12-2008

Evolution of the Set of Signal Transduction Proteins in 10 Species of *Shewanella*

Harold Arthur Shanafield
University of Tennessee - Knoxville

To the Graduate Council:

I am submitting herewith a thesis written by Harold Arthur Shanafield entitled "Evolution of the Set of Signal Transduction Proteins in 10 Species of *Shewanella*." I have examined the final electronic 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.

Igor Jouline, Major Professor

We have read this thesis and recommend its acceptance:

Ed Uberbacher, Frank Larimer, Russell Zaretzki

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a thesis written by Harold Arthur Shanafield, III entitled "Evolution of the Set of Signal Transduction Proteins in 10 Species of *Shewanella*." I have examined the final electronic 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.

Igor Jouline, Major Professor

We have read this thesis
and recommend its acceptance:

Ed Uberbacher

Frank Larimer

Russell Zaretzki

Accepted for the Council:

Carolyn R. Hodges, Vice Provost and
Dean of the Graduate School

Evolution of the Set of Signal Transduction Proteins in 10 Species of *Shewanella*

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

Harold Shanafield
December 2008

ACKNOWLEDGEMENTS

The last several years have been some of the most interesting and challenging of my life. These years have been a journey of both intellectual and self enlightenment and I have been fortunate to have crossed paths with so many terrific and helpful people. Many individuals have played an important role in my journey and I'd especially like to thank Igor Jouline for accepting me into his lab and providing me with the opportunity to complete this journey. That lab and the people who have been a part of it including Luke Ulrich, Brian Cantwell, Bhanu Rekapalli, Davi Ortega, and Kirill Borziak were instrumental in helping me to achieve my goals. I'd also like to thank the Genome Science and Technology program run jointly by the Oak Ridge National Lab and the University of Tennessee for their financial support. Finally, I'd like to thank my wonderful wife Janie for all of her love and support. I know my journey would not be nearly as interesting or exciting if she were not by my side.

ABSTRACT

The recent completion of the sequencing of several species of the *Shewanella* genus provides a unique opportunity for comparative genomics studies. We chose the first 10 fully sequenced *Shewanella* genomes to investigate the evolution of signal transduction proteins (ST). ST is a universal and highly regulated system, and as a very well-studied system provides an excellent starting point for investigation. Furthermore, Shewanella have been shown to have a large number of two-component systems and diguanylate cyclases relative to their genome size. In this study we investigate the evolution of signal transduction across several *Shewanella* strains by utilizing a domain-level approach for determining homology and orthology of the parent proteins. Proteins were broken down into their constituent domains and domain sized sequences and compared using a reciprocal best BLAST hit approach to determine homology between all of the species. Analysis of homologous domains and proteins revealed several levels of conservation and a core group of signal transduction proteins common to all members. Further analysis of domain homology provided putative annotations of previously unrecognized sequences and highlighted deficiencies in specific Pfam domain models. Analysis of paralogous domains and proteins showed agreement with 16s rRNA based estimates of evolution, although the position of *S. oneidensis* MR-1 was novel.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER I: Introduction and General Information

Signal Transduction

All living things must sense and adapt to changes in their environment at the cellular level. Response to environmental stimuli plays a critical role in the adaptive fitness of any organism. Consequently many systems have evolved to sense and respond to environmental change. This is especially critical for bacteria, single-celled organisms with few abilities to change their local environment. As a result, bacteria have evolved sensory capabilities to transduce environmental information and affect the proper responses, both genetically and physically. Specific single and multiple protein systems have evolved to perform this function in and around the cell. The processes in which these proteins are involved are broadly classified as Signal Transduction (ST) systems. These processes including sporulation, chemotaxis and virulence are some of the most thoroughly studied ST systems.

ST systems come in several varieties including one-component, two-component, hybrid, and multi-component systems. Two-component systems were the first to be widely recognized and classified. While the role of transcription factors was understood, the larger context within which transcription factors interacted was less clear. Beginning with work done on the nitrogen regulation (NR) system in *Escherichia coli* responsible for controlling the genetic response to nitrogen availability(Ninfa and Magasanik 1986), and then expanding by recognition that the functional protein elements in the NR system were similar to other systems that performed signal transduction functions and prevalent in several other organisms(Nixon, Ronson et al. 1986), a paradigm was born(Stock, Stock et al. 1990).

Two-component systems typically include a membrane bound sensor histidine protein kinase (HPK) and response regulator (RR). The sensor proteins contain a domain evolved to sense the specific environmental characteristic (e.g. ion concentration, redox levels) and a second domain that can autophosphorylate and transfer that phosphoryl group to the response regulator in a reaction catalyzed by the response regulator. Examples of sensor domains include the PAS, GAF and CHASE families. The HPK domains act as dimers while the regulator usually takes the form of a DNA binding protein whose function is controlled through phosphorylation by its paired HPK. An example is the *ompR/envZ* system in which the sensor HPK EnvZ monitors osmolarity and creates a genetic response through the actions of the transcription factor OmpR. Other examples include nitrite metabolism (Nar), nitrogen regulation (Ntr), phosphate regulation (Pho) and citrate uptake and catabolism (Cit) (Hoch and Silhavy 1995).

Initial research into the proteins of the two-component systems began to reveal the modular nature of ST systems. In fact, it was this modularity which led to the recognition of the widespread nature of the two-component system. Nixon et al found large conserved regions in the C-terminal sequences of *Klebsiella pneumoniae ntrB*, *E. coli envZ*, *cpxA*, and *phoR*, and *Agrobacterium tumefaciens* *virA*. This conservation was also found in *E. coli cheA* (Nixon, Ronson et al. 1986). These regions were later named the HisKA (Bilwes et al. 1999) and HATPase_c domains. These two domains bind ATP (HATPase_c), autophosphorylate a conserved histidine residue, and provide structure for dimerization (HisKA). These two domains are found in all HPK's in two component systems and together form the kinase core.

This relatively simple paradigm of conveying information through phosphoryl transfer also lends itself to more complex configurations including those built on additional phosphoryl transfers. Two-component hybrid systems include an extra transfer within the initial HPK mediated by an extra receiver domain aptly named Respone_reg (Pao and Saier 1995), similar to the receiver domain in the response regulator which catalyze the phosphotransfer from the HPK to the RR. This extra receiver domain then interacts with another phosphorelay domain to transfer the phosphoryl group eventually to the response regulator. One example is the ArcA and ArcB two-component system in *E. coli*. ArcB, the HPK, contains an additional response_reg and HPT domain (Matsushika and Mizuno 1998) that serves as the second site of phosphorylation at a conserved histidine residue (Matsushika and Mizuno 1998).

Further expansion in the form of additional protein phosphorelay intermediates leads to multi-protein systems like those regulating chemotaxis or sporulation. Chemotaxis employs four main proteins required for signal transduction: the chemoreceptor MCP, the histidine kinase CheA, a scaffold protein CheW, and the response regulator CheY(Wadhams and Armitage 2004).

Additional proteins have evolved in different evolutionary branches of this system to regulate the system. CheR and CheB modulate the sensitivity of the sensor through methylation and demethylation of the MCP. CheV contains a CheW domain and a response regulator domain and may be a form of CheW whose function is under regulation(Karatan, Saulmon et al. 2001). CheC and CheD are believed to interact to regulate methylation of MCP's and the adaptation pathway(Rosario and Ordal 1996) and CheC has been shown to aid in the dephosphorylation of CheY-P(Kirby, Kristich et al. 2001). Finally, CheX, and more commonly, CheZ are the phosphatases responsible for dephosphorylating the response regulator CheY(Hess, Oosawa et al. 1988; Motaleb, Miller et al. 2005).

In addition to two component systems, other paradigms of signal transduction have evolved. Adenylate and diguanylate cyclases create cyclic AMP (cAMP) and 3'-5'-cyclic diguanylic acid (c-di-GMP) respectively as messenger molecules as opposed to the direct phosphorylation of a receiver domain on a response regulator protein(Camilli and Bassler 2006). The response regulators of these less common adenylate cyclase systems are identified by the cyclic nucleotide binding domain. The diguanylate cyclase systems also have characteristic protein domains, with the diguanylate cyclases and associated phosphodiesterases containing GGDEF and EAL domains respectively, named for their characteristic polypeptide motif(Jenal and Malone 2006). Finally, even less common are the serine/threonine and tyrosine protein kinases. Proteins containing any variant of the pkinase domain target specific exposed serine or threonine residues which are recognized based on the larger motif in which they reside. Originally thought to be a eukaryotic specific domain, small but significant numbers of proteins containing these domains have been found throughout the bacterial kingdom(Leonard, Aravind et al. 1998).

As knowledge of the number of ST systems and their inclusion in diverse branches of life grew, researchers realized the modularity of signal transduction systems was adaptable to one-component systems. Single proteins that removed the phosphorelay components and instead combined the sensor and output domains together were found(Ulrich, Koonin et al. 2005). In fact, one-component systems were found to be more prevalent and ancient than their two-component relatives, the main difference between the two groups being that one-component systems are cytoplasmic whereas two-component are typically membrane bound.

It has become increasingly apparent that signal transduction systems can be viewed and understood simply from a domain perspective(Galperin and Gomelsky 2005). Protein domains are defined as the smallest independently folding tertiary structures from a single contiguous polypeptide sequence. All ST systems are made up of proteins that contain combinations of a specific subset of domains and different signaling paradigms such as adenylate cyclases and histidine kinases have been shown to utilize the homologous domains for similar functions (e.g. sensory domain CHASE2)(Zhulin, Nikolskaya et al. 2003).

As might be expected input and output domains are highly variable and input domains are especially diverse in particular due to the necessity of adapting to sensing various inputs, e.g. small ligands, redox levels, etc. Since response regulators generally function to regulate gene expression, the output domains function in a DNA-binding capacity, and consequently take the form of the helix-turn-helix (HTH) structure, and are less variable. However, there are examples of output domains which interact with other proteins to convey a signal. The conserved kinase core is much more highly conserved based on its conserved function and is comprised of the transmitter, receiver and Hpt domains.

Recent work has been completed to create a database of domains utilized for signal transduction further aiding in the annotation of newly sequenced genomes and the discovery of novel systems(Ulrich and Zhulin 2007). The Microbial Signal Transduction Database (MiST) contains annotations for Pfam and Smart domain models for every protein in every fully sequenced and published microbial genome. Further, it highlights domains shown to be utilized in signal transduction systems and greatly enhances the ability to recognize novel ST proteins and systems in newly sequenced organisms.

ST protein abundance has also been used to profile the abilities of different bacteria. Cataloging of two-component ST systems in bacteria allowed investigators to use the census information to compute an “IQ” for the various organisms(Galperin 2005). The IQ value represents the ST protein complement normalized for genome size. Not surprisingly, highly motile gram-negative bacteria that had the ability to use a wide variety of electron donors and acceptors scored the best based on the large complement of two-component and one-component systems. In contrast, other signal transduction systems such as adenylate and diguanylate cyclases have not been shown to have a correlation between abundance and genome size.

The overall number of and ratio between one and two-component systems and the overall size of the organisms genome can provide some interesting statistics related to that organisms survival strategies. Previous studies have shown that there is a positive correlation between genome size and the number of regulatory proteins (van Nimwegen 2003; Konstantinidis and Tiedje 2004), while the ratio of transmembrane receptors to intracellular sensors is indicative of an organism's sensitivity to its external environment versus its internal homeostasis. Galperin termed these classes 'extroverts' for organisms more attentive to external factors and 'introverts' for those more concerned with homeostasis(Galperin 2005).

Shewanella

The genus *Shewanella* comprises a group of Gram-negative, aquatic, α-Proteobacteria. Members are motile through the use of a single polar flagellum. As more *Shewanella* have been isolated and studied, their diverse metabolic requirements and abilities have come to light. Most *Shewanella* prefer lactate and other products of fermentations as initial carbon sources and not surprisingly, most *Shewanella* are syntrophic partners of fermentative microbes (Nealson and Scott, 2006). However, some species, most notably *S.*

frigidimarina NCIMB 400, have shown the ability to utilize glucose and other sugars and actually ferment them without aid(Bowman et al., 1997)(Venkateswaran et al., 1999)(Reid and Gordon, 1999). This diverse set of abilities makes it difficult to phenotypically identify different species of *Shewanella*, consequently they are grouped solely based on 16s rRNA sequence.

More than 20 members of the genus *Shewanella* have had their genomes completely sequenced so far, owing to the desire to understand more about organisms with *Shewanella's* exceptional respiration flexibility. *Shewanella* have demonstrated the ability to utilize most electron acceptors more electronegative than sulfate in addition to oxygen. The combination of *Shewanella's* close evolutionary distance to the well-studied E. coli and its extraordinary respiration abilities makes the group extremely well suited for bioremediation tasks. The most important characteristic of *Shewanella* is the ability to easily manipulate the genus under aerobic conditions and utilize them in anaerobic conditions aided by knowledge of closely related systems in *E. coli*. Furthermore, species have been found in habitats ranging from deep ocean sediments to freshwater lakes to food spoilage and include both psychro and piezotolerant members(Kato and Nogi 2001) thereby providing a wide-ranging set of host-adapted environments**.**

Interest in *Shewanella oneidensis* MR-1 was initially driven by the discovery that it was capable of dissimilatory metabolism of manganese and iron oxides(Myers and Nealson 1988). Owing to these initial discoveries and the ease of genetic manipulation, this species quickly became a model organism for metal reduction and has been the main recipient of research attention thus far. With respect to ST, previous work has shown that MR-1 has more than 5 times as many chemoreceptors as *E. coli* indicating a greatly enhanced ability identify and gravitate toward various substances, and a greater number of overall ST proteins and systems, leading to a higher bacterial ‘IQ’(Galperin 2005).

