the three genes 3954, 3955, and 3956.

These three components interact with each other, ATP and the substrate in a concerted fashion. This interaction allows transport of the substrate across the cytoplasmic membrane (Koster, 2001). The hypothetical protein 3954 has homology to the FhuD or MotA type binding protein, which is soluble, found in the periplasm, binds the substrate for transport, and associates with the cytoplasmic membrane and FhuB.

Both the FhuACDB and ModABC proteins are found on the same operon and are co-regulated by *fur* as a result of changing iron concentrations in *E. coli*. ABC transporter genes are often found on the same operon (Self et al., 2001). The genes are typically found in the following order: periplasmic binding protein, inner membrane protein, and ATPase. The hypothetical gene 3954 and associated genes 3955 and 3956 are found in this order. Genes of transport operons are often translationally coupled. These genes may have overlapping start and stop codons, which could be important to guarantee the proper stoichiometry of the transport components (Self et al., 2001). Genes 3954, 3955, and 3956 have overlapping start and stop codons.

The *Shewanella* annotation database and Artemis were used to assess whether the gene 3954 was possibly on the same operon as other genes identified to be components of an ABC transport system. The gene 3954 is located on the reverse strand and is the first in a group of eight genes which appear to present on the same operon. Some of the other genes are homologs to other ABC transporter components from *V. cholera* and to molybdenum metabolism proteins (Fig. 8).

Most of the eight genes in this suspected operon overlap or have a very small intergenic space. There is a rho-independent transcriptional stop site located after this group between genes 3961 and 3962 as identified by TransTerm. Three hundred base pairs of the upstream region from gene 3954 were searched for the presence of *fur* binding sites. A possible *fur*-binding site was found approximately 125 base pairs upstream from the transcriptional start site ATG, as shown below together with the known *fur* consensus in *E. coli*:

E. coli consensus =	GAT	AAT	GAT	AAT	CAT	—	TATC
S. oneidensis 3954 =	GTG	AAG	GTA A	AAG	CAT	GAA	TATC

The ModABC transport genes were compared to the genes 3954, 3955, and 3956. ModB has five TMDs (Self et al., 2001). Protein 3955 also was identified as having five TMDs by SOSUI. ModB contains three ABC permease signature sequences (Self et al., 2001). The hypothetical protein 3955 also contains these signatures. ModC contains Walker A and B ATPase signature sequences. These are also present in protein 3956.

Based on the results from the computational tools the functional annotation suggested by the PROSPECT threading result is supported. The gene 3954 is likely to be the periplasmic component of a molybdate ABC transporter.

DISCUSSION

Sequence data is being generated at an exponential rate. One of the primary steps in cataloguing genes of a newly sequenced genome is to identify the hypothetical genes and functionally annotate them. As shown in this paper, functional annotation based on structural prediction, in conjunction with other computational

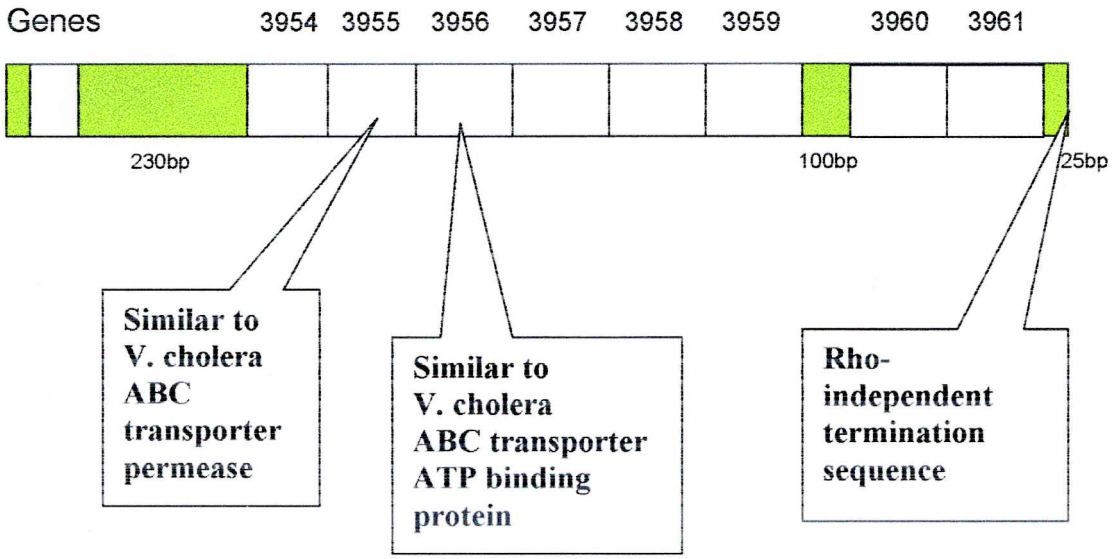

FIG. 8. The operon of *Shewanella* containing, for example, the genes 3954, 3955, and 3956. Other genes that appear to be on the same operon are functionally annotated as possible homologs to ABC transporter components.

