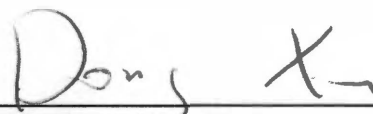
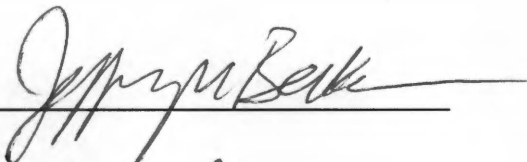
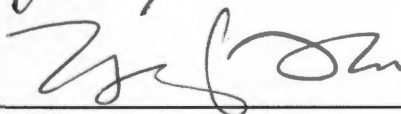


To the Graduate Council:

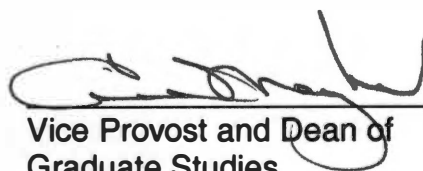
I am submitting herewith a thesis written by Christal Yost entitled "A computational study of *S. oneidensis* MR-1: Discovering the role of hypothetical proteins." I have examined the final paper copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Life Sciences.


Dong Xu, Major Professor

We have read this thesis and
recommend its acceptance:

Acceptance for the Council:


Vice Provost and Dean of
Graduate Studies

Thesis
2002
.4577

**A computational study of *S. oneidensis* MR-1:
Discovering the role of hypothetical proteins**

**A Thesis
Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville**

**Christal Yost
August 2002**

Acknowledgements

Thanks to all who helped me complete this Master of Science degree in Bioinformatics.

I Thank Dr. Dong Xu for his patient direction in every aspect of this learning experience. Thanks to Frank Larimer and Loren Hauser for help with Artemis and evaluation of the gene structures. Most computer related matters including programs written in java, perl and HTML, were accomplished with the help of Dr. Scott Yost.

Thanks to all of the members of the Protein Informatics Group and to the members of the Microbial Genomics and Ecology Group.

Abstract

The genomes of many organisms have been sequenced in the last five years. The characterization of these proteomes is the next step in the process of attaining a global understanding of the workings of the pathways and protein complexes of organisms. It is difficult to computationally assign functional annotations for between thirty and fifty percent of hypothetical proteins from a newly sequenced genome. Sequence based search methods sometimes return little or no information about hypothetical proteins. In these situations three dimensional structural predictions combined with a suite of computational tools can suggest possible functions for these hypothetical proteins. The premise for development of this method is therefore that sequence-structure-function annotations give experimentalists information that is not available using sequence searches alone. The ability to predict function for more of these hard to annotate hypotheticals will allow better characterization of proteomes, will help experimentalists choose possible candidate genes related to their area of research for further study, and will help with interpretation of microarray data.

By predicting the secondary structures and then the three-dimensional structures using the protein threading program PROSPECT, three hypothetical proteins of *S. oneidensis* MR-1 were assigned preliminary functional annotations. Once these preliminary predictions were made using protein threading, literature searches and computational tools were used to complete the functional predictions, and evaluate the accuracy of the preliminary predictions. Computational tools were used to perform transmembrane domain predictions, coiled coil predictions, signal peptide predictions, sub-cellular localization predictions, searches for conserved functional motifs, and operon structure evaluations. Results supported the preliminary sequence-structure-function predictions for the three hypothetical *S. oneidensis* MR-1 proteins suggesting that additional functional characterization may be performed using this computational method.

This method provides an easy to understand protocol which uses freely available computational tools. Molecular biologists may use this method to gain insight into possible functions for hypothetical proteins of interest.

Table Of Contents

Chapter	Page
1 Introduction.....	1
2 Method.....	8
2.1 Protocol.....	8
3 Results and Conclusions.....	23
3.1 Gene 4840.....	23
3.2 Gene 4719.....	36
3.3 Gene 3954.....	44
4 Discussion.....	53
List of References.....	56
References.....	57
Web Tool links.....	60
Vita.....	61

List Of Figures

Figure		Page
1	Flow chart sequence-structure-function.....	2
2	PSIpred prediction example.....	9
3	Prospect output.....	12
4	Predicted vs X-ray models.....	15
5	PSI-BLAST and CD search graphics for 4840	19
6	Method flowchart.....	24
7	Secondary structure prediction, 4840.....	25
8	Three dimensional model, 4840.....	27
9	Bag domain alignment, 4840.....	29
10	ATP binding signature, 4840.....	31
11	Coiled-coil, 4840.....	33
12	Domains, 4840.....	35
13	Secondary structure prediction, 4719.....	37
14	Three dimensional model, 4719.....	38
15	Operon structure, 4719.....	41
16	Secondary structure prediction, 3954.....	45
17	Three dimensional structure, 3954.....	47
18	Operon structure, 3954.....	51

1. Introduction

It has become clear from surveying the literature that predicting the three dimensional structure of a protein from its amino acid sequence is not only possible but, as the number of known structures grows and structural information becomes more readily available, prediction accuracies are increasing. Presently the ability to predict a verifiable three dimensional structure in assessments using the protein threading program PROSPECT is approximately sixty-five to seventy percent for fold recognition (Xu et al. 2000). Basic biochemistry teaches that function is directly related to structure. Given the good prediction accuracy and the structure-function relationship, it has been suggested by many researchers that three dimensional structures predicted by protein threading may be used to help predict the functions of proteins (Skolnick et al. 2000, Brenner 2001, Xu et al. 2002). The technique of predicting function from sequence through structural prediction has been described as the sequence-structure-function method of functional gene annotation (Figure 1).

Why is this sequence-structure-function method a good complement to sequence comparison? PSI-BLAST (Position Specific Iterated - BLAST) (Altschul et al., 1997), using multiple sequence profiles and SAM (Karplus et al. 1998) using hidden Markov models (HMM) are able to identify more remote homologs than predecessors that used only sequence alignments, but they still cannot identify as many remote homologs as protein threading. Over evolutionary time genes experience mutations. These

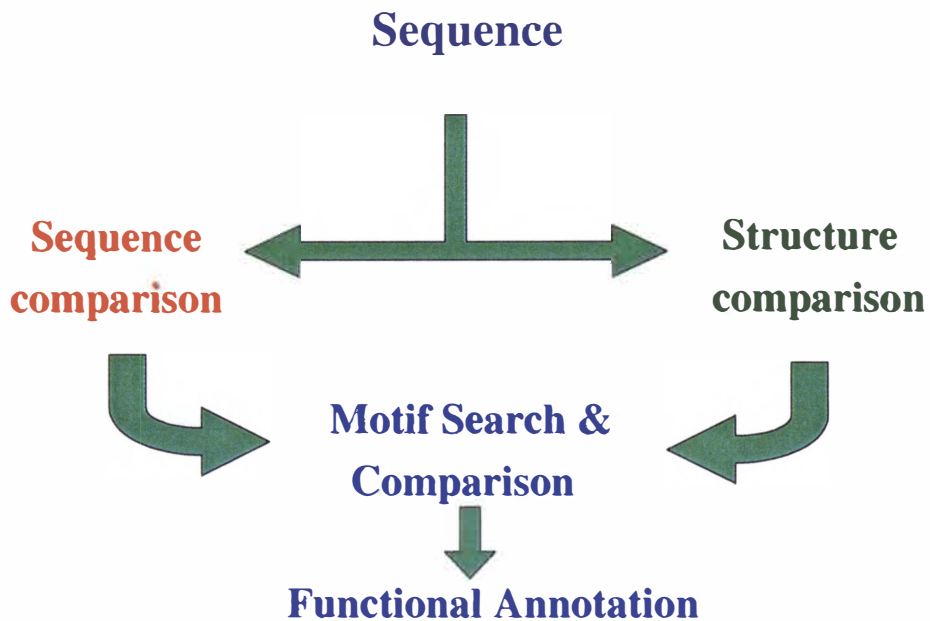


Figure 1
Flow chart showing that functional annotation by sequence-structure-function complements functional annotation by sequence-sequence-function.

mutations can cause a change in both the DNA sequence and the amino acid residues that are incorporated into the final protein product. When changes in amino acid sequence are tolerated, (i.e. an acidic residue is exchanged for an acidic residue etc.) the final protein product can maintain its structure or function. Protein threading can identify these more remote homologs because it uses the superimposition of three dimensional structures to detect similar structures. The classic example of this scenario is horse haemoglobin and whale myoglobin. These two proteins have very similar structure yet they have very little sequence similarity (Brenner 2001). The sequence-structure-function method is a good alternative when sequence comparison fails. Threading compares the secondary structure and pairwise interaction between residues in three dimensions, and it will allow identification of homologs that are undetectable using sequence comparison alone (Xu et al. 2000).

A new sequence-structure-function method was used to predict the function of hypothetical genes from *S. oneidensis* MR-1. Using a practical example to explain this new method allows potential users to better understand its use, strengths and limitations. Because understanding the merit of the results requires knowledge about *S. oneidensis* MR-1 and its importance, some background information is presented.

There are many contaminated waste sites that contain mixtures of metals, chlorinated solvents, and hydrocarbons. In anoxic environments some organisms are able to use terminal electron acceptors such as iron,

nitrate, manganese, nitrite, and fumarate in respiration. One of the most abundant metals at contaminated waste sites is insoluble ferric iron (Fe^{3+}) in the form of iron oxides. Iron oxides are essentially insoluble in water at pH 7, and in many anoxic near surface environments Fe^{III} (ferric iron) is a major element involved in the formation of the crystal structure of mineral deposits (Myers 1997, Nealson 1994). A large number of dissimilatory iron-reducing bacteria have been identified. Among the many Genera of these bacteria are *Shewanella*, *Thermus*, *Ferrimonas*, *Pelobacter*, *Geovibrio*, *Geospirillum*, *Geothrix*, *Desulfuromusa* and *Desulfuromonas* (Kieft et al. 1999, Nealson et al. 1994). The potential of some of these bacteria to be useful for removal of some percentage of the undesirable toxic metals and halogenated organics from polluted sites in bioremediation has been noted. For example; *Geobacter* is known to completely oxidize benzene and toluene to CO_2 , and some iron respirers as well as other types of bacteria are able to reduce soluble toxic metals such as uranium(VI), chromium, cobalt, and selenium, forming precipitates (Venkateswaran et al. 1999, Dobbin et al. 1999, Nealson et al. 1994, Lovely et al. 1991, Lovely 2001).

Shewanella oneidensis MR-1 (formerly *Shewanella putrefaciens* MR-1) is a gram negative, facultatively anaerobic, dissimilatory metal-reducing bacterium. *S. oneidensis* MR-1 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 (Nealson et al. 1994, Myers et al. 1990, Pessanha et al. 2001). Because the use of *S. oneidensis* MR-1 could produce unexpected environmental changes, it is important to understand the biochemical pathways and proteins involved in iron metabolism.

Ferric uptake regulator protein *fur* has been identified primarily as a transcriptional regulator that suppresses or increases expression of proteins related to iron metabolism. Genes whose expression levels vary in the presence or absence of *fur* are most likely related to iron metabolism. When iron levels are high enough that diffusion provides the bacteria with sufficient amounts of iron, *fur* complexed with Fe^{2+} negatively regulates genes related to iron uptake metabolism. In the *fur* mutant these genes are expected to be up-regulated. Ferric uptake regulator protein, *fur*, is also implicated in positive regulation of genes like super-oxide dismutase. (SOD) 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, whose expression levels are increased in the presence of *fur* complexed to Fe^{2+} , are expected to show lower expression levels (Escobar et al. 1999, Braun 2001, Dubrac et al.

