

Figure 1. Modular high-throughput platform for fast and parallel total chemical synthesis, mass-spectrometric purification and single-molecule spectroscopic assay to annotate function for newly predicted proteins.

We have previously described the use of x,y,z robotics and laboratory automation and efficient Fmoc chemistry SPPS protocols for the simultaneous parallel synthesis of the key peptide-thioester building blocks needed for chemical protein synthesis. This made use of a recently reported novel resin linker [Blanco-Canosa JB, Dawson PE: An efficient Fmoc-SPPS approach for the generation of thioester peptide precursors for use in native chemical ligation. *Angew Chem Int Ed Engl.* 2008, 47:6851-5]. Typical data are shown in Figure 2 (Top).

Validation of Genome Sequence Annotation

223

Robotic Chemical Protein Synthesis for the Experimental Validation of the Functional Annotation of Microbial Genomes

Stephen Kent and Kalyaneswar Mandal* (kmandal@uchicag.edu)

Institute for Biophysical Dynamics, University of Chicago, Ill.

Project Goals: Robotic total chemical synthesis to make proteins and protein domains, for the validation of functional annotation of predicted open reading frames.

Modern total protein synthesis has evolved from the 'chemical ligation' methods introduced by the Kent laboratory [Kent SBH. Total chemical synthesis of proteins. *Chemical Society Reviews* 2009; 38: 338-51.]. Unprotected synthetic peptide segments, spanning the amino acid sequence of the mature polypeptide chain derived from a predicted open reading frame, are covalently joined to one another by chemo-selective reaction. Native chemical ligation, the thioester-mediated covalent bond-forming chemoselective reaction of unprotected peptides at a Cys residue, is the most robust and useful ligation chemistry developed to date. The synthetic protein is then used to experimentally validate the predicted biochemical function, and in selected cases to determine the Xray structure of the protein molecule (Figure).

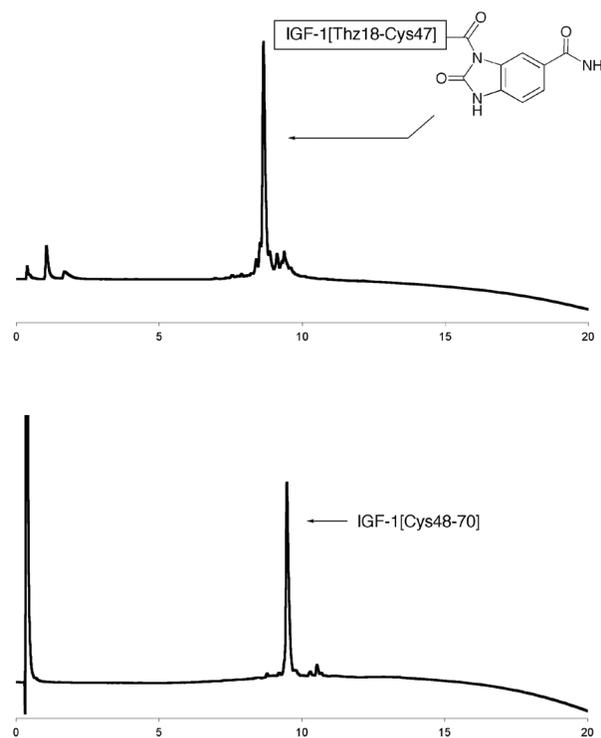


Figure 2. Automated robotic Fmoc SPPS preparation of peptide-thioesters. HPLC-electrospray MS of crude products are shown.

Ready preparation of peptide-thioesters enables the straightforward total chemical synthesis of proteins by native chemical ligation. Proof-of-concept total chemical

syntheses of predicted proteins from microbial and plant genomes will be presented.

224

Using Deep RNA Sequencing for the Structural Annotation of the *Laccaria bicolor* Mycorrhizal Transcriptome

Peter E. Larsen,^{1*} Geetika Trivedi,² Avinash Sreedasyam,² Vincent Lu,¹ Gopi K. Podila,² and Frank R. Collart¹ (fcollart@anl.gov)

¹Biosciences Division, Argonne National Laboratory, Lemont, Ill. and ²Dept. of Biological Sciences, University of Alabama, Huntsville

http://www.bio.anl.gov/molecular_and_systems_biology/proteins.html

Project Goals: To facilitate the process for acquisition of function from complex environmental sequence data sets, we developed methods to utilize RNA-seq data to validate current gene model intron-exon boundary, correct errors in the structural annotation and extend the boundaries of the current gene models using assembly approaches.