Investigations into ST systems overlap nicely with work being done to understand transcription regulatory networks (TRN) and respiration. Work has been done to develop a genome-wide TRN for *S. oneidensis* MR-1 by applying the mutual information algorithms to a transcriptional profiles(Fredrickson, Romine et al. 2008). Research has also elucidated the highly diverse electron-transport chain that includes as many as 42 c-type cytochromes in *S. oneidensis* MR-1 and the link to the metal reduction process mediated by proteins CymA, MtrB, and MtrC(Myers and Myers 2000; Myers and Myers 2001). This work led to possible applications in biological fuel cells(Fredrickson, Romine et al. 2008) and provides a glimpse of the potential of *Shewanella*. If ST systems are viewed as an overall control structure for other large scale processes like respiration, then greater knowledge of ST systems in Shewanella will only enhance and expedite efforts in other areas.

Chapter II: Materials and Methods

Materials

Pfam Database

Proteins can typically be broken down into one or more regions which fold independently. When these regions are found in multiple proteins and share sequence similarity, they are considered domains. Domains perform consistent functions and can be used to identify and predict aspects of protein function. The Pfam database is a collection of protein domain predictions. These predictions are based on annotations from hidden Markov Models (HMM)(Krogh, Brown et al. 1994; Eddy 1996) created from curated multiple sequence alignments. Version 22.0 was released in July 2007 and contains 9318 families(Finn, Mistry et al. 2006).

MiST Database

The Microbial Signal Transduction (MiST) database(Ulrich and Zhulin 2007) is built from the complete, published genomes of Reference Sequence (RefSeq) database(Pruitt, Tatusova et al. 2007). MiST specializes in the annotation of signal transduction proteins and domains. Signal transduction proteins are identified and classified based on protein domain profiles, i.e. proteins that contain one or more protein domains shown to be utilized in signal transduction processes. It contains the latest annotations of both the Pfam and SMART protein domain databases for all proteins in the published genomes.

MiST also contains both nucleotide and protein sequences and provides predictions for low complexity, transmembrane, coiled coil, and signal peptide regions. Graphical representations of the protein domain structure of each protein and the gene neighborhood for the associated DNA locus are presented through a web interface.

COGS Database

The Cluster of Orthogonal Groups (COGs) database is an effort to create an evolutionary classification of groups of proteins based on orthologous relationships(Tatusov, Fedorova et al. 2003).   These groups are based on sequence and structural similarity and provide implied functional annotations.

Gene Ontology Database

The Gene Ontology (GO) database is a collection of annotations based on a predefined, structured dictionary (Ashburner, Ball et al. 2000). Annotations can be made in one of three areas: Cellular Compartment, Molecular Function, and Biological Process. The dictionary consists of a hierarchical set of terms (GO terms) that become more specific at deeper levels. The dictionary forces consistent descriptions which lead to enhanced comparative power.

DAVID

The (DAVID) database is designed as tool for the interconversion of biological information available in various databases and repositories (Sherman, Huang da et al. 2007). DAVID provides a universal unique ID that can be used to translate or compare in one biological database to any annotations in any other. DAVID maintainers provide a web interface through which a small list of starting ID's (several hundred) can be translated at a time. The information sources available range from structural (PDB) to sequence (Refseq) to functional (COGS) in nature. Annotation information relating to *Shewanella oneidensis* MR-1 from the DAVID 2007 version was downloaded and searched.

Shewanella *species*

Table 1 lists the 10 species chosen for this study. These species were the first ten *Shewanella* species or strains to be sequenced completely.

Methods

BLAST

The Basic Local Alignment Search Tool (BLAST) compares an input sequence against a specified database of sequences and returns a list of statistically significant and locally similar sequences based on the search parameters(Altschul, Gish et al. 1990). Scoring of similarity is based on a user-configurable matrix, and the BLOSUM62 was used in this study. Sequences can be either nucleotides or proteins, and any available sequence database can be searched. BLAST is very flexible in that it can also perform pre-search translations from nucleotides to proteins and vice versa BLAST is maintained by the National Center for Biotechnology Information (NCBI) and source is freely

Table 1. Shewanella species and strains used in this study.

Shewanella Strain	Location	Isolation Environment	Reference
Shewanella sp. ANA-3	Woods Hole, Massachusetts, United States	Brackish water; arsenic-treated wooden pier	(Saltikov, Cifuentes et al. 2003)
Shewanella sp. MR-4	Black Sea	Sea-water; oxic zone; 16oC; 5 m	(Nealson, Myers et al. 1991)
Shewanella sp. MR-7	Black Sea	Sea-water; anoxic zone; high NO3; 60 m	(Nealson, Myers et al. 1991)
Shewanella sp. W3-18-1	Washington coast, Pacific Ocean	Marine sediment; under 997 m of oxic water	(Murray, Lies et al. 2001)
Shewanella amazonensis SB2B	Amapa River, Brazil	Sediment; suboxic redox conditions; 1 m	(Venkateswaran, Dollhopf et al. 1998)
Shewanella denitrificans OS217	Baltic Sea	Sea-water; oxic–anoxic interface; 120 m	(Brettar, Christen et al. 2002)
Shewanella frigidimarina NCIMB 400	Coast of Aberdeen, United Kingdom	Sea-water; North Sea	(Bowman, McCammon et al. 1997)
Shewanella loihica PV-4	Hawaiian Sea mount, United States	Iron-rich mat; hydrothermal vent; 1,325 m	(Gao, Obraztova et al. 2006)
Shewanella oneidensis MR-1	Lake Oneida, New York, United States	Sediment; anaerobic; Mn(IV) reduction	(Myers and Nealson 1988)
Shewanella putrefaciens CN-32	Albuquerque, New Mexico, United States	Subsurface; shale sandstone; 250 m	(Fredrickson, Zachara et al. 1998)

downloadable. In addition to aiding in the identification of members of gene families, BLAST is a valuable tool in the process of elucidating functional and evolutionary relationships at the sequence level.

PSI-BLAST

Position Specific Iterative BLAST (PSI-BLAST) is another tool for finding related sequences. PSI-BLAST takes a single sequence, either nucleotide or protein, and returns a list of statistically significant sequences similar to the input sequence(Altschul, Madden et al. 1997). PSI-BLAST differs from BLAST in the mechanism by which it determines similarity. After an initial BLAST of the input sequence, PSI-BLAST uses the resulting list to building a position-specific scoring matrix (PSSM) that is unique to the input sequence. PSI-BLAST then uses this PSSM to search the appropriate sequence database for further matches and after each search iteratively revises the PSSM for the next search.

As a process, PSI-BLAST lends itself to parallelization very easily. Using the Tiger supercomputer facilities at the Oak Ridge National Lab, Dr. Bhanu Rekapali has developed a tool to automate the parallelization of PSI-BLAST. This tool will take a list of input sequences and search each sequence through 4 iterations and return a list of statistically significant hits. An e-value of 0.001 was used with the BLOSUM62 scoring matrix without any other filters. This automation and parallelization of this process saved large amounts of time and effort.

Determination of Homologous Relationships

Based on the annotations available in the MiST database, proteins believed to play a role in signal transduction were selected from ten strains of the genus *Shewanella* (table 1). These protein sequences were broken down into domain sequences, again obtained from the MiST database, based on annotations from PFAM database version 22(Finn, Tate et al. 2008). In cases where portions of a signal transduction protein were not annotated and there was an open stretch, the sequences were broken into sequences roughly 80-100 amino acids long.

The process for determining homologous relationships is similar to that employed by Tatusov et. al(Tatusov, Koonin et al. 1997), with the exception that reciprocity of best BLAST hits is mandated. In summary, each domain sized sequence was searched using BLAST against each of the other ten species, one species at a time. The best hit from each species was then compared back against the original species through a BLAST search. If that second, reciprocal BLAST search returned the original sequence as the best hit, the two are deemed reciprocal best hits and homologous. Three best hit pairs for a given sequence are required to be considered a homologous group (i.e. the original sequence and sequences from two other organisms as reciprocal best hits to the original). Groups that share common pairs are joined to form larger groups. Homologous groups are then assigned unique ID’s and stored in the database (see figure 1). This initial step was designed and carried out by Luke Ulrich.

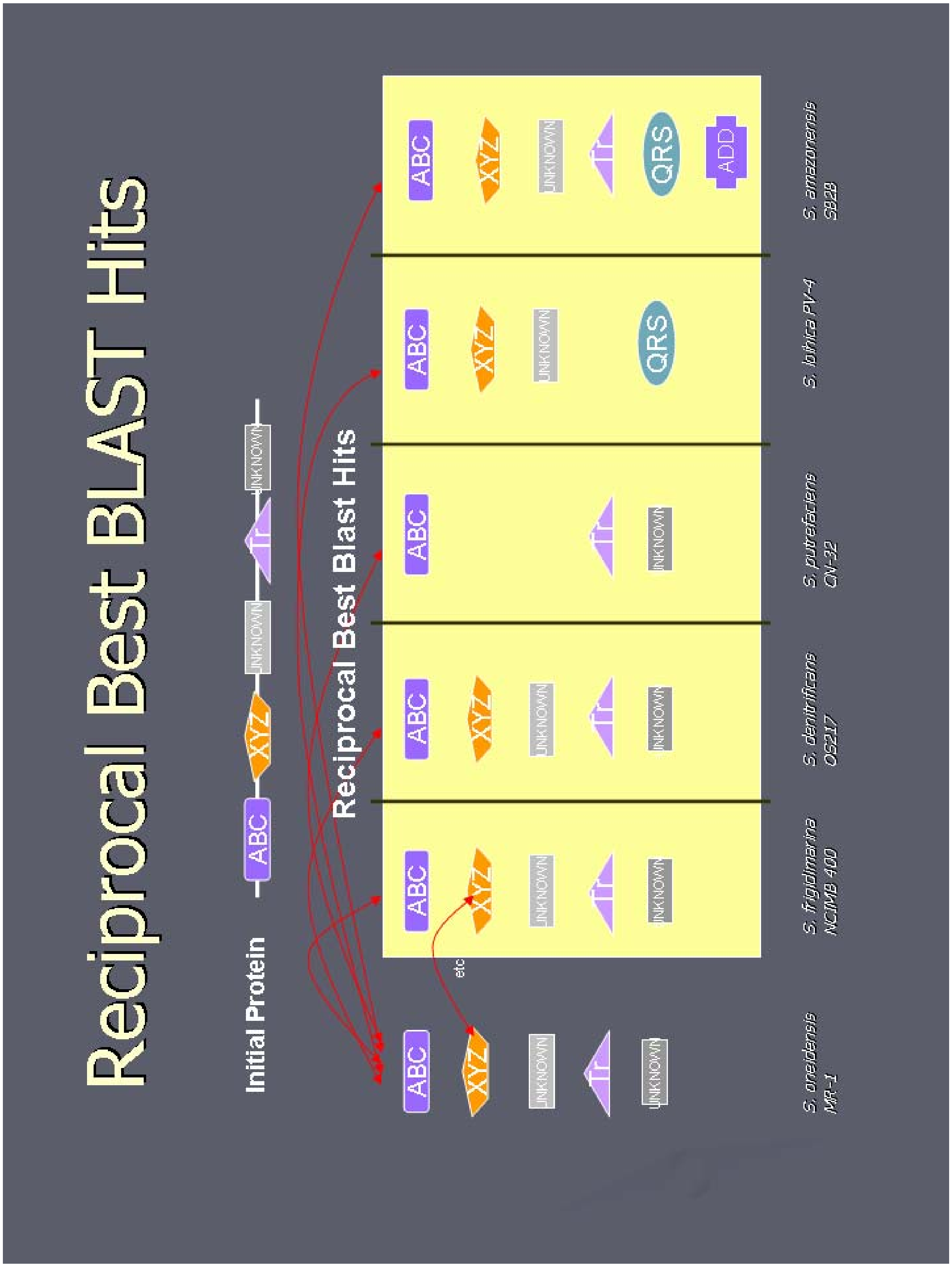

Figure 1. Diagram of Methodology. This diagram represents the process by which homology is determined.

The domain and domain size sequences are then recombined into whole protein sequences and the reassembled proteins were then assessed for the overall patterns of conservation and homology at the domain levels. Proteins classified on the percentage of domain similarity/orthology they shared with other proteins and grouped. Protein groups that shared similarity at each and every domain were considered to be orthologous or paralogous while proteins that shared similarity at the majority of domains were considered to show “significant similarity”. Proteins that only shared similarity at one or fewer than half of their domains were considered to show “limited similarity”. Orthologous protein groups that had representatives in each *Shewanella* species were deemed to be members of the “core” signal transduction apparatus of the genus.