2000). 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). Sequences of three genes with notable changes in expression pattern from this mutant were functionally annotated using this new computational sequence-structure-function method.

The genome of *S. oneidensis* MR-1 was recently sequenced by TIGR (Preliminary sequence data was obtained from The Institute for Genomic Research website at <http://www.tigr.org>. Sequencing of *Shewanella oneidensis* (*putrefaciens*) MR-1 was accomplished with support from the Department of Energy.). High throughput computational tools presently available were used to identify and functionally annotate a mostly complete set of expected genes in the *Shewanella* genome. These annotations were available in house as the *Shewanella* files (Larimer et al. 2002). Computational identification of genes in a newly sequenced genome refers to the process of using computer algorithms to identify DNA sequence that contains components that allow a sequence to be identified as a prospective gene. Examples of these components would be upstream regulatory protein binding sites such as sigma or *fur* binding sites, TATA or GAGA box, ATG and or GTG transcriptional start sites, and transcriptional stop sequences. Once these hypothetical genes have been identified some of them can be given predicted functional annotations

based on sequence based search methods. The *S. oneidensis* MR-1 genome appears to have approximately four thousand five hundred genes. About sixty percent of the genes in the *Shewanella* files have functional annotations (Larimer et al. 2002). Approximately forty percent of the newly identified genes had no sequence similarities to previously studied and documented genes in the databases and were labelled hypothetical. It would take a tremendous amount of time and expense for the experimentalists to express and verify the functions of all of the hypothetical genes identified in *S. oneidensis* MR-1. It is more efficient for experimentalists to be able to target a limited set of protein products for expression and further study. Therefore, it has become apparent that an easy to use computational method, other than sequence based annotation, is necessary to allow experimentalists to attain the functional predictions they need to identify genes of interest for further study (Xu et al. 2002, Brenner 2001, Fetro et al. 2001).

Most methods that have been described for making functional predictions using structure are difficult to understand. A user friendly method that gives good results was evaluated using *S. oneidensis* MR-1 and is described here. The computational tools used in this method are readily available, easy to use and their use is free.

2 Method

2.1 Protocol

1. Database Search: Sequences of three genes from *Shewanella oneidensis* MR-1, with notable changes in expression levels, from the (FUR1) mutant were searched against the *Shewanella* files (Larimer et al. 2002). The *Shewanella* files is an in house database that contains functional annotations, for genes with identifiable homologs, that were obtained using sequence based searches. No functional annotation was found for any of the three genes.

2. Secondary Structure Prediction: The amino acid sequences for the hypothetical proteins were used to predict the secondary structures using PHD (Rost et al. 1993) and PSIPRED (McGuffin et al. 2000). The PSIPRED protein structure prediction server was used to carry out a secondary structural prediction on the proteins of interest (Figure 2). The PSIPRED protein structure prediction server allows submission of a protein sequence via the internet, and the results of the prediction are returned via E-mail. The actual averaged accuracy of PSIPRED is around eighty percent. This program incorporates an average of the output from up to 4 separate neural networks in the prediction process. These perform an analysis on output obtained from PSI-BLAST. The most widely used prediction server for secondary structure is PHD which uses JPRED (Cuff

Conf: Pred:
 Pred: CCCCCCECCCCCHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCEEEEEECCCCCCCCCHHHHHHHHHHCCCCCEEEEBCCC
 AA: LSWNPIFYQGVSMNLVKSIMKSLGLVVAASMLTVLPAQSAEVIKLATTSTENSLQLQELLPKFESASGYKVQVIATGT
 10 20 30 40 50 60 70 80

Conf: Pred:
 Pred: CHHHHCCCCCCCCCEEEBECCHHHHHHHHHHCCCCCCCCCEECCEEEBECCHHCCCCCHHHHHHHHHHHHCCCCCEEEBEC
 AA: GKALKLGEQGDVDMTHAPSAEAKFVADQFGVEPRGIMENDFVVLGPKNDPAKLRESKTAEAEFAKIAKSGLPPLSRGD
 90 100 110 120 130 140 150 160

Conf: Pred:
 Pred: CCCCCCHHHHHHHHHHCCCCCCCCCCCCCHHHHHHHHHHCCCCCECCCCCEEEECCEEEEEECCCCCCCCCEEEEC
 AA: NSGTHIKELEVWKAAGVTPDFKGYTSVGQMGKTLMLANE LQGYTSLDRGTFVAYKTKLDLGVPDGGKTLANPYQVILI
 170 180 190 200 210 220 230 240

Conf: Pred:
 Pred: CCHHCCCCCHHHHHHHHHHHHCHHHHHHHHHHCEBCCCEEEEEECCC
 AA: NPAKYPDLNHHKAKAFSDWLISKGGQAMINSFKVEGEQLFKATYSF
 250 260 270 280

 Helix
 Strand
 Coil

 - +
 Confidence of Prediction

Conf: confidence of prediction Pred: predicted secondary sequence AA: target sequence

The PSIPred prediction server will return a post script file which will allow visualization of the secondary structures of the amino acid sequence. The structures shown are α helix shown as green barrels, β sheet shown as yellow arrows. The loop regions are not highlighted. Simple visualization of this secondary structural prediction can be informative. A helix near the N terminus can suggest there may be a signal sequence and large areas of coil or helix can also suggest what types of structures might be expected such as coiled coil. It is important to note that paying attention and making these types of observations is necessary. These observations can help save time and avoid error in the same way as making approximations when solving math or chemistry problems. For example if the result from the threading program includes structures that contain almost all helix and the secondary structural prediction was predominantly beta sheet the user should re-check the submission and parameters chosen PSIPRED (McGuffin et al. 2000).

et al. 1999) for the secondary structural prediction (ssp). This server uses a combination of neural networks and multiple sequence alignments. Like PSIPred, PHD also allows the user to submit sequence via the internet, and it returns the results to an e-mail address, typically within a few hours. The secondary structural predictions from PHD/JPRED have an averaged accuracy of around 80%, and these PHD/JPRED predictions were used as input for PROSPECT 1.0 (Xu et al. 2000).

3. Threading and Modeling:

a) **Threading:** Two programs were used for protein threading. GenTHREADER (Jones 1999) is a fold recognition program available at the PSIPRED protein structure prediction server. It was run so that the results could be compared with those of the PROSPECT threading program. GenTHREADER is somewhat similar to PROSPECT in that it returns a list of PDB templates with high probability of having a similar three dimensional structure to the query.

PROSPECT (Xu et al. 2000), PROtein Structure Prediction and Evaluation Computer Tookit, is a threading-based protein structure prediction system. Protein threading is a powerful tool for making protein structural predictions, because most globular proteins contain different combinations of conserved folds that are similar to known structures. Using a global approach to find alignments PROSPECT uses secondary structural comparisons and electrostatic, hydrophobic, steric, and other interactions that effect how the protein may fold (Xu et al. 2002).

Constraints such as disulfide bonds, gaps or active sites can be added to the search. PROSPECT recognizes similar folds between the query and templates. Templates are catalogued in databases like the protein database PDB (Berman 2000) and the Fold classification based on Structure-Structure alignment of Proteins (FSSP) database (Holm 1996).

After all template candidates for the threading result have been returned in PROSPECT, a final evaluation based on compactness and neural network assessment of threading reliability determines which will be included in the output (Xu et al. 2002) (Figure 3).

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 such as <http://compbio.ornl.gov/~csecrest/3208-nterm/> generated by PROSPECT. The two columns from the left include a link to the sequence alignment and a link to the three dimensional template via CHIME. Using the web interface allows the user to visually compare the sequence and three dimensional alignment of the template to the query. To display the alignment of the query to the template in CHIME the user can right click on the three dimensional structure and choose Color and then Temperature. The red area represents the length of the backbone where the template and the query are most likely to have a similar backbone structure.

Once the output has been reviewed by the user, the templates with the best scores and visual alignments are chosen. Scores for templates

PROSPECT Results for whole:

Protein	raw	Zscore	C-ndx	mut	sing	pair	gap	ss-fit	Rank	Function
lc1gd	-12050	-9.99	10.37	-1925	-888	67	1079	-10382	1	tropomyosin
lc1gb	-11874	-9.99	10.38	-1084	-737	18	165	-10237	2	tropomyosin
lc1ga	-11830	-9.99	10.41	-1060	-704	33	165	-10265	3	tropomyosin
lc1gc	-11758	-9.99	10.42	-1056	-735	-46	165	-10086	4	tropomyosin
lc1i	-11248	-9.99	4.07	-1351	2	-511	2005	-11393	5	colicin ia biological unit
lbj5	-10469	-9.99	1.86	-1178	-1579	-69	3976	-11618	6	human serum albumin biological unit
lqgra	-10448	-9.99	1.87	-1428	-1315	-690	5225	-12239	7	importin beta subunit (karyopherin beta-1, nuclear factor)
lb3ua	-10269	-9.99	2.46	-2354	-944	-585	5124	-11510	8	protein phosphatase pp2a fragment
lavc	-9941	-9.99	1.83	-2088	-1143	-513	6014	-12211	9	annexin vi (p68, protein iii, 67-kda-calmodulin)
ldg3a	-9891	-9.99	2.81	-1228	-786	-384	4106	-11600	10	interferon-induced guanylate-binding protein 1 fragment (gua)
3bct	-9225	-9.99	2.17	-933	-508	-321	2732	-10194	11	beta-catenin fragment
lee4a	-9109	-9.99	2.09	-1363	-511	-523	3224	-9936	12	karyopherin alpha fragment (serine-rich RNA polymerase i)
lqsa	-9109	-9.99	1.97	-1262	-517	-219	4306	-11416	13	soluble lytic transglycosylase slt70 fragment
lhb5a	-8989	-9.99	2.14	-1221	-698	-427	3250	-9894	14	karyopherin alpha fragment (importin alpha, erlp)
lpprm	-8453	-9.99	1.79	-674	-724	-319	1691	-8427	15	peridinin-chlorophyll protein (pcp)
5eat	-8254	-9.99	1.68	-1781	66	-158	3793	-10176	16	5-epi-aristolochene synthase (5-epi-aristolochene cyclase)
lqkb	-8213	-9.99	1.86	-1673	-720	-396	4539	-9964	17	karyopherin beta2 fragment ran fragment
lqqea	-8142	-9.99	2.29	-580	-625	-122	1279	-8093	18	vesicular transport protein sec17 Mutant biological unit
lcuna	-8117	-9.99	2.96	-679	-735	-87	566	-7182	19	alpha spectrin fragment Mutant
lquua	-8108	-9.99	3.44	-343	-634	-204	574	-7501	20	human skeletal muscle alpha-actinin 2 fragment
lc3ca	-8024	-9.99	2.34	-1854	-928	-431	3776	-8587	21	adenylosuccinate lyase
luby	-7822	-9.99	1.74	-663	-530	-350	2420	-8700	22	farnesyl diphosphate synthase (fps) Mutant biological unit
2occa	-7794	-9.99	1.56	374	-414	-466	3358	-10646	23	cytochrome c oxidase (ferrocytochrome c)

Alignment
of query
and
template

Threading
template
view using
CHIME

Figure 3
Prospect output

that should be considered are as follows: Mutation energy (mut) should be a relatively low number in comparison to other templates under consideration. A lower Mutation energy indicates a higher sequence similarity. The ssfit (secondary structural fit) value should also be lower for the chosen template. This indicates that the query and template have significant similarity in secondary structure. The Reliability for the chosen template should be above 0.7, but it is important to look at the big picture. When most of results are good and the template is also returned with good values by other threading programs the top reliability values can become acceptable even if they are below threshold. The c-ndx is an indication of the ability of the protein to fold into an acceptably compact structure, and any value below 3 is good. When the template has a coiled coil it becomes acceptable for this score to be a little above three.