Advances in sequences technology have enabled deep sequence interrogation of individual organisms as well as complex systems. This capability has led to an improved appreciation of the biological diversity associated with specific ecosystems and the complexity of the molecular systems involved in perception and response to external stimuli. Mapping these signaling pathways is challenging however in sequence data sets from environmental and/or metagenomic projects where uncharacterized organisms often represent a high proportion of the sequence data. To facilitate the process for acquisition of function from complex environmental sequence data sets, we are developing methods to utilize RNA-seq data to correct errors in the structural annotation and extend the boundaries of current gene models using assembly approaches. To validate the methods, we used a transcriptomic data set derived from the fungus *Laccaria bicolor* which develops a mycorrhizal symbiotic association with the roots of many tree species, in which the fungus provides nutrients to the tree in exchange for photosynthetically-derived sugars. This fungal-plant symbiosis is a widespread process of major ecological importance and knowledge of the molecular events and expressed protein sequences associated with the development of the mycorrhizal system is essential for our understanding of natural biological processes related to carbon sequestration, carbon management, sustainability and bioenergy.

We generated >30 million RNA-seq reads from *Laccaria* grown in culture. Our study used the 20614 gene “best model” set and 65-megabase *Laccaria* genomic DNA sequence from the publically available FTP site at the Joint Genome Institute. Our analysis focused on the subset of 1501 gene models that are differentially expressed in the mycorrhizal transcriptome and are expected to be important

elements related to carbon metabolism, membrane permeability and transport, and intracellular signaling.

Our analysis of the intron-exon boundaries in current JGI best gene model set indicates the quality of *L. bicolor* structural annotation is enabling for homology-based comparison applications, but has severe limitations for experimental studies. For every intron-exon boundary in JGI Best Model set, we generated an 18-mer ‘probe’ sequence consisting of 9 bp up and down-stream of the intron-exon boundary. This intron-spanning sequence was used to search the set of RNA-seq reads. At least one read-containing probe was considered validation of gene model intron-exon boundary. Using these criteria, we were able to validate ~80% of the intron-exon boundaries within the gene model boundaries. This level of validation is notable in view of the complexity of the fungal genome (*L. bicolor* genes contain an average of 5.4 introns) and the annotation limitations arising from the relatively small number of sequenced fungal genomes. However, the combination of the error rate and intron density means that 42% of the current gene models contain intron/exon boundaries that do not map to the mRNA sequence data. Also, 58% of gene model 5’ and/or 3’ boundaries did not agree with the collected transcriptomic data. Inaccurate representations of the protein coding sequence are a consequence of these inconsistencies. Accurate coding sequences are essential for experimental approaches to characterize protein function and also to enhance the utility of tools that enable identification cellular localization signals and functional domains. Substantial changes to predicted UTRs also affect the ability to predict the regulatory mechanisms of mycorrhizae-specific genes. To improve the experimental utility of the gene model set, we developed algorithms that use the RNA-seq data to extend the boundaries of the current gene model set where appropriate, identify those intron-exon boundaries that can be validated by the transcriptomic data, and to generate novel intron-exon boundaries to bridge those regions of the gene models that are not supported by RNA-seq data. This extended and bridged contiguous expressed sequence was then aligned to the genome using a modified Smith-Waterman algorithm to recover gene model’s structural annotation. Of the set of 1501 gene models, 1439 (96%) successfully generated modified gene models in which all error flags were successfully resolved and sequences aligned to genomic sequence. The remaining 4% (62 gene models) either had deviations from transcriptome data that could not be spanned or generated sequence that did not align to genomic sequence. We considered a gene model significantly changed if at least one of three criteria were met: 1) an inconsistency in the original gene model was successfully bridged and aligned to scaffold, 2) the revised gene model contained a change in the total number of exons, and/or 3) we observed an absolute change in expressed gene size of more than 10%. Based on application of these criteria to the set of 1439 revised models, 974 (69%) of gene models required changes to match the transcriptomic consensus sequence. Additionally, for 465 (31%) of the models in the original best gene model set, we did not detect any inconsistencies and therefore have independently confirmed the previously published ‘BestModel’ annotation. Of those 62 gene models that could not be adequately

validated by the method proposed here, a number appear to have multiple isoforms in the expressed transcriptome data identifying them as genes of potential biological interest.