Core Annotation

Those groups with representatives in each of the 10 species constitute the core signal transduction apparatus of *Shewanella*, and as such determine the basic functionality of any member of the *Shewanella* genus. Consequently, understanding the makeup and abilities of this group is of paramount importance. To that end several different sources of information have been searched. First, COG annotations for the core proteins in *Shewanella oneidensis* MR-1 were determined using Reverse Position Specific (RPS) BLAST against predefined COG PSSM’s. In RPS-BLAST search sequences are queried against the COGs models. Next, searches for GO annotations were conducted through DAVID. These annotations were combined to determine the best and most thorough descriptions for the proteins involved, and were especially necessary in cases where the protein was annotated as a conserved hypothetical protein.

Identification of Paralogs

In the process of determining reciprocal best hits, only three pairs of best hits are required to create a homologous group. Furthermore, these original three can have independent reciprocal best hits in other organisms that are not necessarily best hits to the other two original members. These new reciprocal best hits can then have reciprocal best hits in one or both of the original organisms that are different from the original sequences. In this way a given organism can have multiple sequences in a homologous group, and these duplicate sequences are considered paralogous. However, a minimum of five organisms and six sequences is required in order to define paralogs by this method.

Figure 2 demonstrates a graphic example. Each colored node represents a protein with a single domain in an organism, and the edges connecting nodes represents reciprocal best BLAST hits between them. Nodes with the same color represent paralogs, like the graph on the left. Proteins with multiple domains require congruent overlapping graphs.

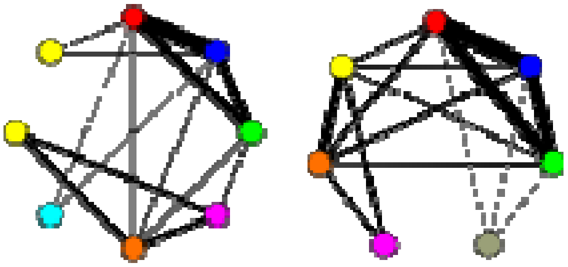

Figure 2. Diagram of paralog identification. Nodes represent domains and edges represent reciprocal best BLAST hits.

Phylogenetic Analysis

A 16s ribosomal RNA (rRNA) tree was constructed based on sequences obtained from the Silva database, a comprehensive online resource of up-to-date, quality controlled rRNA sequence information(Pruesse, Quast et al. 2007). All annotated, full-length 16s rRNA sequences were downloaded and aligned using ClustalW in the Mega package and a tree was created using the neighbor-joining algorithm.

The paralog data was determined based on analysis of the reconstructed protein information gathered from earlier steps. Protein domains in the same organism that were grouped based on reciprocal best BLAST hits were deemed paralogs. There were 56 separate groups of homologous protein groups with paralogs i.e. multiple representatives in a single organism. Five organisms were required to have reciprocal best BLAST hits to discover paralogs.

A matrix of paralog information was created with organism's paralog information as a row and each homologous protein group as a column. The pairwise distance between each organism's row was computed using the pdist function (both Euclidean and cosine distance measures) of Matlab and a tree was built using the both the neighbor-joining function seqneighjoin (using the ‘equivar’ option) and the linkage function (using the ‘ward’ method).

Chapter III: Results

Signal Transduction Conservation

Figure 3 shows the results from the initial survey of signal transduction proteins in the 10 species of Shewanella as annotated in MiST (see Materials and Methods). The first column in blue shows the number of proteins in the given organism with significant similarity to proteins in at least two other organisms. The second column in orange shows the total number of proteins annotated as ST proteins. The total ST protein counts range from 303 to 417 while the homologous counts range from 256 to 384. The percentage of ST proteins with significant similarity ranges from 85% to 99%. The genome size, shown by the yellow line, varies between roughly 4.5 Mb and 5.5 Mb. It is apparent from Figure 1 that S. denitrificans OS217 has undergone a significant loss of ST proteins without a large net reduction in genome size.

The Core

To be included in the set of core proteins an orthologous group must meet several criteria. First, the group must have an invariant protein domain organization. Second, each domain must be represented in every other protein as the reciprocal best BLAST hit. Finally, the group must a have a representative

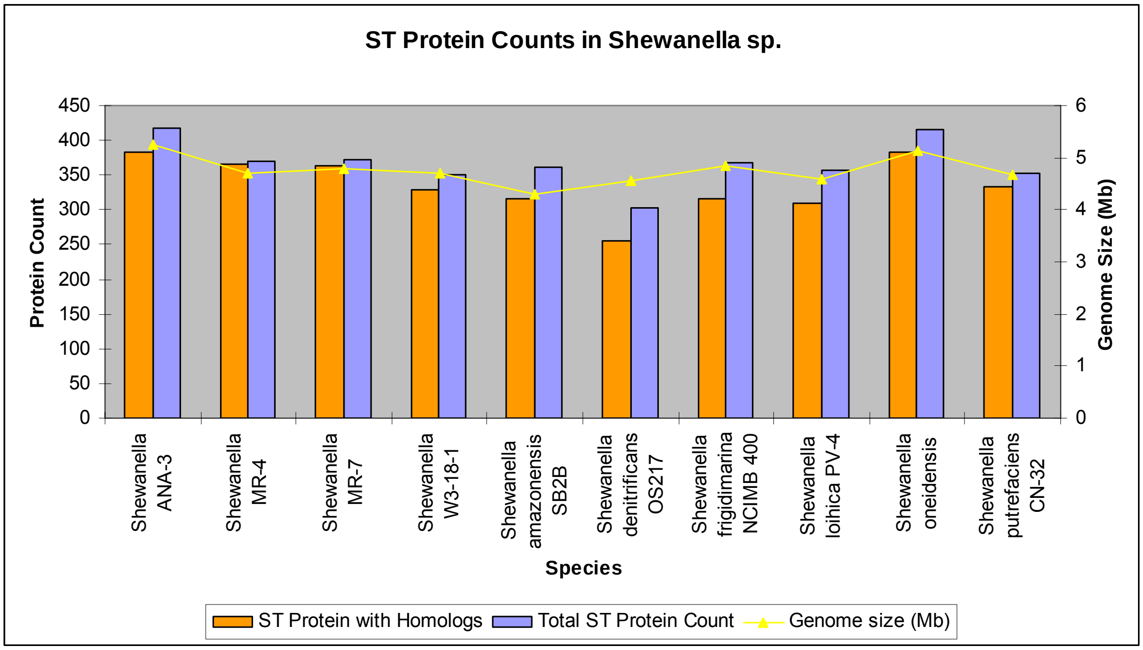

Figure 3. Signal Transduction Protein Counts in Shewanella.

protein in every species. Ninety-nine protein groups met these criteria for the 10 species of *Shewanella* surveyed (see Appendix A).

Of the 99 proteins in the core group in *Shewanella oneidensis* MR-1, 66 are labeled as one-component in the MiST database, and the other 33 are labeled as two-component. Most striking about the list of core proteins is the lack of knowledge from traditional biochemical or genetic techniques, i.e. experimental data. Forty-two of the 66 one-component proteins are generally uncharacterized with only automated annotation such as domain name. Fifteen of the 33 two-component proteins are similarly sparsely annotated. For several proteins “hypothetical conserved” is the extent of the information provided representing putative homology to genes or proteins in other organisms, while others don’t go further than domain annotations. Other sources of information were equally ambiguous. Gene Ontology (GO) term annotation was not much more descriptive than what could be found from glancing at domain information.

There are several familiar groups represented in the core protein group. First is an almost complete chemotaxis system including CheB, CheR, CheW, 3 CheV, 4 MCP's, CheY and CheZ. The multiple CheB, CheR and CheV proteins taken together with the abundant MCP's (more than 20 in most of the species) highlight the diversified chemotaxic abilities of the *Shewanella* and the highly evolved control mechanisms needed to integrate the increased and wide ranging sensitivity.

Also parts of the core ST protein group are several two-component systems. The list includes systems responsible for scavenging for phosphate and nitrogen: phoR and phoB, and ntrB and ntrC. The envelope stress response system is also present in cpxA and cpxR. Finally, ompR and envZ are found in tandem as members of the core.

Significant Similarity

After the core group of ST proteins, the next most conserved groups of proteins were those that showed significant similarity. These protein groups had more than 50% of their domains as reciprocal best hits and in some cases had 100% but were missing a representative in one of the species. There were 132 protein groups in the former and 166 in the latter.

CheA was found in this group. The reason for its exclusion from the core group stems from its sequence variability in the region after the Hpt domain and before the H_kinase_dim domain, roughly amino acids 110 to 315. There are several low-complexity subsequences in this region and their spacing and length is variable across the 10 species. This region is analogous to the P2 region of the *E. coli* CheA and is known to be divergent. This variability lead to mismatches with respect to determining reciprocal best BLAST hits and consequently to an incomplete set of homologous domains.

Again, the list of well-characterized protein representatives is sparse. Of the 298 different homologous protein groups there were 237 proteins in *S. oneidensis* MR-1, only 23 proteins have been annotated beyond automated means.

Limited and No Similarity

A list of the totals for each grouping appears in Table 2. ‘Limited similarity’ proteins have domain homology for fewer than half their constituent domains. ‘No similarity’ proteins have no domains with any similarity to any others in any of the organisms as defined by the reciprocal BLAST best hit methodology. As noted previously S. denitrificans OS217 has a significantly smaller amount of similarity, but interestingly has a relatively high number of unique signal transduction proteins. The smaller number of unique proteins for the MR-4 and MR-7 strains is most likely due to their close evolutionary distance as proteins have not had enough time to diverge significantly.

Table 2. Similarity Totals from *Shewanella*

Species	Limited	Significant	Core	Sum	No Similarity	Total STP
Shewanella ANA-3	24	235	99	358	59	417
Shewanella MR-4	21	221	99	341	29	370
Shewanella MR-7	20	218	99	337	36	373
Shewanella W3-18-1	23	183	99	305	45	350
Shewanella amazonensis SB2B	18	177	99	294	68	362
Shewanella denitrificans OS217	16	130	99	245	58	303
Shewanella frigidimarina NCIMB 400	15	183	99	297	70	367
Shewanella loihica PV-4	18	177	99	294	64	358
Shewanella oneidensis	25	237	99	361	54	415
Shewanella putrefaciens CN-32	22	189	99	310	42	352

Protein Domain Identification

Pfam domain annotations are based on results derived from profile hidden Markov models (profile hMM). Theses profiles are built from multiple sequence alignments and recognize similar domains based on that sequence similarity. Consequently, evolutionarily distant sequences that share little sequence similarity, but still result in the same folding characteristics and functional use may not be recognized by the appropriate HMM. However, other similarity scores can be used in lieu of the hMM to provide evidence for domain homology.

One way to annotate putative protein domain is to compare them to existing annotations of similar regions in homologous proteins. The groups of orthologous proteins provide an excellent framework in which to perform these comparisons. To reiterate, based on the fact that each domain represents the best reciprocal BLAST hit (See Materials and Methods) for every other domain in the group, and therefore a homologous and potentially orthologous relationship, each domain in a homologous group can be interpreted as a homologous fold and function.

There are 10801 domain or domain sized regions (hereafter domains) investigated in this study and those regions were grouped into 1292 homologous groups with 1447 domains not included in any. There are 4216 domains were unrecognizable by Pfam domain models and are annotated as unknown and 893 domains annotated as unknown were in groups that included at least one annotated member. Figure 4 provides totals for the number of unknown domains which are part of an orthologous group in *Shewanella* as defined previously (see Materials and Methods) with at least one annotated member. Not surprisingly, domains with known sequence divergence, such as HAMP and PAS domains, have the highest totals.

In order to provide evidence for the relationship between annotated and possibly related 'unknown' domains, the bit scores of the BLAST hits are displayed in Figure 5. To test the strength of the relationship between the known domains with annotations and the unknowns believed to be homologous, bit scores between known and related unknown domains and perfect score and 50% scores are provided for comparison. As domains increased in length, scores generally decreased.

Figure 6 displays the results from attempts to recognize domains by going outside of the *Shewanella* genus. Using an automated PSI-BLAST approach (see Materials and Methods) unknown domain regions were searched against

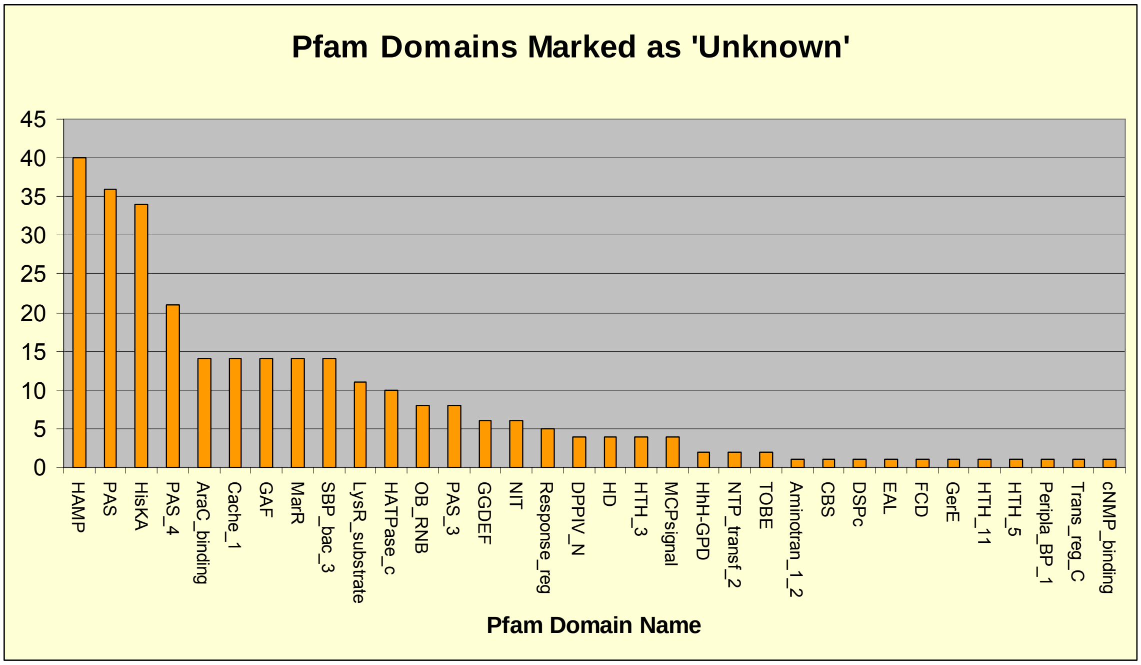

Figure 4. Probable Known 'Unknown' Domains. The domains listed above were found to be homologous to domains marked as ‘unknown’ indicating a high degree of conservation.