Top templates chosen from PROSPECT should next be compared to top hits from the other threading programs such as GenTHREADER and the fold recognition program SAM-T98 (Karplus et al. 1998). If hits from PROSPECT present in all three programs with high rank, then the certainty of a good result becomes very high and the best of those templates should be chosen. Because it is possible that templates returned by PROSPECT are closely related to templates returned by other threading programs, equivalent hits in different threading programs may not be apparent. For example 1gk9 and 1fxv are both structures for penicillin acylase. A list of structurally related templates can be pulled up by entering the PROSPECT

template identification in FSSP. The list of neighbours can then be searched for top hits from other threading programs. In most cases the final templates can be chosen based on common threading results from all threading programs.

b) Model generation: Once the best threading template is chosen, the three dimensional structural prediction for the query protein can be generated. The atomic structures were generated using MODELLER (Sali et al. 1993). It is important to point out that the three dimensional protein structures generated for the protein query are not exactly the same as the three dimensional structures of the protein templates used (Figure 4). The query may contain some residues that do not conform to the secondary structure of the template, and therefore those amino acids will cause some changes in the backbone structure of the three dimensional prediction for the query. As a result of this the predictions for the query are unique. If the template and query are superimposed they should be very similar but they will not be perfectly aligned. Because the majority of the three dimensional structure of the template is conserved in the query, it is possible that the query has similar functions to the template. In other words, if the template is an ATPase and the query has a similar ATPase domain, the query could also have ATPase activity and this should be investigated further using alternative computational methods. In this way the three dimensional predictions can imply possible functions for query

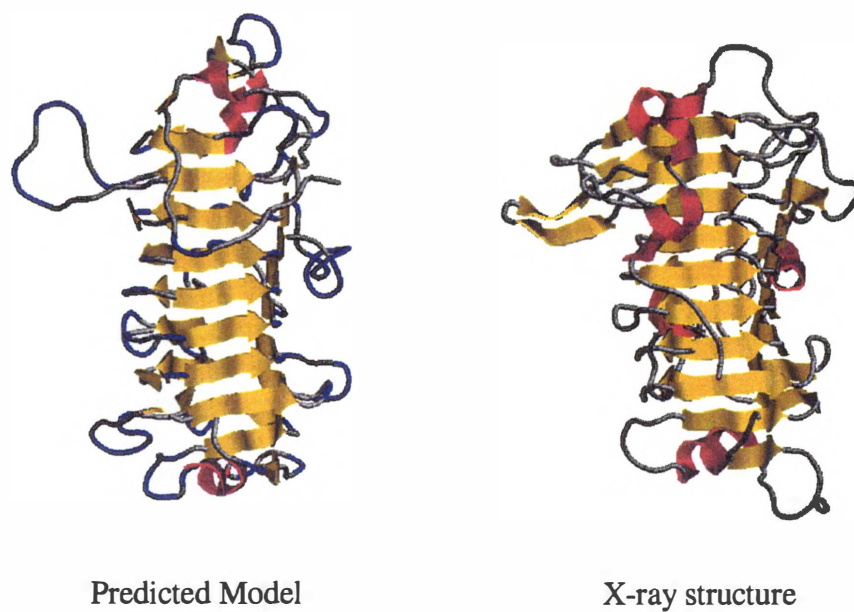


Figure 4

The comparison of two structures. The one on the right of Pectin Methyltransferase: *E. chrysanthemi* was solved by X-ray crystallography. The structure on the left is the model that was generated for a query protein based on the backbone structure of the solved structure on the right.

proteins. Therefore the predicted structure is a powerful tool for functional inference.

4. Literature Search: Once the threading template is chosen literature searches provide information such as possible motifs (signature sequences), adjacent genes, subcellular location, and other information that is known about the template protein. This information will help the user to direct the searches using the successive steps in the method to more easily identify possible functions of the query.

5. PSI-BLAST: The sequence based National Center for Biotechnology Information (NCBI) tool protein PSI-BLAST is much more sensitive than the original protein BLAST to remote homologs that may have less sequence similarity. PSI-BLAST is based on multiple sequence alignments. As was noted in the introduction secondary structure is more highly conserved than sequence. As a result, identification of homologs is not always possible based on conserved sequence alone. PSI-BLAST uses an iterative search in which sequences found in one round of searching are used to build a score model for the next round of searching. Highly conserved positions receive high scores and weakly conserved positions receive scores near zero. The profile is used to perform a second (etc.) BLAST search and the results of each "iteration" are used to refine the profile. This iterative searching strategy results in increased sensitivity based on two dimensional structure.

Five iterations were run for each query protein using PSI-BLAST. Hits from PSI-Blast were searched for the templates selected using the PROSPECT results or any hits with similar function. Literature searches were done on top PSI-BLAST hits and functions of these were noted. If the PSI-blast hits either included the template or proteins related by family or function then the evidence was considered to support the choice of the template from the threading program.

Results from PSI-BLAST searches sometimes include links to the SWISS-PROT database (Bairoch 1999). The SWISS-PROT database provides functional annotations for most of the genes in that database. The database is updated frequently so the information is typically accurate. PSI-BLAST hits that returned values below threshold were not considered significant. In other words, there were no hits with sequence similarities similar enough that they could not have been expected to occur by random chance.

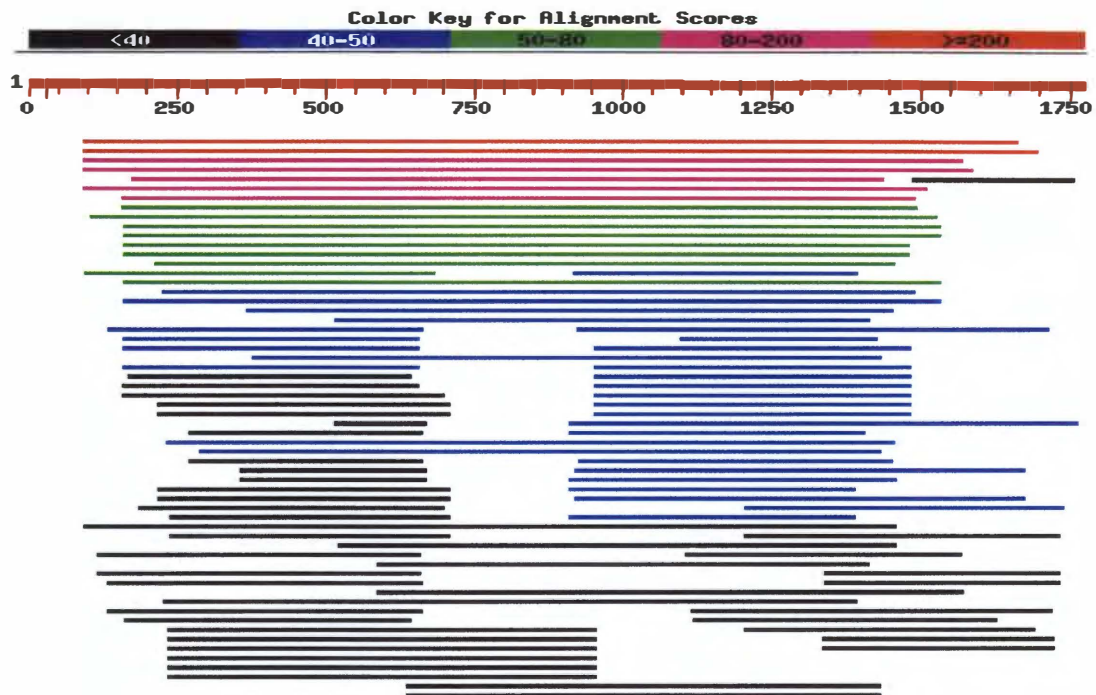
6. Domain parsing: When there is no alignment template for the entire query protein sequence present in the databases it is advantageous to divide the protein into domains and submit those partial sequences to the threading program. Results of a PSI-BLAST search can be used to divide the protein into probable domains (Domain Parsing). When two regions of a protein show sequence homology to other proteins and the area between them is a low complexity region then this region could be considered to be a boundary between two domains. If in Blast or PSI-

BLAST there are a lot of X's showing low sequence homology then it indicates it is possible a low complexity region. There are automated web tools such as the program ProDom (Corpet et al. 2000) which will identify possible domain divisions within a protein sequence (Figure 5).

7. Motif Search: The web based tool MOTIF, (MOTIF) which includes ProSite (Atwood et al. 1999), BLOCKS (Henikoff 1991), ProDom (Corpet et al. 2000), Pfam (Bateman et al. 2002), and PRINTS (Atwood et al. 2002), allows the user to search for identifying motifs (signature sequences) using a query amino acid sequence. Functional domains of proteins sometimes contain a number of perfectly spaced highly conserved amino acids, surrounded by a group of amino acids that have similar physical characteristics (hydrophobicity, size, acidity) across species. 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. It is possible to suggest function with a high degree of accuracy for a protein that contains a conserved signature sequence motif within an expected domain. Motif searches allow identification of signature sequences within a protein.

GeneFIND (Wu et al. 1996) can be used to search for sequence motifs and a family classification. GeneFIND provides an on-line interface that allows direct submission of the query sequence and returns a gene

PSI-BLAST



Conserved Domain search results



Figure 5

The above PSI-BLAST result for gene 4840 shows two main areas of alignment and an obvious region of low complexity between them. The protein should be divided at residue 200 and at residue 400. The three sections should then be submitted for threading separately. PSI-BLAST will also perform a Conserved Domain (CD) search which can be used to identify possible domains. The above results are for the same protein.

family classification of query sequences. GeneFIND uses the ProClass (Atwood et al. 1999) database to return ProSite predictions.

8. TMD search: SOSUI (Hirokawa et al. 1998) allows the prediction of hydrophobic regions of the protein which may either be trans-membrane domains or signal sequence residues. The query is submitted via a user friendly web interface and results are returned within a matter of minutes. The results from this tool include average of hydrophobicity, a prediction of the number of trans-membrane domains and their locations, a hydrophobicity profile, and helical wheel diagrams for predicted trans-membrane segments. These features should correspond to helical regions within the three dimensional structure of the query protein.

9. Signal Sequence: SignalP (Henrik et al. 1997) is 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. SignalP allows the direct submission of the query sequence and returns the result via a user friendly web interface within a matter of minutes.

10. Subcellular Localization: PSORT (Nakai 2000) is a computer program for the prediction of protein localization sites in cells. The amino acid sequence can be submitted directly via the internet interface and the

result is returned within a matter of minutes. The result gives the most probable subcellular localization for the protein of interest.

11. Genomic Sequence analysis: The final step in the protocol was genomic sequence analysis using the *Shewanella* files (Larimer et al. 2002) 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. These types of genes need to be expressed together if the stoichiometry of the components involved in the complex or pathway is important for proper function. In either of these two scenarios it is advantageous for the transcription of these related genes to be co-regulated on the same operon. Often when genes are in an operon, 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 considered.