tools, provides a powerful method to make predictions about hypothetical proteins. Having a predicted three-dimensional structure can suggest possible functions. This method of identifying structures and possible functions can be used by experimentalists to predict possible functions for the protein products of hypothetical genes. It would be even better that some high-throughput experiments, such as mass-spec and microarray measurements, are carried out before the computational studies, as done in our case. In this way, a set of hypothetical proteins can be selected related to particular phenotype, pathway, or cellular role. The relationship of these hypothetical proteins to specific protein pathways may also be predicted. This type of method may also be helpful in interpreting high-throughput data in general.

The computational approach used in this paper is a timely addition to the solutions for genome analysis. We have automated some of our analysis method into a computational pipeline, which is available at http://compbio.ornl.gov/proteinpipeline/ (Shah et al., 2003; Xu et al., 2003). The pipeline consists of seven logical phases, mimicking the manual process used in this paper: (1) preprocessing to identify protein domains in the input sequence, (2) compilation of functional and structural information about a query protein through database search, (3) protein triage to determine which process and prediction branches to use, (4) protein fold recognition for identification of native-like folds of a query protein, (5) protein structure prediction to generate atomic structure models, (6) quality assessment of predicted structure, and (7) prediction result validation. Different processing and prediction branches are determined automatically and employed for each individual protein, by a prediction pipeline manager, based on identified characteristics of the protein. The pipeline has been implemented to run in a heterogeneous computational environment, consisting of Alpha, Solaris and Linux servers, a 64–node Linux cluster and a wide range of ORNL supercomputers as a client/server system with a web interface. A number of genome-scale applications have been carried out on microbial genomes, including *Caenorhabditis elegans*, *Pyrococcus furious*, *Prochlorococcus* sp MED4 and MIT9313, and *Synechococcus* WH8102. The pipeline assess the confidence level of the prediction through either E-value of PSI-BLAST hit (if any) or Z-score of PROSPECT, the latter of which can translate into the probability of structural prediction accuracy, as well as the likelihood of the functional relationship between the query protein and the template hit. Overall, the PROSPECT pipeline identified structural homologs in PDB with reasonable level of confidence (through either PSI-BLAST with E-value less than 10^{-4} or PROSPECT with Z-score 8.0 or above[2]) for 50–55% of all the ORFs in each of a genome

[2]Z-score 8 or above corresponds to at least 63% chance for the structural prediction to be correct.

(Shah et al., 2003; Xu et al., 2003). Together with annotations of membrane proteins, about 75% of all the ORFs in a genome are characterized.

Limitations of this method include the following: (1) The structural template may not be available for a hypothetical protein; currently, one out of 10 newly solved structures of globular proteins does not have a known fold in PDB (www.rcsb.org/pdb/). (2) The resulting functional annotations are still computational results and require experimental validations. The functional annotations given by this method should be used as a tool to allow experimentalists to identify hypothetical genes of interest for further study.

ACKNOWLEDGMENTS

The work was supported by the Office of Biological and Environmental Research, U.S. Department of Energy, under Contract DE-AC05–00OR22725, managed by UT-Battelle, LLC. We would like to thank Dr. Jeffrey M. Becker for helpful discussions. C.Y. acknowledges support from the UT-ORNL Graduate School of Genome Science and Technology. This paper is derived from C.Y.'s thesis under the guidance of D.X.

REFERENCES

ALTSCHUL, S.F., MADDEN, T.L., SCHÄFFER, A.A., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25,** 3389–3402.

ATTWOOD, T.K., BLYTHE M., FLOWER D.R., et al. (2002). PRINTS and PRINTS-S shed light on protein ancestry. Nucleic Acids Res **30,** 239–241.

BATEMAN, A., BIRNEY, E., CERRUTI, L., et al. (2002). The Pfam protein families database. Nucleic Acids Res **30,** 276–280.

BAIROCH, A., and APWEILER R. (1999). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. Nucleic Acids Res **27,** 49–54.

BERMAN, H.M., WESTBROOK, J., FENG, Z., et al. (2002). The ProteinData Bank. Nucleic Acids Res **28,** 235–242.

BRANNIGAN, J., DODSON, G., and WILSON, K. (2002). *Ntn Hydrolases* (York Structural Biology Laboratory. Structural Biology Laboratory Department of Chemistry, University of York, Heslington, York, U.K.).