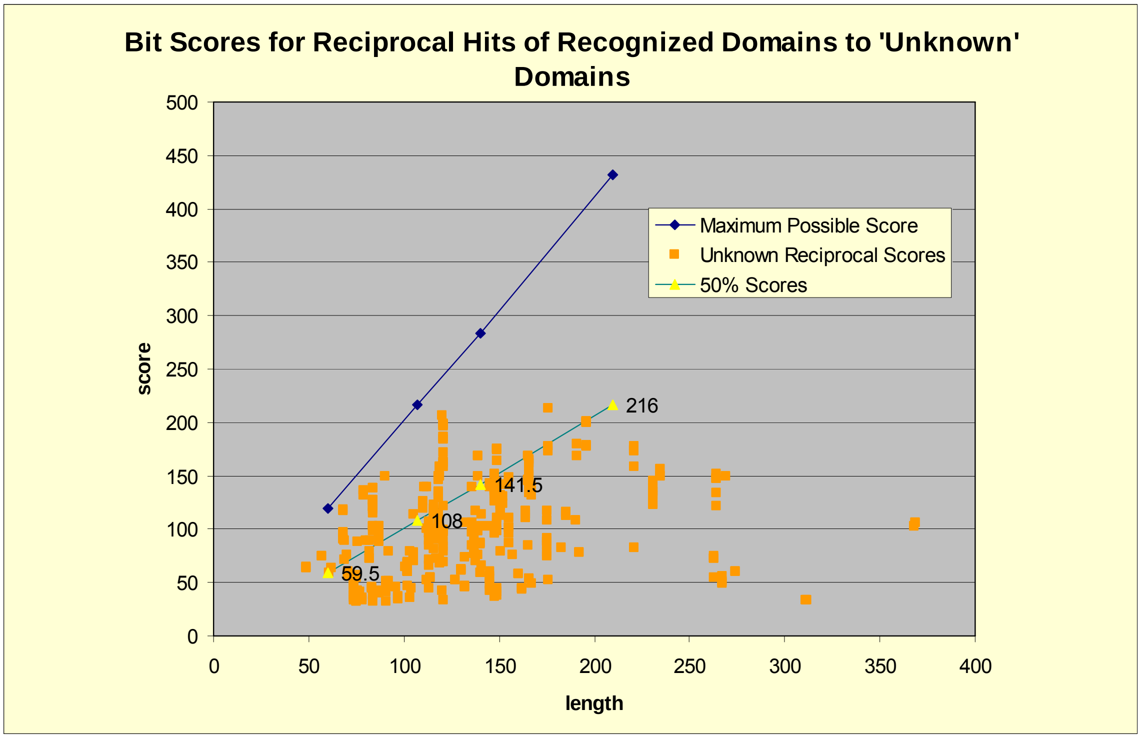

Figure 5. Bit Scores for Reciprocal BLAST hits between Known and Related Unknown Domains. This graph represents the bit scores of annotated domains when compared to unknown domain regions. Each score is a reciprocal best BLAST hit.

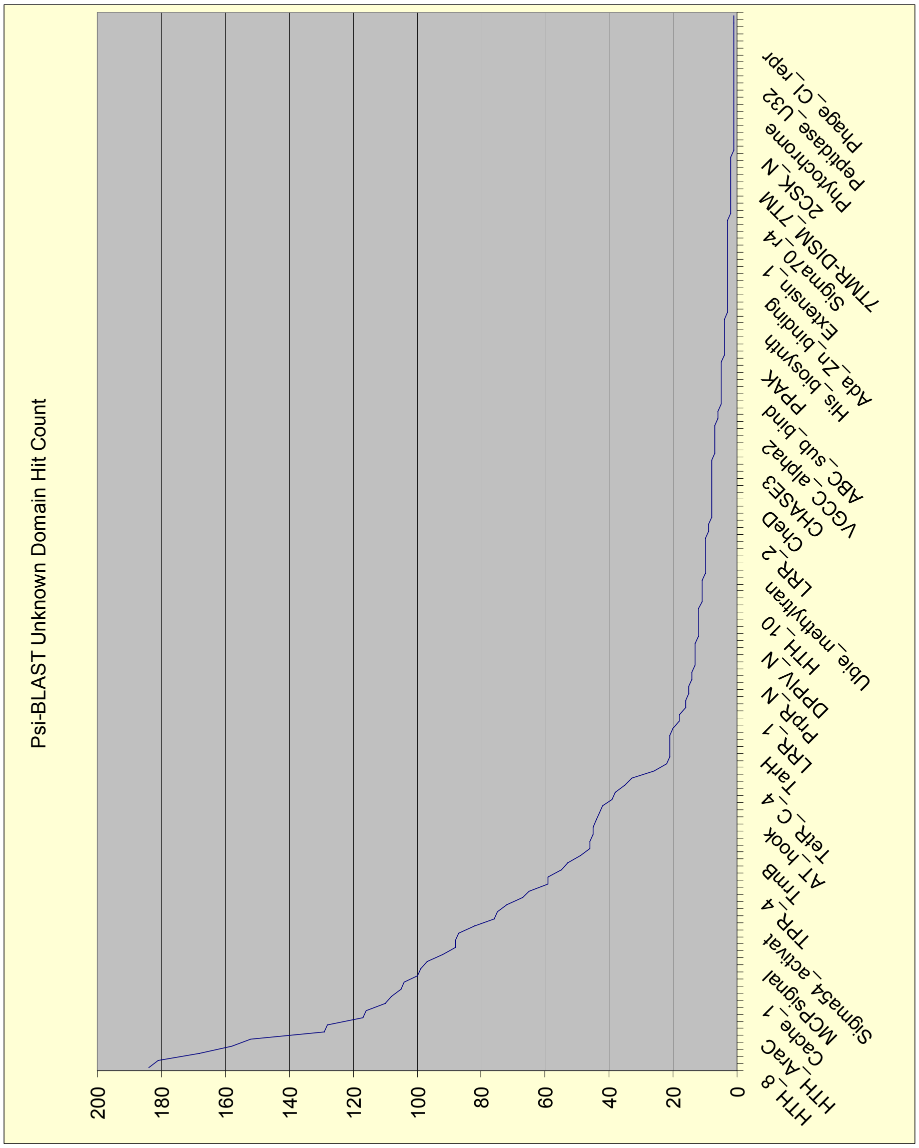

Figure 6. PSI-BLAST Unknown Domain Search

the non-redundant database to determine if they had significant similarity to other regions with existing domain annotations. A total of 3507 regions were searched and 2050 were found to have significant hits to regions previously annotated. Of those sequences, 1457 had no hits to previously annotated regions. The 2050 sequences with hits were found to be similar to 150 different domain models (see Appendix C.2). Again, domains with known variability such as the PAS family predominated.

Phylogenetic Analysis

In conjunction with information about homologous relationships, the reciprocal best hit process provided paralogous information as well. Fifty-six homologous protein domains were found to have paralogs in multiple organisms. This data was clustered and compared to 16s rRNA based phylogenetic data to determine what if any deviance it might show evolutionarily (see Material and Methods).

The relationship between the 10 strains of Shewanella is represented in Figure 7. In general, there are several tight clusters with *S. amazonensis* SB2B and *S. loihica PV-4* being the most distantly related. The individual rRNA gene sequences cluster by species with a few notable exceptions. First, the *S. sp ANA-3, S. sp MR-4, and S. sp MR-7* group primarily in two large clusters indicating their close evolutionary relationship. Second, there is some overlap among the *S. putrefaciens* CN-32 and *S. sp.* W13-18-1. *S. oneidensis* MR-1 is most closely related to the *S. putrefaciens* CN-32 and *S. sp.* W13-18-1 clade.

The tree based on the paralog data (see Appendix F, Materials and Methods) paints a different picture as shown in Figure 8. While *S. sp ANA-3, S. sp MR-4, and S. sp MR-7* cluster together again and the *S. putrefaciens* CN-32 and *S. sp.* W13-18-1 also cluster together, *S. oneidensis* MR-1 has taken a new position relative to the others. It is now most closely paired with *Shewanella frigidimarina* NCIMB 400.

It is interesting to note that *S. oneidensis* MR-1 and *S. frigidimarina* NCIMB 400 share the deepest branch and the most unique paralogous domains. While there are three paralogous domains in common, *S. oneidensis* MR-1 also has three paralogous domains in common with *S. sp.* MR-4 and *S. sp.* MR-7. However, those domains are also shared with several other species in one instance including *S. amazonensis* SB2B and in another instance *S. putrefaciens* CN-32 and *S. sp.* W3-18-1. Visual inspection of the gene neighborhoods of the proteins in *S. oneidensis* MR-1 and *S. frigidimarina* NCIMB 400 that share the paralogous domains shows that whole proteins are intact and flanked by transposable elements. Reconstruction of the paralogous events is also complicated by the fact that *S. oneidensis* MR-1 contains a plasmid a large plasmid that is not shared by *S. frigidimarina* NCIMB 400, and that some of the paralogous sequences are found on this plasmid.

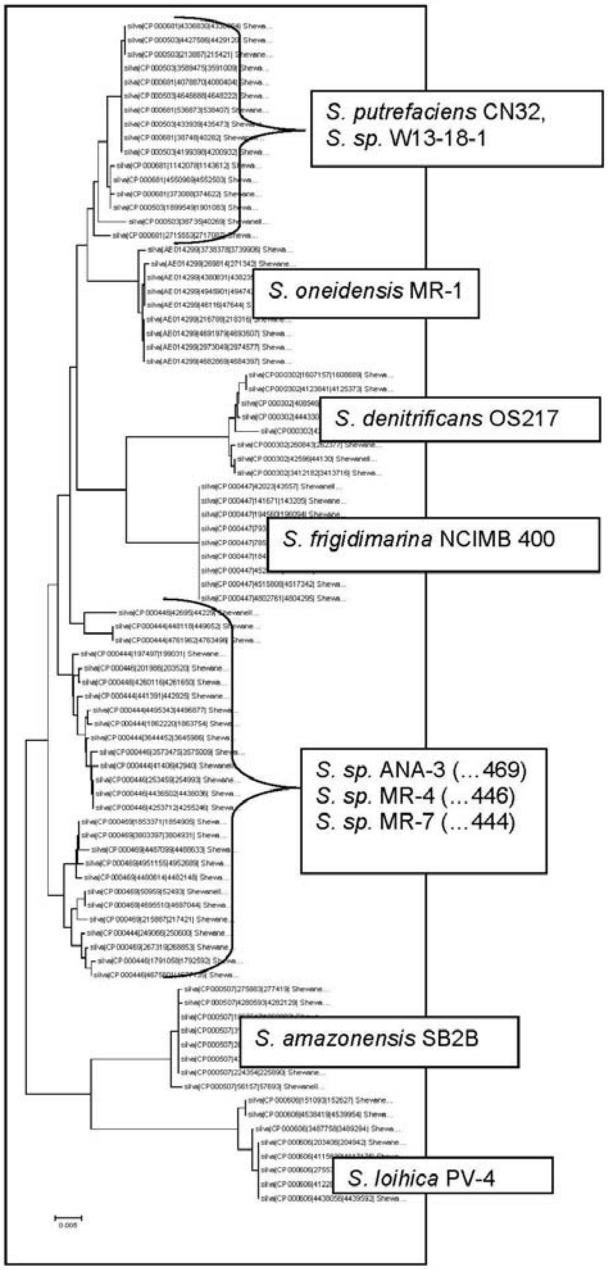

Figure 7. 16s tree of 10 *Shewanella* species. The tree was built with ClustalW in the Mega package using the neighbor-joining algorithm.