If the adjacent genes appeared to be on 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 on an operon can be informative about the functions of surrounding genes. Promoter regions were inspected for TATA boxes, *Sigma* 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 a 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.

12. Other suggested tools: There are a number of other useful tools that may be useful, such as Coils (Lupas et al. 1991) which was used for the annotation of gene 4840. An excellent frequently updated resource for finding links to these tools can be found at the Oak Ridge National Lab, Life Sciences Division, Protein informatics groups Resources of Protein modelling website <http://compbio.ornl.gov/structure/resource/>.

3 Results and Conclusions

The three genes from *S. oneidensis* MR-1 were each analysed using the steps from the flow chart (Figure 6). Each step was described in detail in the Method section. The PROSPECT results for each of the hypothetical genes are available at the following web interfaces:

<http://compbio.ornl.gov/~csecrest/4840-whole> (4840)

<http://compbio.ornl.gov/~csecrest/4719-whole> (4719)

<http://compbio.ornl.gov/~csecrest/3954-whole> (3954)

3.1 Gene 4840

1. Secondary Structure Prediction: Gene 4840 was down-regulated 1.6 fold in the *fur* mutant, FUR1. The hypothetical gene translated to 592 amino acid residues. The amino acid sequence was submitted to PSIPred and PHD for secondary structural prediction (Figure 7). The secondary structural prediction from PHD was used to submit protein 4840 to PROSPECT. The protein sequence was also submitted to PSIPred*/GenTHREADER and SAM.

2. Threading: Threading results in PROSPECT returned templates of proteins, that had similar secondary structure with the query protein. The templates from the PROSPECT threading results that were ranked higher than the chosen template 1dg3a, did not include the entire query protein sequence. Those truncated hits such as tropomyosin only aligned with a

Method for Sequence-structure-function annotation

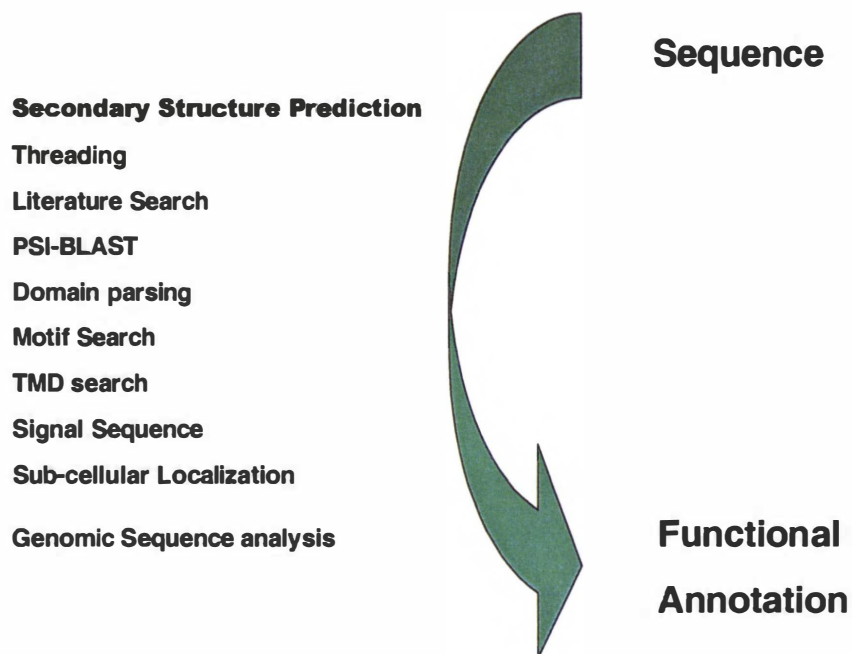


Figure 6
Flowchart for the sequence-structure-function method described.

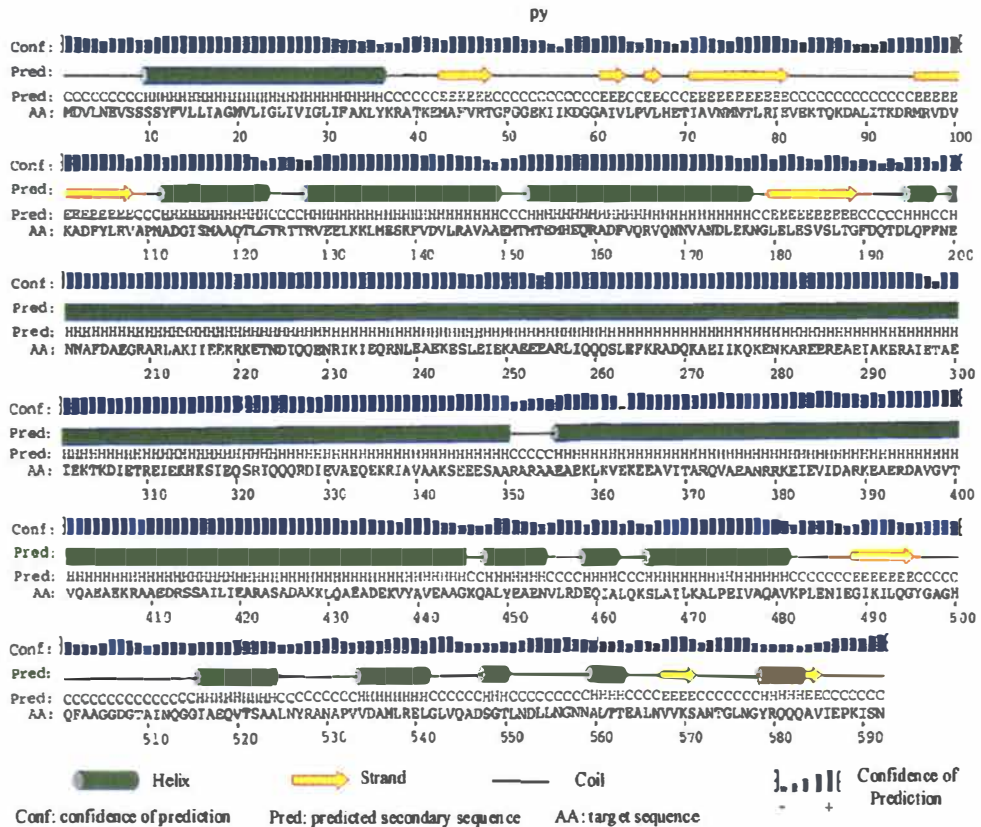


Figure 7

The secondary structural prediction for gene 4840. It is apparent from the graphic returned by PSIPred that this protein has a large percentage of helix. The protein may have a signal sequence since there is a helix of around 16 to 20 residues near the N terminus. The protein also has long stretches of helix in the C-terminal half which may indicate a coiled coil PSIPRED (McGuffin et al. 2000).

small region and were therefore ignored. The rank of the full alignments were lower because it is more difficult to align a template with the entire query sequence maintaining low energetics and a more significant secondary structure and sequence similarity. The alignment for the template 1dg3a encompassed almost the entire length of the hypothetical query protein, from residue 50 to 490. Prospector results for the alignment of 1dg3a with the hypothetical 4840 had a low Mutation energy, a low ssfit value, a good rank, packing index, and reasonable reliability. 1dg3a was also returned as a favourable template by the programs GenTHREADER and SAM. Because of the consensus in the three computational tools, 1dg3a was selected as the template for gene 4840, and a unique three dimensional model (Figure 8) was created for gene 4840 using MODELLER (Sali et al. 1993). The N-terminal trans-membrane domain was removed before threading.

3. 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 regions, a pleckstrin homology domain: 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.

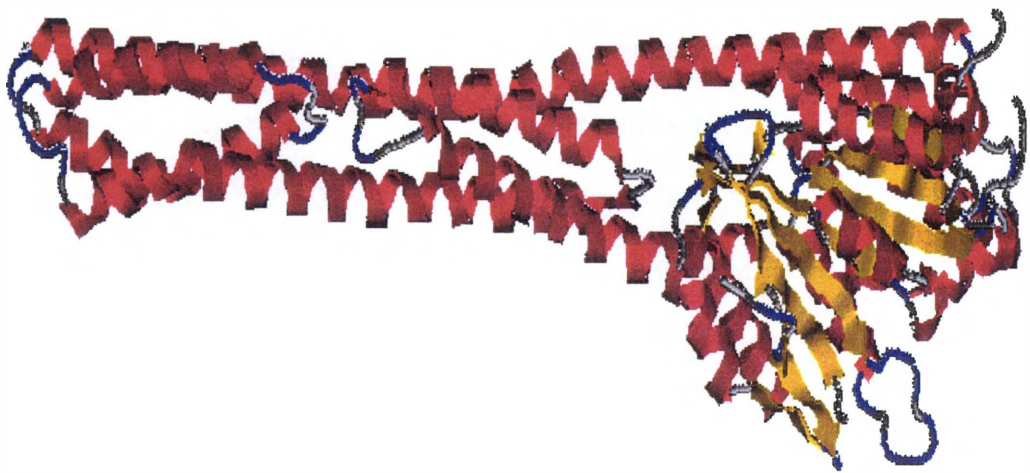


Figure 8
Three dimensional Model for hypothetical protein 4840, MODELLER (Sali et al. 1993).

1dg3a has a yeast homolog Vps1p. It has been studied, and it has been suggest that this homolog could be a part of the protein secretory pathway (Rothman et al. 1990). In an investigation of Eucaryotic heat shock chaperone proteins, the Bag interaction with the heat shock ATPase is described. This Bag-1 nucleotide exchange factor increases interaction between the ATPase and chaperone coiled coil domains. This study focuses on the presentation of the interaction domain (Bag-1) that forms the interface between the ATPase and coiled coil domains. The Bag domain contains a highly conserved sequence of residues (Sondermann et al. 2001). On visual examination of the region between the ATP/GTPase and coiled coil domains in the query protein 4840, the Bag conserved sequence was found (Figure 9). This strongly suggests that interaction between the two domains ATP/GTPase and coiled coil may occur in the hypothetical protein.

4. PSI-BLAST: The results of this search were primarily below threshold. The CD search results indicated the query aligned with three possible families of proteins. The homologs suggested by the CD search were Band 7, PHB and caldesmon (Figure 5). Band 7 and PHB proteins are both similar to the query in sub-cellular localization and the probability of having a single TMD anchor near the N terminus. caldesmon also has similarities to the template protein 1dg3a. It has a high helical content, it is localized to the cytoplasmic membrane. caldesmon is involved in exocytosis.

The signature sequence shown falls between the ATP/GTPase and the coiled coil domains.