BRENNER, S., A. (2001). Tour of structural genomics. Nat Rev Genet **2,** 801–809.

CASP. (1995). Protein structure prediction issue. Proteins Struct Funct Genet **23,** 295–462.

CASP. (1997). Protein structure prediction issue. Proteins Struct Funct Genet Suppl 1 **29,** 1–230.

CASP. (1999). Protein structure prediction issue. Proteins Struct Funct Genet Suppl 3 **37,** 1–237.

CASP. (2001). Protein structure prediction issue. Proteins Struct Funct Genet Suppl 4 **45,** 1–199.

CORPET, F., SERVANT, F., GOUZY, J., et al. (2000). ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. Nucleic Acids Res **28,** 267–269.

DANINO, D., and HINSHAW, J.E. (2001). Dynamin family of mechanoenzymes. Curr Opin Cell Biol **13,** 454–460.

DASSA, E., and BOUIGE, P. (2001). The ABC of ABC's: a phylogenetic and functional classification of ABC systems in living organisms. Res Microbiol **152,** 211–229.

DUBRAC, S., and TAUATI, D. (2000). Fur positive regulation of iron superoxide dismutase in *Escherichi coli*: functional analysis of the *sodB* promoter. J Bacteriol **182,** 3802–3808.

ENZ, S., MAHREN, S., STROEHER, U.H., et al. (2000). Surface signaling in ferric citrate transport gene induction: interaction of the FecA, FecR, and FecI regulatory proteins. J Bacteriol **182,** 637–646.

ERMOLAEVA, M.D., KHALAK, H.G., WHITE, O., et al. (2000). Prediction of transcription terminators in bacterial genomes. J Mol Biol **301,** 27–33.

ESCOLAR, L., PEREZ-MARTIN, J., and DELORENZO, V. (1999). Opening the iron box: transcriptional metalloregulation by the fur protein. J Bacteriol **181,** 6223–6229.

FETROW, J.S., SIEW, N., DI GENNARO, M.M., et al. (2001). Genomic-scale comparison of sequence- and structure-based methods of function prediction: does structure provide additional insight? Protein Sci **10,** 1005–1014.

GODROY, P., RAMOS-GONZALEZ, M.I., and RAMOS, J.L. (2001). Involvement of the TonB system in tolerance to solvents and drugs in *Pseudomonas putida* DOT-TIE. J Bacteriol **183,** 5285–5292.

HANTKE, K. (2001). Iron and metal regulation in bacteria. Curr Opin Microbiol **4,** 172–177.

HEIDELBERG, J.F., PAULSEN, I.T., NELSON, K. E., et al. (2002). Genome sequence of the dissimilatory metal ion-reducing bacterium *Shewanella oneidensis.* Nat Biotechnol **20,** 1118–1123.

HENIKOFF, S., and HENIKOFF, J.G. (1991). Automated assembly of protein blocks for database searching. Nucleic Acids Res **19,** 6565–6572.

HENRIK, N., ENGELBRECHT, J., BRUNAK, S., et al. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng **10,** 1–6.

HIROKAWA, T., SEAH, B.-C., and MITAKU, S. (1998). SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics **14,** 378–379.

HOLM, L., and SANDER, C. (1996). Mapping the protein universe. Science **273,** 595–602.

JONES, D.T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol **287,** 797–815.

KARPLUS, K., BARRETT, C., and HUGHEY, R. (1998). Hidden Markov models for detecting remote protein homologies. Bioinformatics **14,** 846–856.

KIEFT, T.L., FREDRICKSON, J.K., ONSTOTT, T.C., et al. (1999). Dissimilatory reduction of Fe(III) and other electron acceptors by a *Thermus* isolate. Appl Environ Microbiol **65,** 1214–1221.

KOSTER, W. (2001). ABC transporter-mediated uptake of iron, siderophores, heme and vitamin B_{12}. Res Microbiol **152,** 291–301.

LARIMER, F., HAUSER, L., and LAND, M. (2002). Draft analysis of the *Shewanella* genome generated through the microbial annotation pipeline at ORNL (unpublished data).

LUPAS, A., VAN DYKE, M., and STOCK, J. (1991). Predicting coled coils from protein sequences. Science **252,** 1162–1164.

McGUFFIN, L.J., BRYSON, K., and JONES, D.T. (2000). The PSIPRED protein structure prediction server. Bioinformatics **16,** 404–405.

MDL INFORMATION SYSTEMS. (1999). CHIME (MDL Information Systems, Inc., San Leandro, CA).