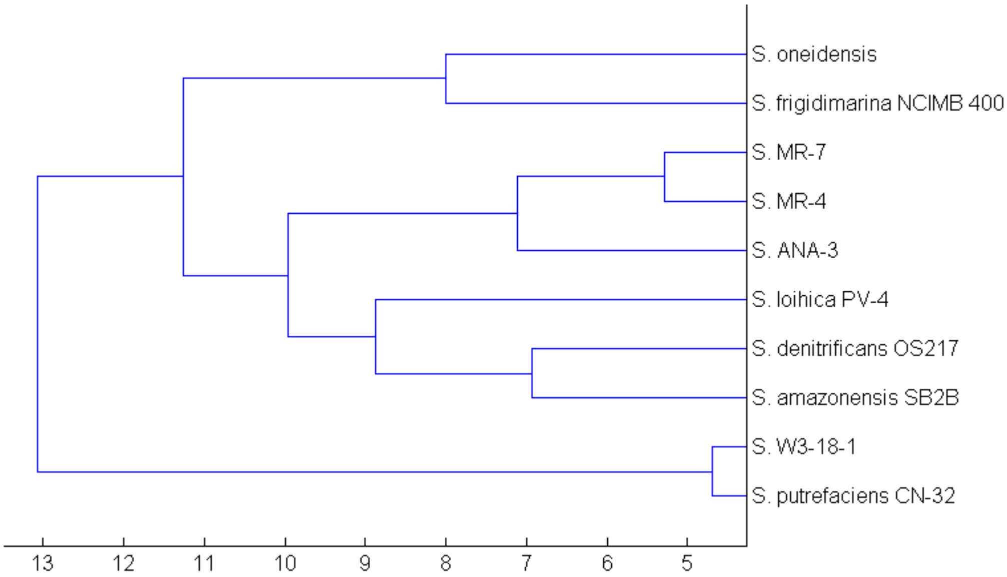

Figure 8. Tree based on paralogous domains data.

Chapter IV: Discussion

This study demonstrates the power of comparative genomics and more specifically, the resolution that can be obtained with access to the genome sequences of a large set of organisms related at the species level. Whereas in previous studies comparisons could only be made at a systems level, having complete genome sequence information from multiple species of the same genus we can shed light on how systems evolve and even how individual proteins evolve in those systems. With the enhanced ability to see finer details we can determine the elements that define groups of organisms and the features that are specific to only some or one. This method for exploiting homology will be increasingly available as more and more gaps are filled in on the evolutionary tree.

The first goal of this research was to define the core set of signal transduction proteins from *Shewanella spp*. and thereby define the innate abilities common to all of the members of this study. The invariant members of this core group represent the mechanisms and processes most tightly controlled through evolution. Specifically, this conserved group demonstrates the importance of chemotaxis to every species in the study. Furthermore, it highlights the basic conserved functionality of osmolarity sensing, nitrogen and phosphate regulation, and the envelope stress response system. All of these are basic system crucial to the survival of any organism and so it's not surprising that they would be members of core set of conserved proteins. Finally the large numbers of putative transcription factors implies a large number of conserved pathways and other conserved processes outside the scope of this study.

The core set of conserved proteins was also notable for the relatively sparse coverage of annotations and information. Two thirds of the core set was only annotated with the most basic information. This would seem to imply that one the greatest utility for to come from this study would be as a starting point for further experimental characterization.

Much like the core set, the ‘significant similarity’ group also highlights interesting features of the evolution of signal transduction in Shewanella. The two categories which comprise this group of proteins each provide insight into how the individual species are evolving. The first group is comprised of proteins that are completely conserved, but are absent from one or more species and this group shows the impact of the large gene loss in *S. denitrificans* OS217. If we exclude *S. denitrificans* OS217 and group only on the remaining 9 species 30 additional protein groups are added to the core group. In contrast, if we exclude *S. loihica* PV-4, the most distantly related species based on 16s phylogeny and regroup, only 3 additional protein groups are added to the core group.

The second category of significant similarity demonstrate some the strengths and weakness of this particular approach. The protein groups have representative proteins with changes in domain architecture, for example additions, deletions, or domains which are no longer reciprocal best BLAST hits. As an example, CheA is obviously integral to chemotaxis, a system whose proteins have already been shown to be members of the core conserved group.

However, CheA is variable enough in the P2 region that it no longer propagates reciprocal best BLAST hits across even the closely related members of this study. Consequently, like many other powerful bioinformatics based approaches, the results are not always straightforward and clear in their interpretation.

While not always clear, this approach of using reciprocal best BLAST hits to demonstrate homology does have the power to shed light on other areas where other tools are lacking. Determining protein domain identification only through profile hidden Markov model (HMM) is dependent upon the initial sequences used to create the alignment upon which the HMM is built. In many cases these sequences are from closely related organisms and the sequences used do not possess a great deal of diversity, especially in regions less critical to function and more critical to structure. However, very similar domain structures can be created by divergent sequences so structures that have maintained their overall structure and possibly function will not be recognized by HMM's built from these initial biased samples.

The analysis of protein domains demonstrates the fallibility of HMM based domain recognition. Not unsurprisingly, domains known for their sequence variability were missed. The PAS domain is a ubiquitous sensor domain capable of binding small ligands or employing a cofactor to sense changes in local characteristics and is known to have a highly divergent sequence(Zhulin, Taylor et al. 1997). There are currently seven different Pfam HMM's based on thousands of sequences employed to recognize this fold and yet there are still a small but significant number of cases where the HMM's fail as the results from this study show. Of the roughly 4000 sequences not recognized by HMM's (roughly 40% of the total sequences), more than half were recognized either by BLAST-based sequence similarity or automated PSI-BLAST. Clearly, by combining the two approaches and using other approaches a higher fraction of coverage can be attained.

The enhanced recognition ability provided by combining profile HMM's and homology study is a great benefit of this method. It becomes increasingly important when our ability to sequence new organisms greatly outstrips our ability to experimentally characterize the resulting data. For signal transduction systems, the problem of missed annotations is compounded by the fact that automated ST protein characterization is highly dependent on the constituent domains. The current situation bears out the need for increased ability to make accurate predictions as 80% the proteins in the 'significant similarity' set only had basic automated annotations.   Orthologous proteins have names that range in descriptive ability from "sensory box protein" to "diguanylate cyclase/phosphodiesterase with PAS/PAC sensor(s)" (GI: 24374900, 114562745). The ability to make better predictions will naturally enhance our ability prioritize our investigations of systems and to characterize organisms.

The diverse respiratory talents of Shewanella make any characterization of their relationships difficult due to the fact that the different methods seem to provide different answers, specifically with respect to *S. oneidensis* MR-1. The traditional method of ribosomal RNA based phylogeny places MR-1 nearest to S. putrefaciens CN-32 and S. sp. W3-18-1 among the 10 members of this study.

However, in a study done by Wang et al. that included the 10 species in this study and using a whole proteome sequence based phylogeny method, MR-1 was found to be closest to *S. sp.* ANA-3, *S. sp.* MR-4 and *S. sp.* MR-7(Wang, Wang et al. 2008).

This position for MR-1 is contradicted by clustering of the paralog data generated from this study where MR-1 is found to be closest to *S. frigidimarina* NCIMB 400. This latest finding may lend some credence to the theory that MR-1 is a recent contaminant of Lake Oneida(Hau and Gralnick 2007). The theory holds that canals built in the 19th century that connect the lake to the Hudson River and Lake Ontario created the possibility of contamination by ocean going ships. Combined with the fact that *S. frigidimarina* NCIMB 400 has the highest number of unique signal transduction proteins suggests that

This novel relationship between these species highlights the power of this comparative genomics study. These findings were made possible by the ability to compare many closely related species. In addition, by defining a core group of conserved signal transduction proteins we have identified processes critical to the function of all Shewanella species and provided a prioritized list for future investigation. This knowledge will aid in the further exploitation of Shewanella by providing insight into the critical processes of signal transduction.

Chapter V: Future Work

The definitions used to determine the core set of conserved signal transduction proteins and the significant similarity group represent conservative estimates. Groups were assigned to provide stringent criteria with respect to conservation and may have erred on the side of caution. The case of CheA is one obvious example where these criteria may have proven too strict. CheA is an integral chemotaxis protein with a conserved function. Because of a region of sequence variability, CheA did not meet the requirements to be included in the core set of conserved proteins.

A review of the method used to generate the data would seem to be a logical place to determine if situations like this could be remedied. The CheA situation was due in large part to the method used to generate the underlying data. Proteins were broken up into smaller sequences based on domain annotations. Regions without annotations were broken up into domain sized sequences of around 100 amino acids long. At this point all of the sequences were treated the same even though domain annotations clearly imply a higher probability of conservation.

Future versions of this method should make a distinction between sequences with and without domain annotations. Perhaps the easiest way would be to investigate first the relationships between sequences with annotations and their reciprocal best BLAST hits in related organisms. A first pass with these annotated sequences would highlight conservation and identifying putative domains in other organisms that are missed by current methods. Due to the current coverage of domain models, it would be reasonable to expect that more than half of the sequences would be recognized. The next step would be to investigate sequence regions that gave no indication of protein domains, either by domain model recognition or similarity to annotated regions. High levels of sequence similarity would indicate possible novel domains while low levels of similarity would indicate areas not being conserved and possibly less important to the overall function of the protein. Regions with low levels of similarity could be searched with more general approaches like PSI-BLAST. And proteins with these low similarity regions would not necessarily have to be excluded from orthologous groups if these regions were recognized and interpreted as highly variable. In this way a multistep approach would reveal as much, if not more information while avoiding some of the shortcomings of the previous approach.

LIST OF REFERENCES

LIST OF REFERENCES

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

Bowman, J. P., S. A. McCammon, et al. (1997). "Shewanella gelidimarina sp. nov. and Shewanella frigidimarina sp. nov., novel Antarctic species with the ability to produce eicosapentaenoic acid (20:5 omega 3) and grow anaerobically by dissimilatory Fe(III) reduction." Int J Syst Bacteriol **47**(4): 1040-7.

Brettar, I., R. Christen, et al. (2002). "Shewanella denitrificans sp. nov., a vigorously denitrifying bacterium isolated from the oxic-anoxic interface of the Gotland Deep in the central Baltic Sea." Int J Syst Evol Microbiol **52**(Pt 6): 2211-7.

Camilli, A. and B. L. Bassler (2006). "Bacterial Small-Molecule Signaling Pathways." Science **311**(5764): 1113-1116.

Eddy, S. R. (1996). "Hidden Markov models." Curr Opin Struct Biol **6**(3): 361-5.

Finn, R. D., J. Mistry, et al. (2006). "Pfam: clans, web tools and services." Nucleic Acids Res **34**(Database issue): D247-51.

Finn, R. D., J. Tate, et al. (2008). "The Pfam protein families database." Nucl. Acids Res. **36**(suppl_1): D281-288.

Fredrickson, J. K., M. F. Romine, et al. (2008). "Towards environmental systems biology of Shewanella." Nat Rev Microbiol **6**(8): 592-603.

Fredrickson, J. K., J. M. Zachara, et al. (1998). "Biogenic iron mineralization accompanying the dissimilatory reduction of hydrous ferric oxide by a groundwater bacterium." Geochimica et Cosmochimica Acta **62**(19-20): 3239-3257.

Galperin, M. Y. (2005). "A census of membrane-bound and intracellular signal transduction proteins in bacteria: bacterial IQ, extroverts and introverts." BMC Microbiol **5**: 35.

Galperin, M. Y. and M. Gomelsky (2005). "Bacterial signal transduction modules: from genomics to biology." ASM News **71**(7): 8.

Gao, H., A. Obraztova, et al. (2006). "Shewanella loihica sp. nov., isolated from iron-rich microbial mats in the Pacific Ocean." Int J Syst Evol Microbiol **56**(Pt 8): 1911-6.

Hau, H. H. and J. A. Gralnick (2007). "Ecology and biotechnology of the genus Shewanella." Annu Rev Microbiol **61**: 237-58.

Hess, J. F., K. Oosawa, et al. (1988). "Phosphorylation of three proteins in the signaling pathway of bacterial chemotaxis." Cell **53**(1): 79-87.

Hoch, J. and T. Silhavy (1995). Two-Component Siganl Transduction. Washington, D.C., ASM Press.

Jenal, U. and J. Malone (2006). "Mechanisms of Cyclic-di-GMP Signaling in Bacteria." Annual Review of Genetics **40**(1): 385-407.

Karatan, E., M. M. Saulmon, et al. (2001). "Phosphorylation of the response regulator CheV is required for adaptation to attractants during Bacillus subtilis chemotaxis." J Biol Chem **276**(47): 43618-26.

Kato, C. and Y. Nogi (2001). "Correlation between phylogenetic structure and function: examples from deep-sea Shewanella." FEMS Microbiol Ecol **35**(3): 223-230.

Kirby, J. R., C. J. Kristich, et al. (2001). "CheC is related to the family of flagellar switch proteins and acts independently from CheD to control chemotaxis in Bacillus subtilis." Mol Microbiol **42**(3): 573-85.

Konstantinidis, K. T. and J. M. Tiedje (2004). "Trends between gene content and genome size in prokaryotic species with larger genomes." Proc Natl Acad Sci U S A **101**(9): 3160-5.

Krogh, A., M. Brown, et al. (1994). "Hidden Markov models in computational biology. Applications to protein modeling." J Mol Biol **235**(5): 1501-31.