This is the signature of the Bag-1 domain which is the interaction domain for the ATPase and coiled coil in the protein family containing DnaK and Hsc70

MDVLNEVSSS SYFVLLIAGM VLIGLIVIGL IFAKLYKRAT KEMAFVRTGFGGEKIIKDGG AIVLPVLHET IAVNMNTLRI
EVEKTQKDAL ITKDRMRVDVKADFYLRVAP NADGISMAAQ TLGTRTTRVE ELKKLMESKF VDVLRVAAEMTMTEM
HEQR ADFVQVRVQNN VANDLEKN GL ELESVSLTGF DQTDLQFFNE NNAFDAEGRA
RLAKIIEEK R KETNDIQEN RIKIEQRNLE AEKESLEIEK AEEEEARIQQ QSLEFKRADQ KAEIHK
QKENKAREEREAEI AKERAIETAE IEKTKDIETR EIEKHKSIEQ SRIQQORDIE VAEQEKRIAV AAKSEESAA RARAAEAEK
L KVEKEEAVIT ARQVAEANRR KEIEVIDARK EAERDAVGVTQAEAEKRAA EDRSSAILIE ARASADAKKL QAEADEKVY
A VEAAGKQALY EAENVLRDEQ IALQKSLAIL KALPEIVAQA VKPLENIEGKILQGYGAGHQFAAGGDGTA INQGGIAEQV
TSAALNYRAN APVVDAMLRE LGLVQADSGT LNDLLNGNNA LTTEALNVVK SANTGLNGYR QQQAVIEPKI SN

Bag-1M	QGFLPKDLQAEALCK-LDRRVKATIEQFMKILEEID-----TLILPEN-FKDSRLKRKGLVKKVQAFLECDTVEQNICQ
Bag-3	-----GKK-TDKKYLMIIEYLTKEALLD-----S-VDPEG-RADVQARRDGVVRKVQTILEKLDQKAIDVPG
Bag-4	-----VGKK-TDKAYWLLLEMLTKELLELD-----S-VETGG-QDSVRQARKEAVCKIQAILKLEKKGL----
Bag-5	-----GEVLSFDGNNR-TDKNYIRLELLTKQLLALD-----A-VDPQG-EEKCKAARKQAVRLAQNLSYLDLKSDEWEY
BAG-1	RGFLEKPKQVEMGKK-LEKKVKYFNEEAERHLETLDGMNIITETTPENQAKRNREKRKTLVNGIQTLNQNDAALLRLQE

4840 VQRVQNNVANDLEKNGLELESVSLTGFDFNAFDAEGRARLAKIIEEKRKETNDIQENRIKIEQRNLE AE

Figure 9

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). The conserved signature is present in the hypothetical protein 4840.

5. Domain parsing: Probable domains were identified using results of the PSI-BLAST search. The search result showed that the protein has at least two possible domains. The PSI-BLAST CD (conserved domain) search was used to divide the protein into three possible domains (Figure 5). The threading results using the divided protein returned no results that were significantly different or better than the hit 1dg3a. All of the domains in 1dg3a were present in the query. The two domains suggested by the CD search are an ATP/GTPase and a large coiled coil domain.

6. Motif Search: A Motif search was completed for 4840 and the BLOCKS and pfam results included heat shock protein and Band 7. PRINTS contained no useful information. Using the web tool GeneFIND, a section of the query protein sequence was identified as a ProSite (*PS01036*) signature pattern related to ATP binding in DNAK (Figure 10). The GeneFIND ProClass best-Matched Family Members hit list also linked the query to DNAK. DNAK is the primary heat shock protein found in *E. coli*. All heat shock-70 proteins bind 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.

7. TMD search: The results of SOSUI suggest that the protein has one (TMD) trans-membrane helix from residue 9 to 31. When there is only

ProSite signature pattern PS01036

```

AC   PS01036

ID   HSP70_3; PATTERN.
DE   Heat shock hsp70 proteins family signature 3.
PA   [LIVMY]-x-[LIVMF]-x-G-G-x-[ST]-x-[LIVM]-P-x-
     [LIVM]-x-[D
PA   EQKRSTA]
DNAK_BACSU+S09500   PST   306   VILVGGSTRIPAVQE
HS7C_PETHY+S03250   PST   340   VVLVGGSTRIPKVQQ
DNAK_BACST+JC4738   PST   305   IILVGGSTRIPAVQE
DNAK_LEPIN+         PSN   177   VILVGGGLRIPAVQE
Query 4840          GFN    55   IIKDGGAIVLPVLHE
                                ::  **.  :*  :::

```

Figure 10
The Prosite alignment for the ATP binding signature in 4840, ProSite (Atwood et al. 1999).

one hydrophobic region and it is near the N terminus it should be evaluated to determine if it could be a signal sequence.

8. 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.

9. Subcellular Localization: The results from PSORT search 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 domain. 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.

10. Coiled Coil: 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 (Figure 11).

11. Genomic Sequence analysis: The *Shewanella* files database and Artemis were used to assess whether the gene 4840 was possibly on

Coiled-coil

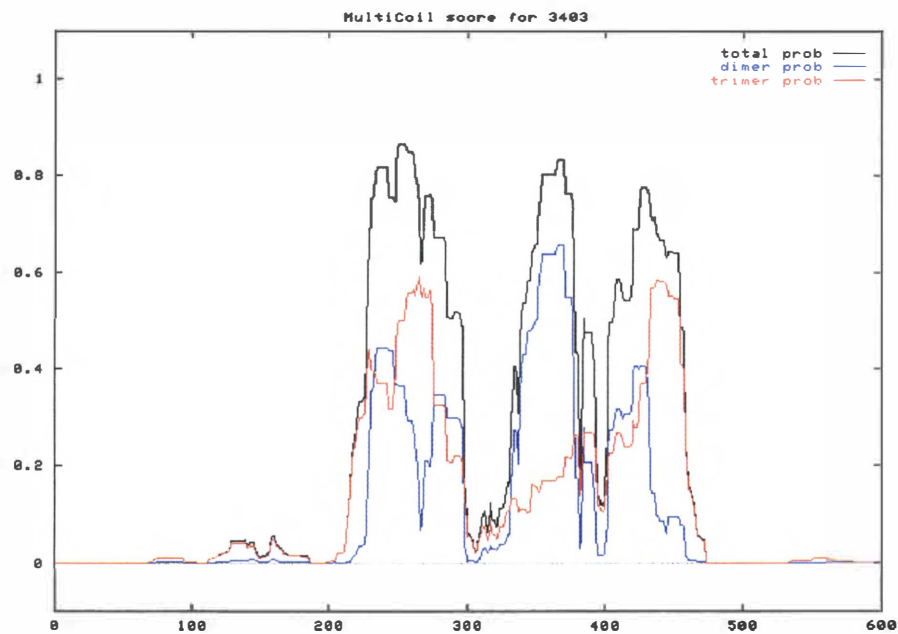


Figure 11
The graphical representation from the web tool coils (Lupas et al. 1991) showing the coiled coil region from residue 200 to about 480.

the same operon as other identified genes. The gene identified in the *Shewanella* files database as 4840 was separated from the gene upstream 4839 by only thirty 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. 300 base pairs of the intergenic region upstream of hypothetical gene 4839 were inspected for possible *fur* and *sigma* 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 be involved in folding and insertion of the protein product of 4839 or associated proteins.

12. 4840 Functional Annotation: Based on the results of the evaluation of gene 4840, the functional annotation suggested by the initial threading template is supported. The hypothetical protein 4840 can be divided into 4 domains based on the findings from computational searches (Figure 12).

All the evidence suggests that the hypothetical protein may have homology to the Dynamin family of proteins and or Heat shock proteins.

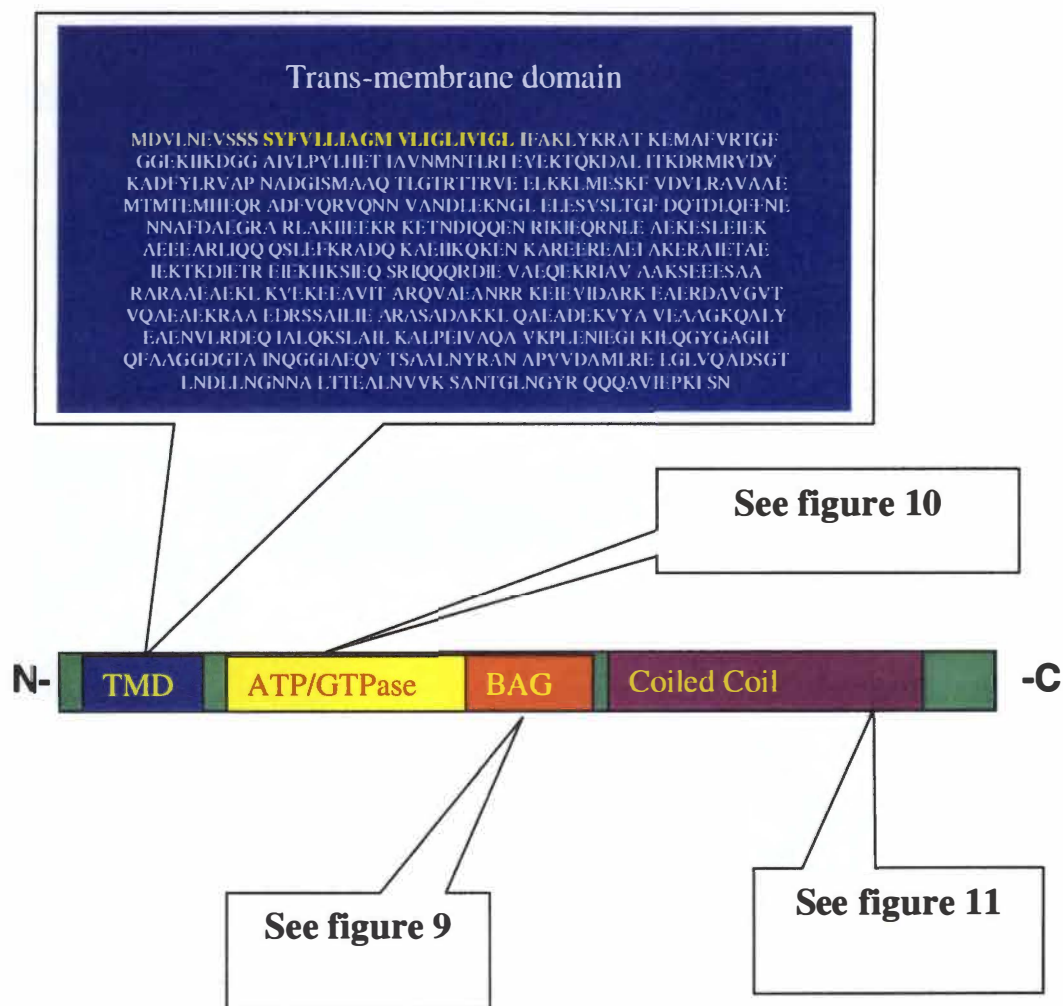


Figure 12

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

This protein may be involved in protein folding (chaperone), signalling, GTP/ATP binding, and or protein translocation.

3.2 Gene 4719

1. Secondary Structure Prediction: Gene 4719 was up-regulated approximately three fold in the *fur* mutant. The hypothetical gene translated to 274 amino acids. The amino acid sequence was submitted to PSIPred and PHD for secondary structural prediction (Figure 13). The secondary structural prediction from PHD was used to submit protein 4719 to PROSPECT.

2. Threading: The hypothetical protein 4719 shows good alignment with residues 33 to 286 of the PROSPECT template 3pvaa. The ranking of the alignment with 3pvaa, which was 3, was good. The PROSPECT results returned a low mutation energy, a low ssfit value, a c-ndx of 1.87, and a reliability for this protein at 0.410. When all other values are within limits and the alignment is good a reliability of between 0.1 and 0.7 may be accepted. The template 3pvaa was not returned by GenTHREADER and SAM, however when the PROSPECT hits were searched using FSSP the template 3pvaa was found in the structural neighbours list of some of the top hits for these other programs. The three dimensional structure of the hypothetical protein 4719 was created using Modeller (Figure 14).

3. Literature Search: The functional part of the template penicillin acylase protein lies within the region of alignment with the query protein.