SAYLE, R.A., and MILNER-WHITE, E.J. (1995). RASMOL: biomolecular graphics for all. Trends Biochem Sci **20,** 374–376.

MYERS, C.R., and NEALSON, K.H. (1990). Respiration-linked proton translocation coupled to anaerobic reduction of manganese(IV) and Iron(III) in *Shewanella prtrefaciens* MR-1. J Bacteriol **172,** 6232–6238.

NAKAI, K. (2000). Protein sorting signals and prediction of subcellular localization. Adv Protein Chem **54,** 277–344.

NEALSON, K.H., and SAFFARINI, D. (1994). Iron and manganese in anaerobic respiration: environmental significance, physiology, and regulation. Annu Rev Microbiol **48,** 311–343.

PESSANHA, M., BRENNAN, L., XAVIER, A.V., et al. (2001). NMR structure of the haem core of a novel tetrahaem cytochrome isolated from *Shewanella firgidimarina*: identification of the haem-specific axial ligands and order of oxidation. FEBS Lett **489,** 8–13.

ROST, B. (1996). PHD: predicting one-dimensional protein structure by profile-based neural networks. Methods Enzymol **266,** 525–539.

ROST, B., and SANDER, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol **232,** 584–599.

ROTHMAN, J.H., RAYMOND, C.K., GILBERT, T., et al. (1990). A putative GTP binding protein homologous to interferon-inducible Mx proteins performs an essential function in yeast protein sorting. Cell **61,** 1063–1074.

RUTHERFORD, K., PARKHILL, J., CROOK, J., et al. (2000). Artemis: sequence visualisation and annotation. Bioinformatics **16,** 944–945.

SAFFARINI, D.A., DICHRISTINA, T.J., BERMUDES, D., et al. (1994). Anaerobic respiration of *Shewanella putrefaciens* requires both chromosomal and plasmid-borne genes. FEMS Microbiol Lett **119,** 271–278.

SALI, A., and BLUNDELL, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol **234,** 779–815.

SELF, W.T., GRUNDEN, A.M., HOSONA, A., et al. (2001). Molybdate transport. Res Microbiol **152,** 311–321.

SHAH, M., PASSOVETS, S., KIM, D., et al. (2003). A computational pipeline for protein structure prediction and analysis at genome scale. Proceedings of the 3rd IEEE Symposium on Bioinformatics and Bioengineering, pp. 3–10.

SKOLNICK, J., and FETROW, J.S. (2000). From genes to protein structure and function: novel applications of computational approaches in the genomic era. TIBTECH. **18,** 34–39.

SONDERMANN, H., SCHEUFLER, C., SCHNEIDER, C., et al. (2001). Structure of a Bag/Hsc70 complex: convergent functional evolution of Hsp70 nucleotide exchange factors. Science **291,** 1553–1557.

THOMPSON, D.K., BELIAEV, A.S., GIOMETTI, C.S., et al. (2002). Transcriptional and proteomic analysis of a ferric uptake regulator (fur) mutant of *Shewanella oneidensis*: possible involvement of fur in energy metabolism, transcriptional regulation, and oxidative stress. Appl Environ Microbiol **68,** 881–892.

VENKATESWARAN, K., MOSER, D.P., DOLLHOPF, M.E., et al. (1999). Polyphasic taxonomy of the genus *Shewanella* and description of *Shewanella oneidensis* sp. nov. Int J System Bacteriol **49,** 705–724.

WU, C., ZHAO, S., CHEN, H.L., et al. (1996). Motif identification neural design for rapid and sensitive protein family search. CABIOS **12,** 109–118.

XU, D., XU, Y., and UBERBACHER, E.C. (2000). Computational tools for protein modeling. Curr Protein Peptide Sci **1,** 1–21.

XU, D., and XU, D. (2002). bComputational studies of protein structure and function using threading program PROSPECT. In *Protein Structure Prediction: Bioinformatic Approach.* I. Tsigelny, ed. (International University Line Publishers, La Jolla, CA), pp. 5–41.

XU, D., KIM, D., DAM, P., et al. (2003). Characterization of protein structure and function at genome scale using a computational prediction pipeline. In *Genetic Engineering, Principles and Methods.* J.K. Setlow, ed. (in press).

XU, Y., and XU, D. (2000). Protein threading using PROSPECT: design and evaluation. Proteins Struct Funct Genet **40,** 343–354.

XU, Y., XU, D., and OLMAN, V. (2002). A practical method for interpretation of threading scores: an application of neural network. Stat Sin **12,** 159–177.

Address reprint requests to:
Dr. Dong Xu
Protein Informatics Group
ORNL
1060 Commerce Park Drive
Oak Ridge, TN 37831-6480

E-mail: xud@ornl.gov