Leonard, C. J., L. Aravind, et al. (1998). "Novel Families of Putative Protein Kinases in Bacteria and Archaea: Evolution of the "Eukaryotic" Protein Kinase Superfamily." Genome Res. **8**(10): 1038-1047.

Matsushika, A. and T. Mizuno (1998). "A dual-signaling mechanism mediated by the ArcB hybrid sensor kinase containing the histidine-containing phosphotransfer domain in Escherichia coli." J Bacteriol **180**(15): 3973-7.

Matsushika, A. and T. Mizuno (1998). "The structure and function of the histidine-containing phosphotransfer (HPt) signaling domain of the Escherichia coli ArcB sensor." J Biochem **124**(2): 440-5.

Motaleb, M. A., M. R. Miller, et al. (2005). "CheX Is a Phosphorylated CheY Phosphatase Essential for Borrelia burgdorferi Chemotaxis." J. Bacteriol. **187**(23): 7963-7969.

Murray, A. E., D. Lies, et al. (2001). "DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes." Proc Natl Acad Sci U S A **98**(17): 9853-8.

Myers, C. R. and K. H. Nealson (1988). "Bacterial Manganese Reduction and Growth with Manganese Oxide as the Sole Electron Acceptor." Science **240**(4857): 1319-1321.

Myers, J. M. and C. R. Myers (2000). "Role of the tetraheme cytochrome CymA in anaerobic electron transport in cells of Shewanella putrefaciens MR-1 with normal levels of menaquinone." J Bacteriol **182**(1): 67-75.

Myers, J. M. and C. R. Myers (2001). "Role for outer membrane cytochromes OmcA and OmcB of Shewanella putrefaciens MR-1 in reduction of manganese dioxide." Appl Environ Microbiol **67**(1): 260-9.

Nealson, K. H., C. R. Myers, et al. (1991). "Isolation and identification of manganese-reducing bacteria and estimates of microbial Mn(IV)-reducing potential in the Black Sea." Deep Sea Research **38**: S907-S920.

Ninfa, A. J. and B. Magasanik (1986). "Covalent Modification of the glnG Product, NRI, by the glnL Product, NRII, Regulates the Transcription of the glnALG Operon in Escherichia coli." Proceedings of the National Academy of Sciences **83**(16): 5909-5913.

Nixon, B. T., C. W. Ronson, et al. (1986). "Two-Component Regulatory Systems Responsive to Environmental Stimuli Share Strongly Conserved Domains with the Nitrogen Assimilation Regulatory Genes ntrB and ntrC." Proceedings of the National Academy of Sciences **83**(20): 7850-7854.

Pao, G. M. and M. H. Saier, Jr. (1995). "Response regulators of bacterial signal transduction systems: selective domain shuffling during evolution." J Mol Evol **40**(2): 136-54.

Pruesse, E., C. Quast, et al. (2007). "SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB." Nucl. Acids Res. **35**(21): 7188-7196.

Pruitt, K. D., T. Tatusova, et al. (2007). "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." Nucleic Acids Res **35**(Database issue): D61-5.

Rosario, M. M. and G. W. Ordal (1996). "CheC and CheD interact to regulate methylation of Bacillus subtilis methyl-accepting chemotaxis proteins." Mol Microbiol **21**(3): 511-8.

Saltikov, C. W., A. Cifuentes, et al. (2003). "The ars detoxification system is advantageous but not required for As(V) respiration by the genetically tractable Shewanella species strain ANA-3." Appl Environ Microbiol **69**(5): 2800-9.

Sherman, B. T., W. Huang da, et al. (2007). "DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis." BMC Bioinformatics **8**: 426.

Stock, J., A. Stock, et al. (1990). "Signal Transduction in Bacteria." Nature **344**: 6.

Tatusov, R. L., N. D. Fedorova, et al. (2003). "The COG database: an updated version includes eukaryotes." BMC Bioinformatics **4**: 41.

Tatusov, R. L., E. V. Koonin, et al. (1997). "A genomic perspective on protein families." Science **278**(5338): 631-7.

Ulrich, L. E., E. V. Koonin, et al. (2005). "One-component systems dominate signal transduction in prokaryotes." Trends Microbiol **13**(2): 52-6.

Ulrich, L. E. and I. B. Zhulin (2007). "MiST: a microbial signal transduction database." Nucleic Acids Res **35**(Database issue): D386-90.

van Nimwegen, E. (2003). "Scaling laws in the functional content of genomes." Trends Genet **19**(9): 479-84.

Venkateswaran, K., M. E. Dollhopf, et al. (1998). "Shewanella amazonensis sp. nov., a novel metal-reducing facultative anaerobe from Amazonian shelf muds." Int J Syst Bacteriol **48 Pt 3**: 965-72.

Wadhams, G. H. and J. P. Armitage (2004). "Making sense of it all: bacterial chemotaxis." Nat Rev Mol Cell Biol **5**(12): 1024-37.

Wang, F., J. Wang, et al. (2008). "Environmental adaptation: genomic analysis of the piezotolerant and psychrotolerant deep-sea iron reducing bacterium Shewanella piezotolerans WP3." PLoS ONE **3**(4): e1937.

Zhulin, I. B., A. N. Nikolskaya, et al. (2003). "Common extracellular sensory domains in transmembrane receptors for diverse signal transduction pathways in bacteria and archaea." J Bacteriol **185**(1): 285-94.

Zhulin, I. B., B. L. Taylor, et al. (1997). "PAS domain S-boxes in Archaea, Bacteria and sensors for oxygen and redox." Trends Biochem Sci **22**(9): 331-3.

APPENDIX

Appendix

Appendix A. Core Conserved Signal Transduction Proteins and Descriptions from *Shewanella oneidensis* MR-1

Gene Locus	COG Symbol	Gene Symbol	Description	COG Description
SO4742	GlpR	SO4742	Transcriptional regulator, DeoR family	Transcriptional regulators of sugar metabolism
SO4711	COG2206	SO4711	HD domain protein	HD-GYP domain
SO4705	HipB	SO4705	Transcriptional regulator, putative	Predicted transcriptional regulator protein
SO4675	AcrR	SO4675	Transcriptional regulator, TetR family	Transcriptional regulator
SO4647	OmpR	SO4647	DNA-binding response regulator	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain
SO4635	Tar	SO4635	Methyl-accepting chemotaxis protein	Methyl-accepting chemotaxis protein
SO4634	BaeS	envZ	Osmolarity sensor protein EnvZ	Signal transduction histidine kinase
SO4633	OmpR	ompR	Transcriptional regulatory protein OmpR	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain
SO4556	LysR	SO4556	Transcriptional regulator, LysR family	Transcriptional regulator
SO4478	BaeS	cpxA	Sensor protein CpxA	Signal transduction histidine kinase
SO4477	OmpR	cpxR	Transcriptional regulatory protein CpxR	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain
SO4472	AtoC	ntrC	Nitrogen regulation protein NR(I)	Response regulator containing CheY-like receiver AAA-type ATPase and DNA-binding domains
SO4471	NtrB	ntrB	Nitrogen regulation protein	Signal transduction histidine kinase nitrogen specific
SO4454	Tar	SO4454	Methyl-accepting chemotaxis protein	Methyl-accepting chemotaxis protein
SO4428	OmpR	SO4428	DNA-binding response regulator	Response regulators consisting of a CheY-

				like receiver domain and a winged-helix DNA-binding domain
SO4427	BaeS	SO4427	Sensor histidine kinase	Signal transduction histidine kinase
SO4350	LysR	ilvY	Transcriptional regulator ilvY	Transcriptional regulator
SO4323	Rtn	SO4323	GGDEF domain protein	FOG: EAL domain
SO4251	AcrR	slmA	HTH-type protein slmA	Transcriptional regulator
SO4172	COG4567	SO4172	DNA-binding response regulator	Response regulator consisting of a CheY-like receiver domain and a Fis-type HTH domain
SO4116	Rtn	mshH	MSHA biogenesis protein MshH	FOG: EAL domain
SO3988	OmpR	arcA	Aerobic respiration control protein ArcA	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain
SO3982	CitB	SO3982	DNA-binding nitrate/nitrite response regulator	Response regulator containing a CheY-like receiver domain and an HTH DNA-binding domain
SO3838	Tar	SO3838	Methyl-accepting chemotaxis protein	Methyl-accepting chemotaxis protein
SO3799	Lrp	asnC	Regulatory protein AsnC	Transcriptional regulators
SO3684	AcrR	SO3684	Transcriptional regulator, TetR family	Transcriptional regulator
SO3660	FhlA	SO3660	Sigma-54 dependent transcriptional regulator/sensory box protein	Transcriptional regulator containing GAF AAA-type ATPase and DNA binding domains
SO3642	Tar	SO3642	Methyl-accepting chemotaxis protein	Methyl-accepting chemotaxis protein
SO3595	BaeS	SO3595	Sensor protein RstB, putative	Signal transduction histidine kinase
SO3582	Tar	SO3582	Methyl-accepting chemotaxis protein	Methyl-accepting chemotaxis protein
SO3538	ArsR	hlyU	Transcriptional regulator HlyU	Predicted transcriptional regulator protein
SO3516	PurR	SO3516	Transcriptional regulator, LacI family	Transcriptional regulators
SO3426	CsrA	csrA	Carbon storage regulator homolog	Carbon storage regulator (could also regulate swarming and quorum sensing)
SO3419	TrpR	trpR	Trp operon repressor	Trp operon repressor
SO3393	AcrR	SO3393	Transcriptional regulator, TetR	Transcriptional

			family	regulator
SO3277	AcrR	SO3277	Transcriptional regulator, TetR family	Transcriptional regulator
SO3252	CheW	cheV-3	Chemotaxis protein CheV	Chemotaxis signal transduction protein
SO3251	CheR	cheR-2	Chemotaxis protein methyltransferase CheR	Methylase of chemotaxis methyl-accepting protein
SO3232	AtoC	flrA	Flagellar regulatory protein A	Response regulator containing CheY-like receiver AAA-type ATPase and DNA-binding domains
SO3230	AtoC	flrC	Flagellar regulatory protein C	Response regulator containing CheY-like receiver AAA-type ATPase and DNA-binding domains
SO3209	AtoC	cheY-3	Chemotaxis protein CheY	Response regulator containing CheY-like receiver AAA-type ATPase and DNA-binding domains
SO3208	CheZ	cheZ	Chemotaxis protein CheZ	Chemotaxis protein
SO3206	CheB	cheB-3	Chemotaxis response regulator protein-glutamate methylesterase group 1 operon (EC 3.1.1.61), Chemotaxis response regulator protein-glutamate methylesterase of group 1 operon	Chemotaxis response regulator containing a CheY-like receiver domain and a methylesterase domain
SO3202	CheW	cheW-3	Purine-binding chemotaxis protein CheW	Chemotaxis signal transduction protein
SO3196	AtoC	SO3196	Response regulator	Response regulator containing CheY-like receiver AAA-type ATPase and DNA-binding domains
SO3123	CheW	cheV-2	Chemotaxis protein CheV	Chemotaxis signal transduction protein
SO3084	COG5001	SO3084	Sensory box protein	Predicted signal transduction protein containing a membrane domain an EAL and a GGDEF domain
SO2885	FadR	fadR	Fatty acid metabolism regulator protein	Transcriptional regulators
SO2862	COG2206	SO2862	HDIG domain protein	HD-GYP domain
SO2852	GntR	SO2852	Transcriptional regulator, GntR family	Transcriptional regulators
SO2725	CitB	SO2725	Transcriptional regulator, LuxR family	Response regulator containing a CheY-like receiver domain and an HTH DNA-binding

				domain
SO2649	LysR	cysB	Cys regulon transcriptional activator	Transcriptional regulator
SO2640	MarR	SO2640	Transcriptional regulator, MarR family	Transcriptional regulators
SO2603	COG1956	SO2603	Hypothetical protein, Hypothetical protein SO2603	GAF domain-containing protein
SO2507	Rtn	SO2507	GGDEF domain protein	FOG: EAL domain
SO2493	AcrR	SO2493	Transcriptional regulator, TetR family	Transcriptional regulator
SO2490	RpiR	SO2490	Transcriptional regulator, RpiR family	Transcriptional regulators
SO2485	Dgt	SO2485	Deoxyguanosinetriphosphate triphosphohydrolase-like protein	dGTP triphosphohydrolase
SO2484	COG1896	SO2484	Hypothetical UPF0207 protein SO2484, UPF0207 protein SO2484	Predicted hydrolase of HD superfamily
SO2438	LysR	SO2438	Transcriptional regulator, LysR family	Transcriptional regulator
SO2305	Lrp	lrp	Leucine-responsive regulatory protein	Transcriptional regulators
SO2263	COG1959	SO2263	Rrf2 family protein	Predicted transcriptional regulator protein
SO2202	LysR	SO2202	Transcriptional regulator, LysR family	Transcriptional regulator
SO2197	COG2199	SO2197	GGDEF family protein	FOG: GGDEF domain
SO2053	LysR	SO2053	Transcriptional regulator, LysR family	Transcriptional regulator
SO2049	PleD	SO2049	GGDEF family protein	Response regulator containing a CheY-like receiver domain and a GGDEF domain
SO1989	CheW	cheV-1	Chemotaxis protein CheV	Chemotaxis signal transduction protein
SO1965	LysR	SO1965	Transcriptional regulator, LysR family	Transcriptional regulator
SO1937	Fur	fur	Ferric uptake regulation protein	Fe2+/Zn2+ uptake regulation protein
SO1898	SoxR	SO1898	Transcriptional regulator, putative	Predicted transcriptional regulator protein
SO1860	CitB	SO1860	DNA-binding response regulator, LuxR family	Response regulator containing a CheY-like receiver domain and an HTH DNA-binding domain
SO1806	FhlA	pspF	Psp operon transcriptional activator	Transcriptional regulator containing GAF AAA-type ATPase and DNA binding domains
SO1669	TyrR	tyrR	Transcriptional regulatory protein TyrR	Transcriptional regulator of aromatic amino acids