4719

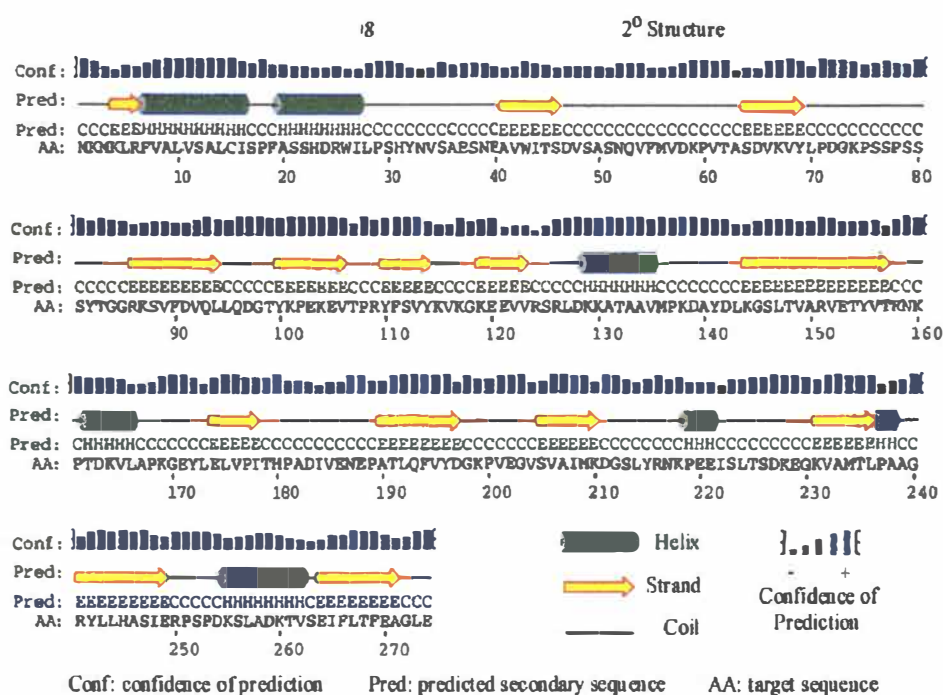


Figure 13

The predicted secondary structure of protein 4719. The secondary structure shows a predominance of beta sheet, PSIPRED (McGuffin et al. 2000).

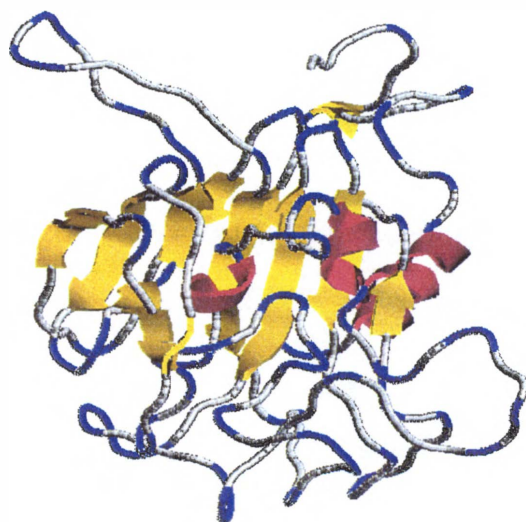


Figure 14
The three dimensional structure generated for protein 4719, MODELLER
(Sali et al. 1993).

The sub-cellular location of penicillin acylase is periplasmic. 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. The members do not have any real sequence similarities (signature). The Ntn hydrolases fold to form an oxyanion hole which contains an N terminal catalytic nucleophilic 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).

4. PSI-BLAST: The protein sequence was used to complete a PSI-BLAST search. The CD search results returned no hits. The results of PSI-BLAST were primarily below threshold and therefore unremarkable.

5. Domain parsing: Results of a PSI-BLAST search did not suggest that this short protein should be divided into separate domains. The CD search of PSI-BLAST returned no hits and none of the hits that could be used to divide the protein were above threshold.

6. Motif Search: The Motif search returned few results. The only remarkable signature for protein 4719 from BLOCKS is a sequence for an amidation site. Some proteins which are found in the periplasmic space are attached to the cell wall peptidoglycan by an amide bond. ProSite, pfam and PRINTS contained no useful information.

7. TMD search: SOSUI predicted the protein to be a soluble protein.

The hydrophobicity plot does indicate that the N terminal signal sequence is very hydrophobic and may be a signal.

8. Signal Sequence: SignalP predicts the signal peptide to be between residues 4 and 23. The signal is predicted to be cleaved between residues 23 and 24.

9. Subcellular Localization: PSORT predicts the protein is localized to the periplasmic space. These results are consistent with the sub-cellular localization and presence of signal on the template protein 3pvaa penicillin acylase.

10. Genomic Sequence analysis: The *Shewanella* files 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 (Figure 15). 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 was inspected for possible *fur* or *sigma* binding sites. There is a possible

Operon structure

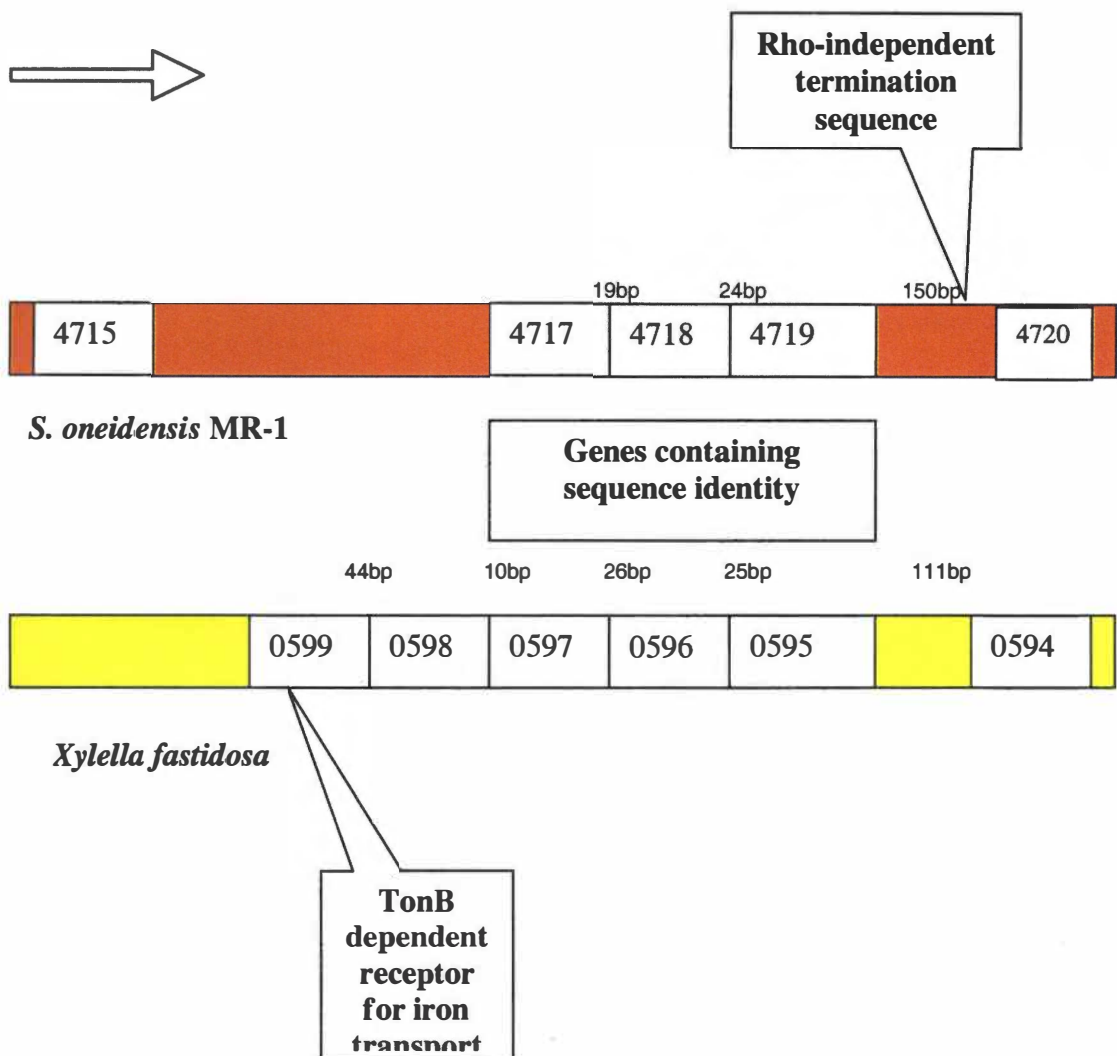


Figure 15

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

(probable) *fur* binding site 118 bp upstream of the ATG transcriptional start site for gene 4717. This possible *fur* binding site is shown below.

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 and iron metabolism. The fact that these three genes appear to be on the same operon also indicates that their functions may be related in some manner.

Proteins 4717 and 4718 were threaded and functional annotation was also given for these. Results were as follows: Gene 4717 has three trans-membrane domains 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. If these genes are on the same operon because they arose by gene duplication they should share some conserved sequence. These three genes do not share conserved sequence.

The *Shewanella* files database was used to search for homologs for the three related proteins. The three proteins each were related with high sequence similarity to *Xylella fastidiosa* proteins XF0597, XF0596 and XF0595. The operon structure of these proteins was inspected using the genome channel. These three *Xylella fastidiosa* proteins were found on the genome in the same order and orientation as the *S. oneidensis* MR-1 proteins. The *Xylella fastidiosa* genes presented on a possible operon with

an annotated gene. (Figure 15) The upstream genes xf0598 and xf0599 in *Xylella fastidiosa* appeared to be on the same operon as the three homologs to the *S. oneidensis* MR-1 genes. The functional annotation of gene xf0599 is given as: 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 cell's 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 bacteriophages, 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 like FhuA that interact with TonB and transport siderophores with bound iron into the cell also bind and transport vitamin B12 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 .

11. 4719 Functional Annotation: Based on the results of the evaluation of gene 4719, the functional annotation suggested by the initial 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.

3.3 Gene 3954

1. Secondary Structure Prediction: Gene 3954 of *S. oneidensis* MR-1 was up-regulated approximately three fold in the *fur* mutant. The hypothetical gene translated to 274 amino acids. The PSIPred postscript secondary structural prediction was reviewed (Figure 16). The secondary structural prediction from PHD was used to submit protein 3954 to PROSPECT.