The outcome of this process is a set of high confidence gene models that can be reliably used for experimental characterization of protein function. This improved annotation process can be extended to other important gene families and will facilitate the process to identify the molecular mechanisms leading to the development of the mycorrhizal symbiosis and its implications in improving carbon sequestration by poplar.

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357.

225 Molecular Approaches for Elucidation of Sensory and Response Pathways in Cells

Sarah Zerbs,^{1*} Sarah E. Giuliani,¹ Elizabeth Landorf,¹ Maureen McNulty,¹ William Studier,² and Frank R. Collart¹ (fcollart@anl.gov)

¹Biosciences Division, Argonne National Laboratory, Lemont, Ill. and ²Biology Division, Brookhaven National Laboratory, Upton, N.Y.

http://www.bio.anl.gov/molecular_and_systems_biology/proteins.html

Project Goals: This program addresses the hypothesis that cellular behavior can be modeled through an understanding of the biological interface with the environment and the cellular responses that originate from the cell/environment interaction. The long term objective of the program is to define cellular sensory and regulatory pathways that respond to environmental nutrients thereby facilitating a system-level model that predicts the cellular response to environmental conditions or changes.

Increased knowledge of protein function enhances our understanding of cellular functions and is ultimately required to model biological activities and systems. This program addresses the hypothesis that cellular behavior can be modeled through an understanding of the biological interface with the environment and the cellular responses that originate from the cell/environment interaction. The long term objective of the program is to define cellular sensory and regulatory pathways that respond to environmental nutrients thereby facilitating a system-level model that predicts the cellular response to environmental conditions or changes. The program uses a parallel strategy of technology development to improve capabilities for extraction of relevant biological information from the sequence data coupled to genome scale approaches for elucidation of protein function and cellular regulatory networks.

One aspect of this program will develop tools to bridge the gap between genomes and systems biology. Progress in sequencing technology has provided molecular validation of the diversity and complexity of environmental systems. However, sequencing capacity has far outpaced computational and experimental methods to fully utilize the genomic data. We are addressing this gap between DNA sequence and the ability to extract relevant biological information from the sequence data by the development of genome scale approaches for elucidation of protein function and cellular regulatory networks. These approaches utilize next generation sequencing technology and high throughput approaches to enable economical and efficient protein production and characterization.

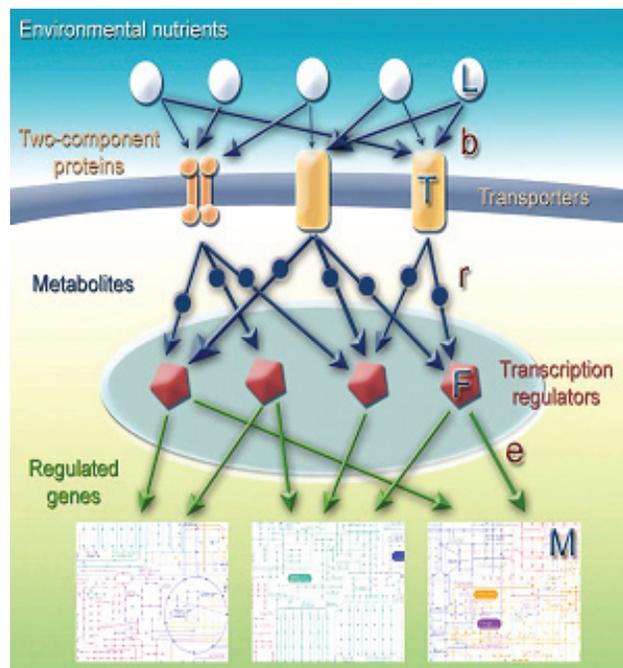


Fig 1. Illustration of experimental approach and application to systems modeling.

The capabilities to improve functional interrogation of sequences are coupled to *in vitro* methods for functional characterization of proteins involved in cellular sensory and response pathways. The functional screens will focus on key proteins that mediate communication between the cell and the environment such as transporters, two-component sensory systems, and membrane receptors (Fig. 1). This functional characterization will be linked to the cellular regulatory network by identification of the transcription factors whose activity is mediated by the environmental ligands or their metabolic derivatives. A coupling of the regulatory ligands with the DNA-binding regions of the transcription factors allows the association of metabolic pathways with the regulatory network. This genome scale process will determine the functional properties and potential of microbes and plants that are central to DOE missions. The functional assignments and ability to define specific sensory and regulatory pathways will increase the predictive capability of current models and support the development of

predictive systems-level models. This increased knowledge of the molecular components and control features of cellular sensory and response pathways is essential for our understanding of natural biological processes related to carbon management, sustainability and bioenergy.