				metabolism
SO1646	COG2199	SO1646	GGDEF family protein	FOG: GGDEF domain
SO1559	VicK	phoR	Phosphate regulon sensor protein PhoR	Signal transduction histidine kinase
SO1558	OmpR	phoB	Phosphate regulon response regulator PhoB	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain
SO1551	COG2199	SO1551	GGDEF domain protein	FOG: GGDEF domain
SO1533	LysR	SO1533	Glycine cleavage system transcriptional activator, putative	Transcriptional regulator
SO1338	LysR	nhaR	Transcriptional activator protein NhaR	Transcriptional regulator
SO1332	PtsP	ptsP	Phosphoenolpyruvate-protein phosphotransferase PtsP	Signal transduction protein containing GAF and PtsI domains
SO1328	LysR	SO1328	Transcriptional regulator, LysR family	Transcriptional regulator
SO1278	Tar	SO1278	Methyl-accepting chemotaxis protein	Methyl-accepting chemotaxis protein
SO1208	COG5001	SO1208	GGDEF domain protein	Predicted signal transduction protein containing a membrane domain an EAL and a GGDEF domain
SO0997	LysR	SO0997	Transcriptional regulator, LysR family	Transcriptional regulator
SO0860	COG3437	SO0860	Response regulator	Response regulator containing a CheY-like receiver domain and an HD-GYP domain
SO0839	LysR	SO0839	Transcriptional regulator, LysR family	Transcriptional regulator
SO0817	LysR	metR	Transcriptional activator protein MetR	Transcriptional regulator
SO0769	ArgR	argR	Arginine repressor	Arginine repressor
SO0624	Crp	crp	Catabolite gene activator	cAMP-binding protein - catabolite gene activator and regulatory subunit of cAMP-dependent protein kinase
SO0570	AtoC	SO0570	Response regulator	Response regulator containing CheY-like receiver AAA-type ATPase and DNA-binding domains
SO0443	SoxR	zntR	Transcriptional regulator, MerR family	Predicted transcriptional regulator protein
SO0423	FadR	pdhR	Pyruvate dehydrogenase complex repressor	Transcriptional regulators

SO0393	Fis	fis	DNA-binding protein fis	Factor for inversion stimulation Fis transcriptional activator
SO0346	GntR	SO0346	Transcriptional regulator. GntR family	Transcriptional regulators
SO0214	BirA	birA	BirA bifunctional protein	Biotin-(acetyl-CoA carboxylase) ligase
SO0198	AcrR	SO0198	Transcriptional regulator, TetR family	Transcriptional regulator
SO0096	PhnF	hutC	Histidine utilization repressor	Transcriptional regulators
SO0045	COG1959	SO0045	Rrf2 family protein	Predicted transcriptional regulator protein
SO0026	ArsR	SO0026	Transcriptional regulator, ArsR family	Predicted transcriptional regulator protein

Appendix B. Proteins in the Significant Similarity Group of *S. oneidensis* MR-1 and Descriptions

24372126	964234	arsR	arsenical resistence operon repressor [Shewanella oneidensis MR-1]
24373681	965666	cheA	chemotaxis protein CheA [Shewanella oneidensis MR-1]
24373686	965671	cheB-1	protein-glutamate methylesterase CheB [Shewanella oneidensis MR-1]
24373685	965670	cheD-1	chemotaxis protein CheD [Shewanella oneidensis MR-1]
24373682	965667	cheW-1	purine-binding chemotaxis protein CheW [Shewanella oneidensis MR-1]
24373680	965665	cheY-1	chemotaxis protein CheY [Shewanella oneidensis MR-1]
24373867	965835	cheY-2	chemotaxis protein CheY [Shewanella oneidensis MR-1]
24374654	966590	dctD	C4-dicarboxylate transport transcriptional regulatory protein [Shewanella oneidensis MR-1]
24373903	965868	etrA	electron transport regulator a [Shewanella oneidensis MR-1]
24374743	966683	flrB	flagellar regulatory protein B [Shewanella oneidensis MR-1]
24373194	965229	glnD	PII uridylyl-transferase [Shewanella oneidensis MR-1]
24374395	966335	iciA	chromosome replication initiation inhibitor protein [Shewanella oneidensis MR-1]
24371659	963795	kdpE	transcriptional regulatory protein KdpE [Shewanella oneidensis MR-1]
24375453	967339	mgtE-2	magnesium transporter [Shewanella oneidensis MR-1]
24375351	967244	modE	molybdenum transport regulatory protein ModE [Shewanella oneidensis MR-1]
24375468	967350	narQ	nitrate/nitrite sensor protein NarQ [Shewanella oneidensis MR-1]
24373510	965510	phoP	transcriptional regulatory protein PhoP [Shewanella oneidensis MR-1]
24373509	965509	phoQ	sensor protein PhoQ [Shewanella oneidensis MR-1]
24372399	964490	rbsK	ribokinase [Shewanella oneidensis MR-1]
24372921	964981	rseA	sigma-E factor negative regulatory protein [Shewanella oneidensis MR-1]
24372809	964874	torR	torcad operon transcriptional regulatory protein TorR [Shewanella oneidensis MR-1]
24372811	964876	torS	sensor histidine kinase/response regulator TorS [Shewanella oneidensis MR-1]
24372123	964232	trpI	trpba operon transcriptional activator [Shewanella oneidensis MR-1]
24375423	967311	vacB	ribonuclease R [Shewanella oneidensis MR-1]

Appendix C.1 Probable ‘Unknown’ Protein Domain Annotations

Pfam Domain Name	Count
HAMP	40
PAS	36
HisKA	34
PAS_4	21
AraC_binding	14
Cache_1	14
GAF	14
MarR	14
SBP_bac_3	14
LysR_substrate	11
HATPase_c	10
OB_RNB	8
PAS_3	8
GGDEF	6
NIT	6
Response_reg	5
DPPIV_N	4
HD	4
HTH_3	4
MCPsignal	4
HhH-GPD	2
NTP_transf_2	2
TOBE	2
Aminotran_1_2	1
CBS	1
DSPc	1
EAL	1
FCD	1
GerE	1
HTH_11	1
HTH_5	1
Peripla_BP_1	1
Trans_reg_C	1
cNMP_binding	1
Total	248

Appendix C.2 Results from Automated PSI-BLAST Search of 'Unknown' Domains

Hit counts of known domains to unknown sequences.

Domain	Hit Count
HTH_8	184
PAS	181
PAS_4	168
TetR_N	158
Reg_prop	152
HTH_AraC	129
HAMP	128
HisKA	117
GGDEF	116
TetR_C_3	110
Cache_1	108
PAS_3	105
TetR_C_2	104
GAF	100
HDOD	99
MCPsignal	97
Response_reg	92
TPR_1	88
LysR_substrate	88
MerR-DNA-bind	87
Sigma54_activat	82
MerR	76
TPR_2	75
HTH_5	72
MarR	67
TPR_4	65
Crp	59
SBP_bac_3	59
TPR_3	55
AraC_binding	53
TrmB	49
LacI	46
HTH_7	46
Cache_2	45
Sel1	45
AT_hook	44
HTH_11	43
HD	42
DUF1956	39

PPR	38
TetR_C_4	35
CheC	33
DUF24	26
RNB	22
MCP_N	21
TarH	21
TetR_C_5	21
HTH_1	21
DUF955	20
Y_Y_Y	18
LRR_1	18
Tetradecapep	16
GFO_IDH_MocA	16
CBS	15
DAGK_acc	15
PrpR_N	14
Sigma70_r4_2	14
STAS	13
HATPase_c	13
Rrf2	13
DPPIV_N	13
NSF	12
SpoIIE	12
Acyl-CoA_dh_N	12
HTH_DeoR	12
HTH_10	12
TonB_dep_Rec	11
BPL_C	11
TetR_C	11
HTH_Mga	11
Ubie_methyltran	10
Peripla_BP_1	10
ABC_tran	10
zf-B_box	10
PD40	10
LRR_2	10
Hpt	9
LexA_DNA_bind	9
NIT	8
SGL	8
CheD	8
KAP_NTPase	8
DAGK_cat	8
Pencillinase_R	8
LRRNT	8
CHASE3	8

AlkA_N	8
CheR	7
PaaX	7
PadR	7
VGCC_alpha2	7
GerE	7
SMC_N	6
SBP_bac_1	6
HTH_3	5
ABC_sub_bind	5
HisKA_2	5
MASE1	5
AraC_N	5
WD40	5
PPAK	5
PT	4
GSPII_E	4
PKD	4
EAL	4
His_biosynth	4
NB-ARC	4
NodS	3
NNMT_PNMT_TEMT	3
DJ-1_PfpI	3
Ada_Zn_binding	3
BPD_transp_1	3
Methyltransf_1N	3
MORN_2	3
H-kinase_dim	3
Extensin_1	3
AAA_5	3
DEAD_2	3
Pkinase	3
HhH-GPD	3
Sigma70_r4	3
Methyltransf_2	2
LeuA_dimer	2
NMT1	2
DNA_binding_1	2
7TMR-DISM_7TM	2
OGFr_III	2
OpuAC	2
AraC_E_bind	2
HWE_HK	2
2CSK_N	1
HemolysinCabind	1
WIF	1

FAINT	1
GlnD_UR_UTase	1
Phytochrome	1
RCSD	1
ACT	1
RNA_pol_Rpb1_R	1
Wzz	1
Peptidase_U32	1
HMA	1
DUF258	1
SSF	1
Peripla_BP_2	1
Phage_CI_repr	1
Filament	1
Homeobox	1
ELK	1
MEKHLA	1

Appendix D. Gene Ontology Annotations

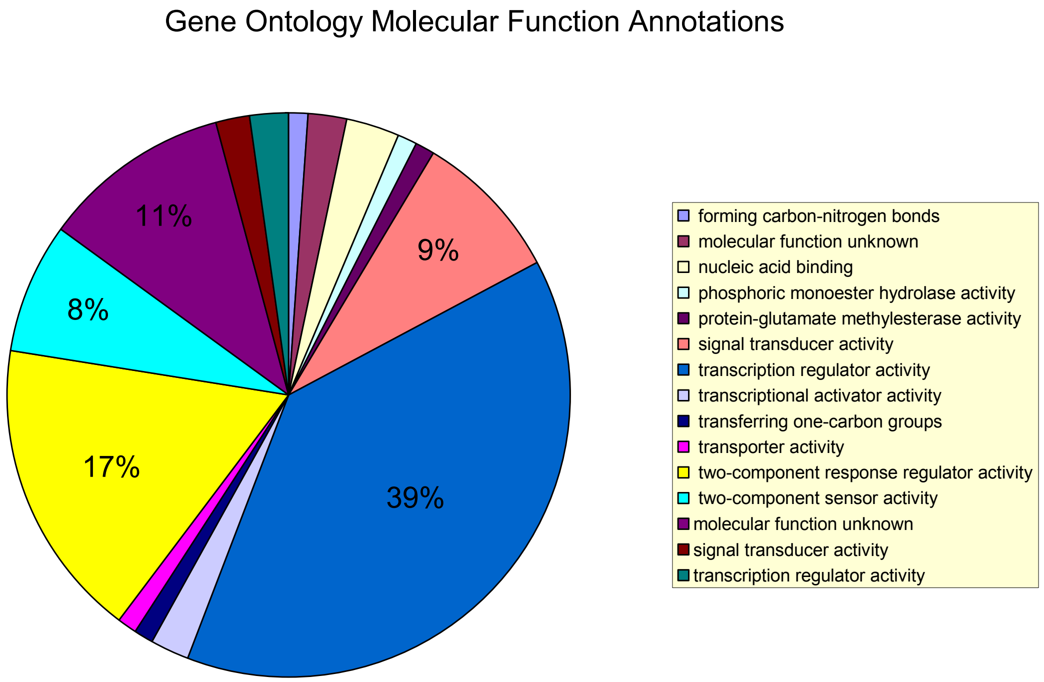

Figure 9. GO Molecular Functions Annotations.