2. Threading: The hypothetical protein 3954 shows alignment with residues 10 to 270 of the PROSPECT template 1atg. Both the rank of the alignment with 1atg was 6. The PROSPECT results returned a low mutation energy, a low ssfit value, and a c-ndx of 1.73. The PROSPECT template 1atg was also returned by the threading program GenTHREADER and SAM-T98. The certainty for GenTHREADER was 1.0, >99%. These results in

3954

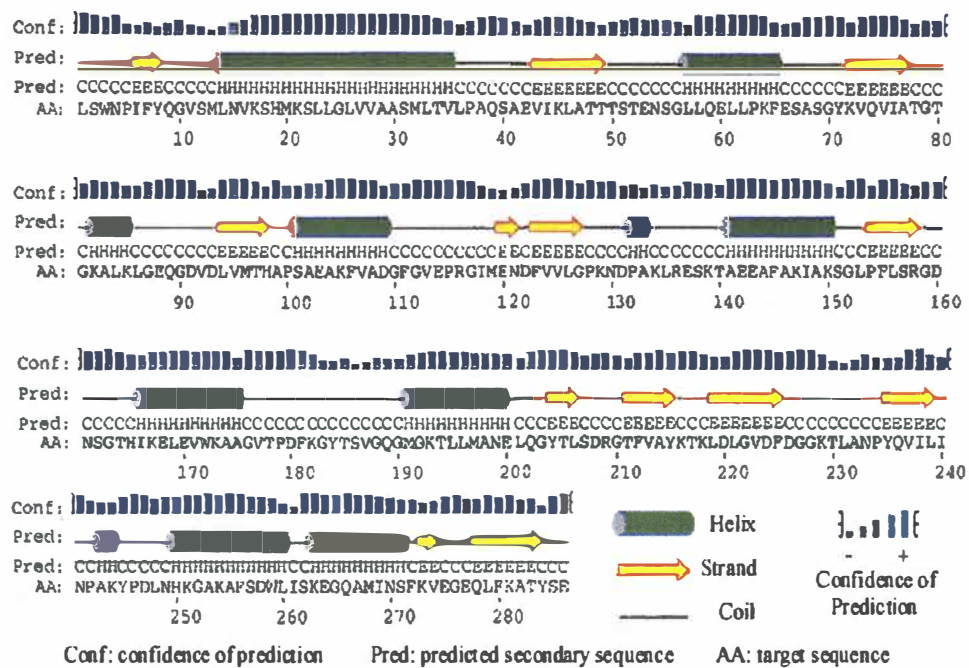


Figure 16
Secondary structure prediction for gene 3954, PSIPRED (McGuffin et al. 2000).

combination were enough to allow the acceptance of the template 1atg. The three dimensional structure of the hypothetical protein 3954 was created using Modeller (Figure 17).

3. 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. 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 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. And FhuC is provides

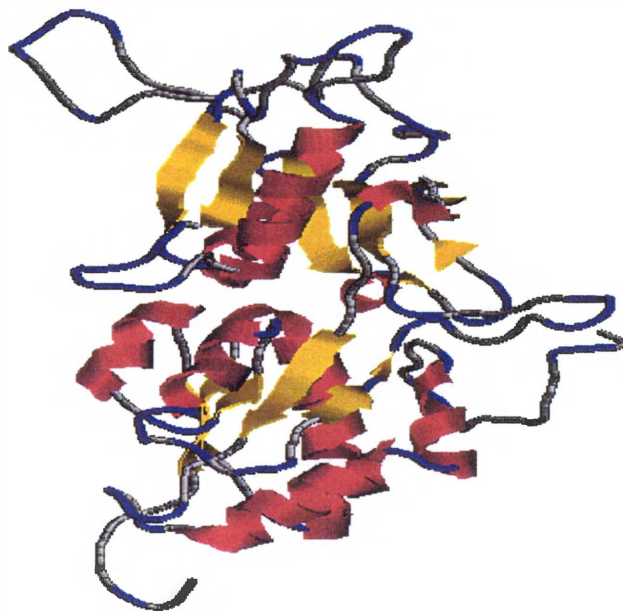


Figure 17
The three dimensional structure generated for the protein encoded by Gene 3954, MODELLER (Sali et al. 1993).

energy via hydrolysis of ATP and is localized to the inner side of the cytoplasmic membrane. The ATPase domain is very 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 but not always found on the same operon (Self, Grunden 2001). The genes are typically found in the following order: periplasmic Binding Protein, inner membrane protein, 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, Grunden 2001). Genes 3954, 3955 and 3956 have overlapping start and stop codons.

4. PSI-BLAST: The protein sequence was used to complete a PSI-BLAST search. The results of PSI_BLAST were primarily below threshold and unremarkable. This original annotation process was begun for gene 3954 in the spring of 2001. The PSI_BLAST search was repeated in the spring of 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 one year prior.

5. Domain parsing: Results of a PSI-BLAST search did not suggest that this short protein should be divided into separate domains. The CD search of PSI-BLAST returned no hits, and none of the hits that could be used to divide the protein were above threshold.

6. Motif Search: The Motif search returned few results. There were no remarkable signatures for protein 3954 from BLOCKS. ProSite, pfam and PRINTS contained no useful information.

7. TMD search: The protein was predicted to be soluble by SOSUI.

8. Signal Sequence: A cleavable N-terminal signal sequence from residue 5 to 27 was predicted by SignalP.

9. Subcellular Localization: The subcellular localization of gene 3954 given by PSORT was periplasmic.

10. Genomic Sequence analysis: The *Shewanella* files 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 (Figure 18).

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 was searched for the presence of *fur* or *sigma* binding sites. A possible *fur* binding site was found approximately 125 base pairs upstream from the transcriptional start site ATG. This possible *fur* binding site is shown below.

<i>E. coli</i> consensus	=	GAT	AAT	GAT	AAT	CAT	---	TATC
<i>S. oneidensis</i> 3954	=	GTG	AAG	GTA A	AAG	CAT	GAA	TATC

The ModABC transport genes were compared to the genes 3954, 3955 and 3956. ModB has 5 trans-membrane domains (Self et al. 2001). Protein 3955 also was identified as having 5 trans-membrane domains by SOSUI. ModB contains three ABC permease signature sequences (Self et al. 2001). The hypothetical protein 3955 also contains these signatures .

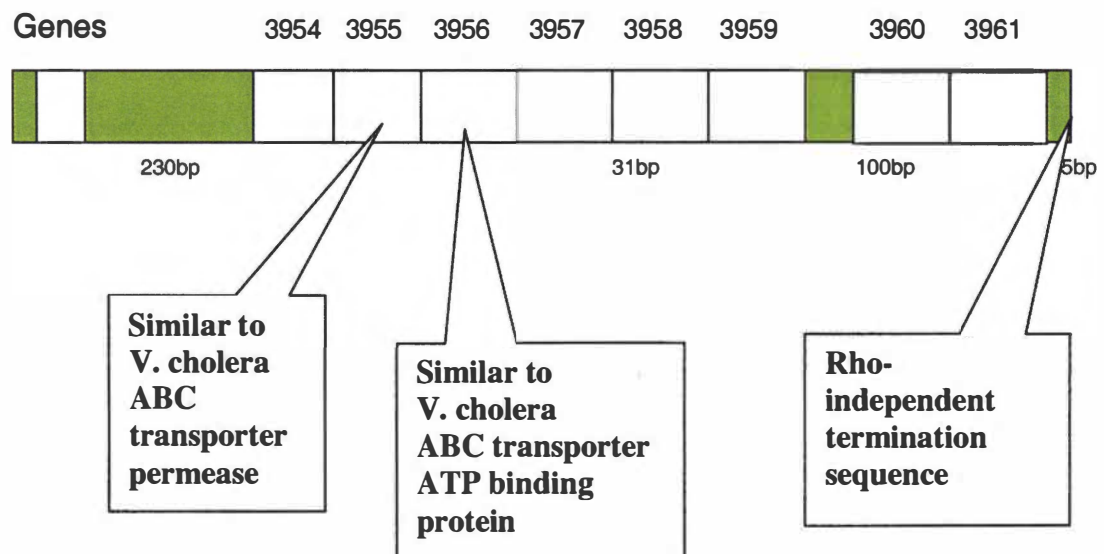


Figure 18

The operon of *Shewanella* containing the genes 3954, 3955, 3956 etc. Other genes which appear to be on the same operon are functionally annotated as possible homologs to ABC transporter components.

ModC contains Walker A and B ATPase signature sequences. These are also present in protein 3956.

12. 4840 Functional Annotation: Based on the results from the computational tools the functional annotation suggested by the original PROSPECT threading result is supported. The gene 3954 is likely to be the periplasmic component of a molybdate ABC transporter.

4 Discussion

The amount of sequence data that is being generated seems to be increasing exponentially, therefore new methods for evaluating genomic sequence data must constantly be developed. One of the primary steps in cataloguing genes of a newly sequenced genome is to identify the hypothetical genes and functionally annotate them. The sequence based methods presently available are only capable of annotating fifty to seventy percent of these hypothetical genes (Skolnick et al. 2000). Because it is desirable to predict possible functions for all or at least most of the hypothetical genes in a newly sequenced genome, methods that will complement the present sequence based methods are needed. Sequence-structure-function methods may fill this gap (Xu et al. 2002).

Sequence-structure-function methods may be able to predict more remote homologs than sequence based methods alone. It is possible for proteins to contain almost no sequence similarities and still have the same three dimensional structure. These proteins may contain only a small signature sequence (Motif) which sequence based methods often cannot identify. In these cases, when structure is more highly conserved than sequence, the conserved signature sequence is a very small part of the protein sequence, and the sequence alignment will not identify them. Signatures within the same family can also contain variation and be difficult to identify. Having a three dimensional structure prediction can suggest possible functions, and it gives the user some idea of what kinds

of signatures to look for in a sequence. Therefore this method, of identifying structures and then signatures that are associated with known functions, may be a first step in elucidation of function of many hypothetical genes.

The simple and elegant method described in this study is easy to understand and use. The web tools suggested give excellent results, and they are freely available on the internet. Most of the suggested tools allow queries to be submitted and returned through user friendly interfaces. Because the results of computational studies are intangible (not experimentally proven), it is often difficult to explain their merit to experimentalists. The description of results obtained for three hypothetical genes using this method is important for two reasons: (1) These descriptions give practical examples of the use of the method and its simplicity. (2) These descriptions give tangible examples of how the combination of results, may in many cases, lend support to the functional annotations given to hypothetical genes using the method.

The primary limitation of this method is that the Protein structure databases contain only template structures of those proteins that have been studied and identified using some method such as crystallography or NMR. Fortunately the number of structures in these databases is constantly growing and as the number of available templates increases so does the accuracy of threading results. It is also likely that there is a limited number of conserved functional domains, since nine out of ten

newly solved structures have a known fold in PDB (Berman 2000). As databases containing known structures grow, threading programs like PROSPECT continue to return more accurate initial predictions from which quick initial functional annotations can be implied. A second limitation is that this method requires some user evaluation, but there is not an automated alternative available at this time. As the results of sequence-structure-function methods improve more automated methods should become available. It is important to remember that 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. This description of a user friendly method, for functional annotation of hypothetical genes, is a timely addition to the solutions for genome analysis.

List Of References

References

- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res.* 1997;25:3389-3402.
- Attwood, T. K., Flower D. R., Lewis A. P., Mabey J. E., Morgan S. R., Scordis P., Selley J., Wright W., *Nucleic Acids Research* 1999; 27:220-225.
- Attwood, T.K., Blythe M., Flower D.R., Gaulton A., Mabey J.E., Maudling N., McGregor L., Mitchell A., Moulton G., Paine K., Scordis P., "PRINTS and PRINTS-S shed light on protein ancestry.", *Nucleic Acids Research*, 2002;30: 1: 239-241.
- Bateman A, Birney E, Cerruti L, Durbin R, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL, The Pfam protein families database. *Nucleic Acids Research* 2002;30:1:276-280
- Bairoch, A., Apweiler R., *Nucleic Acids Research* 1999; 27:49-54.
- Berman H. M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The ProteinData Bank. *Nucleic Acids Research* 2002;28:235-242.
- Brannigan J., Dodson G., Wilson K., Ntn Hydrolases. York Structural Biology Laboratory 2002. Structural Biology Laboratory Dept. of Chemistry University of York Heslington YORK UK YO10 5DD.
<http://www.ysbl.york.ac.uk/>.
- Braun J., Iron uptake mechanisms and their regulation in pathogenic bacteria. *Int. J. Med Microbiol.* 2001; 291:67-79.
- Brenner, S., A Tour of Structural Genomics. *Nature Reviews, Genetics* 2001; Vol 2:801-809.
- Corpet F, Servant F, Gouzy J, Kahn D., ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.* 2000;28:267-269.
- Cuff J. A., Barton G. J. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction, *PROTEINS: Structure, Function and Genetics.* 1999;34:508-519
- Danino D. Hinshaw J. E., Dynamin family of mechanoenzymes. *Current Opinion in Cell Biology* 2001; 13:454-460.
- Dassa E., Bouige P., The ABC of ABC's: a phylogenetic and functional classification of ABC systems in living organisms. *Res. Microbiology* 2001; 152:211-229.
- Dobbin, P. S., Butt, J. N., Powell, A. K., Reid, G. A., Richardson, D. J., Characterization of a flavocytochrome that is induced during the anaerobic respiration of Fe³⁺ by *Shewanella frigidimarina* NCIMB400. *Journal of Biochemistry* 1999; 342:439-448.
- Dubrac S., Tauati D., Fur Positive Regulation of Iron Superoxide Dismutase in *Escherichia coli*: Functional Analysis of the *sodB* Promoter. *Journal of Bacteriology* 2000; 182:13:3802-3808.
- Enz S., Mahren S., Stroeder U. H., Braun V., Surface Signaling in Ferric Citrate Transport Gene Induction: Interaction of the FecA, FecR, and FecI Regulatory Proteins. *Journal of Bacteriology* 2000; 182:3:637-646.
- Ermolaeva M. D., Khalak H. G., White O., Smith H. O., Salzberg S. L., Prediction of Transcription Terminators in Bacterial Genomes. *J Mol Biol* 2000;301:1:27-33.
- Escolar L., Perez-Martin J., DeLorenzo V., Opening the Iron Box: Transcriptional Metalloregulation by the Fur Protein. *Journal of Bacteriology* 1999; 181:20:6223-6229.
- Fetrow, J. S., Siew, N., Di Gennaro, M. M., Dyson, H. J., Skolnick, J., Genomic-scale comparison of sequence- and structure-based methods of function prediction: Does structure provide additional insight?. *Protein Science* 2001; 10:1005-1014.

- Godroy P., Ramos-Gonzalez M. I., Ramos J. L., Involvement of the TonB System in Tolerance to Solvents and Drugs in *Pseudomonas putida* DOT-TIE. *Journal of Bacteriology* 2001; 183:18:5285-5292.
- Henikoff S., Henikoff J. G., "Automated assembly of protein blocks for database searching", *NAR* 1991;19:6565-6572.
- Henrik N., Engelbrecht J., Brunak S., Gunnar von Heijne, Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering* 1997;10:1-6.
- Hirokawa T., Seah B.-C., Mitaku S., SOSUI: Classification and SecondaryStructure Prediction System for Membrane Proteins., *Bioinformatics* (formerly known as CABIOS) 1998 May; 14:4: 378-379.
- Holm L., Sander C., Mapping the protein universe, *Science* 1996;273:595-602.
- Jones, D.T. GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J.Mol. Biol.* 1999;287, 797-815.
- Karplus K., Barrett C., Hughey R., "Hidden Markov Models for Detecting Remote Protein Homologies", *Bioinformatics* 1998; 14:10:846-856.
- Kieft, T. L., Fredrickson, J. K., Onstott, T. C., Gorby, Y. A., Kostandarithes, H. M., Bailey, T. J., Kennedy, D. W., Li, S. W., Plymale, A. E., Spadoni, C. M., Gray, M. S., Dissimilatory Reduction of Fe(III) and Other Electron Acceptors by a *Thermus* Isolate. *Applied and Environmental Microbiology* 1999; Mar:1214-1221.
- Koster W., ABC transporter-mediated uptake of iron, siderophores, heme and vitamin B₁₂. *Res. Microbiology* 2001; 152:291-301.
- Larimer F. , Hauser L., Land M., Draft analysis of the *Shewanella* genome generated through the Microbial annotation pipeline at ORNL 2002; unpublished results. Artemis interface and the Web interface Shewanella files database.
- Lovely, D. R., Phillips, E. J. P., Gorby, Y. A., Landa, E. R., Microbial reduction of uranium. *Nature* 1991; 350:413-416.
- Lovely. D. R, Anaerobes to the Rescue. *Science* 2001; 293:1444-1446.
- Lupas A., Van Dyke M., Stock J., Predicting Coiled Coils from Protein Sequences, *Science* 1991;252:1162-1164.
- McGuffin LJ, Bryson K, Jones, D.T. , The PSIPRED protein structure prediction server. *Bioinformatics* 2000; 16: 404-405.
- MDL Information Systems, CHIME. MDL Information Systems, Inc., San Leandro, California; 1999.
- Sayle R. A., Milner-White E. J., RASMOL:biomolecular graphics for all. *Trends in Biochemical Sciences* 1995; 20:374-376.
- Myers C. R., Myers J. M., Outer Membrane cytochromes of *Shewanella putrefaciens* MR-1: spectral analysis, and purification of the 83-kDa c-type cytochrome. *Biochimica et Biophysica Acta* 1997; 1326:307-318.
- Myers C. R., Nealson K. H., Respiration-Linked Proton Translocation Coupled to Anaerobic Reduction of Manganese(IV) and Iron(III) in *Shewanella putrefaciens* MR-1. *Journal of Bacteriology* 1990; Nov.:6232-6238.
- MOTIF. Institute for Chemical Research 1999. Kyoto University, Kyoto, Japan.
- Nakai, K, Protein sorting signals and prediction of subcellular localization, *Adv. Protein Chem.* 2000;54, 277-344.
- Nealson K. H., Saffarini D., Iron and Manganese in Anaerobic Respiration: Environmental Significance, Physiology, and Regulation. *Annu. Rev Microbiol.* 1994; 48:311-43.

- Pessanha M., Brennan L., Xavier A. V., Cuthbertson P. M., Reid G. A., Chapman S. K., Turner D. L., Salgueiro C. A., NMR structure of the haem core of a novel tetrahaem cytochrome isolated from *Shewanella firgdimarina*: identification of the haem-specific axial ligands and order of oxidation
- Rost B., PHD: predicting one-dimensional protein structure by profile based neural networks. Meth. in Enzym., 1996;266: 525-539.
- Rost B., Sander C., Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol., 1993, 232, 584-599.
- Rothman JH, Raymond CK, Gilbert T, O'Hara PJ, Stevens TH. A putative GTP binding protein homologous to interferon-inducible Mx proteins performs an essential function in yeast protein sorting. Cell 1990; 61:1063-74.
- Rutherford K., Parkhill J., Crook J., Horsnell T., Rice P., Rajandream M-A., Barrell B., Artemis: sequence visualisation and annotation. Bioinformatics 2000;16:10: 944-945.
- Saffarini D. A., DiChristina, T. J., Bermudes, D., Neelson, K. H., Anaerobic respiration of *Shewanella putrefaciens* requires both chromosomal and plasmid-borne genes. FEMS Microbiology Letters 1994; 119:271-278.
- Sali A. & T.L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815, 1993.
- Self W. T., Grunden A. M., Hosona A., Shanmugam K. T., Molybdate transport. Res. Microbiol. 2001; 152:311-321.
- Skolnick, J., Fetrow, J. S., From genes to protein structure and function: novel applications of computational approaches in the genomic era. TIBTECH 2000; 18:34-39.
- Sondermann H., Scheufler C., Schneider C., Hohfeld J., Hartl F., Moarefi I., Structure of a Bag/Hsc70 Complex: Convergent Functional Evolution of Hsp70 Nucleotide Exchange Factors. Science 2001; 291:1553-1557.
- Thompson D. K., Beliaev A. S., Giometti C. S., Tollaksen S. L., Khare T., Lies D. P., Neelson K. H., Lim H., Yates III J., Brandt C. C., Tiedje J. M., Zhou J., 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. Applied and Environmental Microbiology 2002; 68:2: 881-892.
- Venkateswaran, K., Moser, D. P., Dollhopf, M. E., Lies, D. P., Saffarini, D. A., MacGregor, B. J., Ringelberg, D. B., White, D. C., Nishijima, M., Sano, H., Burghardt, J., Stackebrandt, E., Neelson, K. H., Polyphasic taxonomy of the genus *Shewanella* and description of *Shewanella oneidensis* sp. nov. International Journal of Systematic Bacteriology 1999; 49:705-724.
- Wu, Cathy, S. Zhao, H. L. Chen, C. J. Lo and J. McLarty. (1996). Motif identification neural design for rapid and sensitive protein family search. CABIOS 1996;12:2:109-118.
- Xu, Y., Xu, D., Protein Threading using PROSPECT: Design and Evaluation. PROTEINS: Structure Function and Genetics 2000; 40:000-000.
- Xu D., Xu Y., Uberbacher E. C., Computational Tools For Protein Modeling. Current Protein and Peptide Science 2000; 1:1-21.
- Xu Y., Dong Xu, and Victor Olman. A Practical Method for Interpretation of Threading Scores: An Application of Neural Network. Statistica Sinica (special issue in bioinformatics) 2002; 12:159-177.
- Xu D., and Ying Xu. Computational Studies of Protein Structure and Function Using Threading Program PROSPECT. In Protein Structure Prediction: Bioinformatic Approach, edited by Igor Tsigelny. International University Line publishers (IUL), La Jolla, CA. Invited publication 2002;5-41.

Web links

The Institute for Genomic Research

TIGR

www.tigr.org

Protein Structure Prediction Server

PSIPRED

GenTHREADER

<http://bioinf.cs.ucl.ac.uk/psipred/>

PredictProtein server

PHD

<http://dodo.cpmc.columbia.edu/predictprotein/>

JPRED

<http://jura.ebi.ac.uk:8888/>

Protein Database

PDB

<http://www.rcsb.org/pdb/>

Fold classification based on Structure-Structure alignment of Proteins FSSP

<http://www2.ebi.ac.uk/dali/fssp/>

Sequence Alignment and Modeling System

SAM-T98

<http://www.cse.ucsc.edu/research/compbio/sam.html/>

Protein Domain Database

ProDom

<http://prodes.toulouse.inra.fr/prodom/doc/prodom.html>

ORNL Resources of Protein Modeling page http://compbio.ornl.gov/protein_group/resource/

Position-Specific Iterated BLAST

PSI-BLAST

<http://www.ncbi.nlm.nih.gov/BLAST/>

SWISS-PROT database

<http://www.ebi.ac.uk/swissprot/>

MOTIF

<http://motif.genome.ad.jp/>

GeneFIND

<http://pir.georgetown.edu/qfserver/Welcome.html>

SOSUI

www.sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html

Signal sequence Prediction

SignalP

<http://www.cbs.dtu.dk/services/SignalP/index.html>

PSORT

<http://psort.nibb.ac.jp/>

TransTerm (TIGR software)

<http://www.tigr.org/software/transterm.html>

Coils

<http://www.york.ac.uk/depts/biol/units/coils/coilcoil.html>

The Genome Channel

<http://compbio.ornl.gov/GC/help/ack.html>

Vita

Christal Yost was born in Dallas Texas in 1961. Christal completed most of her undergraduate coursework at Old Dominion University in Norfolk, Virginia. Christal received a B. S. in Biochemistry cellular and molecular biology from the University of Tennessee in Knoxville in 1999.

Christal is currently completing her M. S. in Bioinformatics at the University of Tennessee.