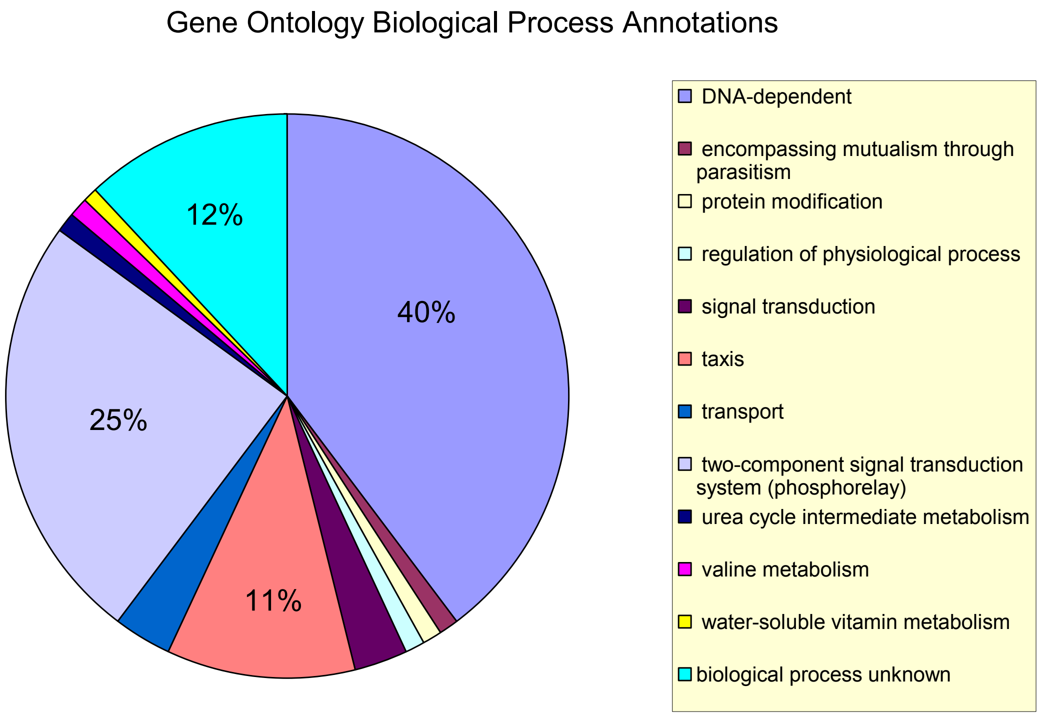

Figure 10. GO Biological Process Annotations.

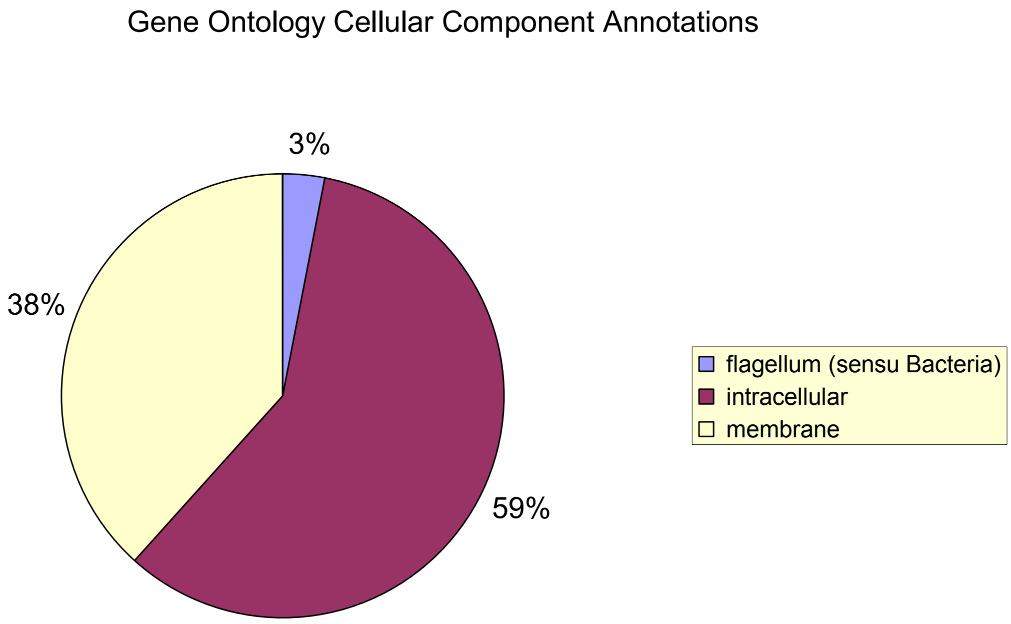

Figure 11. GO Cellular Component Annotations.

Appendix E: GO Annotations for Proteins in Core Group in *Shewanella oneidensis* MR-1

Protein GI	Molecular Function	Biological Process	Cellular Component
24371626	transcription regulator activity	DNA-dependent	intracellular
24371645	molecular function unknown	biological process unknown	
24371696	transcription regulator activity	DNA-dependent	intracellular
24371798	transcription regulator activity	DNA-dependent	
24371812	forming carbon-nitrogen bonds	water-soluble vitamin metabolism	
24371989	transcription regulator activity	DNA-dependent	
24372018	transcription regulator activity	DNA-dependent	intracellular
24372038	transcription regulator activity	DNA-dependent	intracellular
24372163	two-component response regulator activity	two-component signal transduction system (phosphorelay)	
24372215	transcription regulator activity	DNA-dependent	intracellular
24372358	transcription regulator activity	urea cycle intermediate metabolism	
24372406	transcription regulator activity	DNA-dependent	
24372428	transcription regulator activity	DNA-dependent	
24372790	signal transducer activity	signal transduction	membrane
24372859	signal transducer activity	taxis	membrane
24372906	transcription regulator activity	DNA-dependent	
24372910	transporter activity	transport	intracellular
24372916	transcription regulator activity	transport	
24373106	transcription regulator activity	DNA-dependent	
24373122	molecular function unknown	biological process unknown	
24373128	two-component response regulator activity	two-component signal transduction system (phosphorelay)	
24373129	two-component sensor activity	two-component signal transduction system (phosphorelay)	membrane
24373214	molecular function unknown	biological process unknown	
24373237	transcription regulator activity	DNA-dependent	

24373371	transcriptional activator activity	DNA-dependent	
24373425	two-component response regulator activity	two-component signal transduction system (phosphorelay)	intracellular
24373463	transcription regulator activity	DNA-dependent	intracellular
24373501	transcription regulator activity	DNA-dependent	
24373529	transcription regulator activity	DNA-dependent	
24373553	two-component response regulator activity	two-component signal transduction system (phosphorelay)	intracellular
24373609	molecular function unknown	biological process unknown	
24373752	molecular function unknown	biological process unknown	
24373757	transcription regulator activity	DNA-dependent	
24373816	nucleic acid binding	DNA-dependent	
24373857	transcription regulator activity	transport	intracellular
24373985	transcription regulator activity	DNA-dependent	
24374028	molecular function unknown	biological process unknown	
24374029	triphosphoric monoester hydrolase activity	primary metabolism	
24374034	transcription regulator activity	DNA-dependent	
24374037	transcription regulator activity	DNA-dependent	
24374051	molecular function unknown	biological process unknown	
24374146	molecular function unknown	biological process unknown	
24374181	transcription regulator activity	DNA-dependent	intracellular
24374190	transcription regulator activity	DNA-dependent	
24374266	transcription regulator activity	DNA-dependent	intracellular
24374381	transcription regulator activity	DNA-dependent	intracellular
24374391	molecular function unknown	biological process unknown	
24374414	transcription regulator activity	DNA-dependent	
24374604	two-component sensor activity	two-component signal transduction system (phosphorelay)	
24374641	two-component response regulator activity	two-component signal transduction system (phosphorelay)	intracellular

24374708	two-component response regulator activity	two-component signal transduction system (phosphorelay)	
24374714	signal transducer activity	taxis	intracellular
24374718	protein-glutamate methylesterase activity	taxis	
24374720	molecular function unknown	taxis	flagellum (sensu Bacteria)
24374721	two-component response regulator activity	two-component signal transduction system (phosphorelay)	
24374742	two-component response regulator activity	two-component signal transduction system (phosphorelay)	
24374744	transcriptional activator activity	DNA-dependent	
24374762	transferring one-carbon groups	taxis	
24374763	two-component response regulator activity	two-component signal transduction system (phosphorelay)	intracellular
24374788	transcription regulator activity	DNA-dependent	
24374904	transcription regulator activity	DNA-dependent	
24374929	transcription regulator activity	DNA-dependent	intracellular
24374936	nucleic acid binding	regulation of physiological process	
24375020	transcription regulator activity	DNA-dependent	intracellular
24375042	transcription regulator activity	DNA-dependent	intracellular
50261353	signal transducer activity	taxis	membrane
24375141	signal transducer activity	taxis	membrane
24375159	two-component sensor activity	two-component signal transduction system (phosphorelay)	
24375182	transcription regulator activity	DNA-dependent	
24375292	transcription regulator activity	DNA-dependent	intracellular
24375328	signal transducer activity	taxis	membrane
24375469	two-component response regulator activity	two-component signal transduction system (phosphorelay)	intracellular
24375475	two-component response regulator activity	two-component signal transduction system (phosphorelay)	
24375602	molecular function unknown	encompassing mutualism through parasitism	
24375658	two-component response regulator activity	two-component signal transduction system (phosphorelay)	

24375735	transcription regulator activity	DNA-dependent	
24375805	signal transducer activity	signal transduction	membrane
24375831	transcription regulator activity	valine metabolism	
24375905	two-component sensor activity	two-component signal transduction system (phosphorelay)	membrane
24375906	two-component response regulator activity	two-component signal transduction system (phosphorelay)	
24375932	signal transducer activity	taxis	membrane
24375949	two-component sensor activity	two-component signal transduction system (phosphorelay)	membrane
24375950	two-component response regulator activity	two-component signal transduction system (phosphorelay)	
24375955	two-component response regulator activity	two-component signal transduction system (phosphorelay)	
24375956	two-component sensor activity	two-component signal transduction system (phosphorelay)	membrane
24376030	transcription regulator activity	DNA-dependent	
24376106	two-component response regulator activity	two-component signal transduction system (phosphorelay)	
24376107	two-component sensor activity	two-component signal transduction system (phosphorelay)	membrane
24376108	signal transducer activity	taxis	membrane
24376120	two-component response regulator activity	two-component signal transduction system (phosphorelay)	
24376147	transcription regulator activity	DNA-dependent	
24376177	nucleic acid binding	biological process unknown	
24376183	molecular function unknown	biological process unknown	
24376214	transcription regulator activity	DNA-dependent	intracellular

Appedix F Paralogous Domain Data.

Paralogous Domain ID	amazonensis SB2B	ANA-3	denitrificans OS217	frigidimarina NCIMB 400	loihica PV-4	MR-4	MR-7	oneidensis	putrefaciens CN-32	W3-18-1
128								2		
926	2								2	2
927	2	2			3	2	2		2	2
948							2		2	2
949		2					2		2	2
950									2	2
951		2								
1056		2			2				2	2
1253				2						
1818				2						
1955			2							
2186				2				2		
2187				2				2		
2325							2			
2546				2						
2647		2								
2648		2								
2790									2	
3512	2			2						
3518		2				2	2	2		
3527		2				2	2		2	2
3554		2		2		2	2	2	3	3
3943				2						
4488		2			2	2	2			
4506					2					
4696									2	
4748				2						
4766					2					
4767					2					
4897									2	2
4963								2		
5096	2	2			2	2		2	2	2
5247	2	2				2	2			
5357				2						
5358	2									
5359	2									
5424	2	2			2	2	2	2		
5456									2	
5493								2		
5522	2	2								
5535					2					

5677				2						
5742	2	2		2	2		2			
5921	2	2	2		2					
5948	2									
5961		2			2					
5995		2				2	2		2	2
6038	2	2	2			2	2			
6040	2	2	2			2	2			
6215							2			
6295									2	
6702										2
6801		3			2	2			2	3
7708									2	3
7922				2						
7983					2					
Grand Total	28	43	8	26	29	24	30	18	35	31

VITA

Harold Arthur Shanafield III was born January 16, 1975 in West Berlin, West Germany. At the age of 3 he moved with his parents to Evanston, Illinois. As a teenager, he attended Evanston Township High School where was a National Merit Scholar and made the honor roll several semesters. Active in many extracurricular activities, Harold played in the band, played several sports including earning a varsity letter and captaining the volleyball team, and was a sports editor for the school newspaper.

Based on his high school achievements, Harold earned a scholarship to attend Northwestern University. He double majored in Biology and Environmental Science and earned a degree in four years. He was again active in extracurricular activities as a brother in the Delta Tau Delta fraternity and a member of the championship intramural basketball team his junior year.

After graduation Harold accepted a job as a Vendor Manager for the internet startup pcOrder.com. After a year-and-a-half of successfully demonstrating his abilities, Harold was promoted to the position of Program Manager in the Content Division. Harold successfully led pcOrder.com's development of the office products data market covering information on over 25000 products. Harold was also responsible for initiating the partnership and integration between pcOrder.com and Deja.com, a leading internet comparison shopping site.

After three-and-a-half years at pcOrder.com, tough times struck many dot.com's and pcOrder.com was acquired by its parent company Trilogy. Harold left the new company and worked briefly for Bearingpoint developing their automated archiving systems before returning to school. Harold chose to attend the Genome Science and Technology program run jointly by the University of Tennessee and the Oak Ridge National Lab. Harold completed a Master's of Science in Statistics in addition to the Master's of Science in the Life Sciences department with a focus on bioinformatics.

Harold currently resides in Knoxville, TN with his wife of seven years and his 3 year-old daughter.