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory (“Argonne”). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. Brookhaven National Laboratory is operated under Contract No. DE-AC02-98CH10886.

226

Functional Linkage of ABC Transporter Profile with Metabolic Capability in *Rhodopseudomonas palustris*

Sarah E. Giuliani,^{1*} Ashley M. Frank,¹ Catherine Seifert,¹ Lisa M. Miller,² Loren Hauser,³ and **Frank R. Collart**¹ (fcollart@anl.gov)

¹Biosciences Division, Argonne National Laboratory, Lemont, Ill.; ²National Synchrotron Light Source, Brookhaven National Laboratory, Upton, N.Y.; and ³Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://www.bio.anl.gov/molecular_and_systems_biology/proteins.html

Project Goals: We suggest that the functional profile of the genome set of transporter proteins is predictive of metabolic capabilities and ecological niche of organisms. To test this hypothesis, we profiled the genome set of ABC transporters for *Rhodopseudomonas palustris* CGA009 to define the relationship between the transporter profile and metabolic capability for *R. palustris* CGA009.

Transporter proteins are an organism’s primary interface with the environment. The expressed set of transporters mediates cellular metabolic capabilities and influences signal transduction pathways and regulatory networks. The role and impact of different transporters families differ in eukaryotic and prokaryotic organisms and the absolute number of transporters is dependent on the characteristics of the ecological niche. We suggest that the functional profile of the genome set of transporter proteins is predictive of metabolic capabilities and ecological niche of organisms. To test this hypothesis, we profiled the genome set of ABC transporters for *Rhodopseudomonas palustris* CGA009. In the *R. palustris*, ABC-type transporters represent approximately 45% of all transporters encoded in the genome. The ABC transporters family is widely distributed in soil organisms and can transport a variety of substrates such as metals, small ions, mono- and oligosaccharides, peptides, amino acids, iron-siderophores, polyamines, and vitamins.

An ABC transporter complex consists of a permease, ATPase, and a solute binding protein. The ligand specificity is determined by the solute binding protein which in

some cases can utilize multiple membrane permeases. The genome of *R. palustris* CGA009 encodes approximately 117 ABC type transporters as determined by the number of encoded solute binding proteins. The functional properties of these transport proteins are largely unknown and less than 10% have specific functional assignments. The largest group of binding proteins is annotated as “branched-chain amino acid” binding protein. To improve the utility of the function annotation, we expressed and purified the set of binding proteins from *R. palustris* and are characterizing ligand-binding specificity using ligand libraries consisting of environmental and cellular metabolic compounds and high throughput binding screens, including fluorescence thermal shift, small angle x-ray scattering, x-ray absorption spectroscopy, circular dichroism spectroscopy, and infrared spectroscopy. To date, this process resulted in the assignment of specific binding ligands for approximately 60% of the purified and screened proteins. In most cases, the binding was observed for specific compound classes and was observed for only 1-3 compounds from the entire ligand library. For approximately 20% of the screened proteins, a specific binding ligand was not observed, which we attribute to the limited scope of the screening library relative to the complexity of compounds in the natural environment.

The impact of these studies is two-fold. First, our screening method generated specific functional annotations for an important group of uncharacterized or incorrectly annotated transporter proteins. For example, six proteins encoded by genes annotated as branched chain amino acid binding proteins were demonstrated to bind various aromatic compounds derived from lignin degradation. Analysis of the flanking genomic regions reveals the co-localization of these transporter genes with metabolic genes associated with utilization of the transported compounds. Similar functional insight was obtained for previously uncharacterized proteins associated with the transport of fatty acids, dicarboxylic acids, oligopeptides, metals, and additional small molecule compounds. This functional insight can be used to improve the annotation of related organisms and provides a route to evaluate the evolution of the important and diverse group of transporter proteins.

Second, the results of this study also provide important biological insight for the metabolic capabilities and environment fitness of this organism. The profile and number of transport proteins specific for aromatic compounds is consistent with ecological and laboratory studies which demonstrate the capabilities of this organism for the utilization of plant degradation products such as lignin-derived aromatic compounds.

One of these binding proteins, RPA1385, showed high affinity and selectivity for vanadate, which is a catalytic component of a nitrogenase protein complex. *R. palustris* is a nitrogen fixing bacteria and has been shown to utilize a vanadium nitrogenase (V-nitrogenase) as a metabolic alternative when molybdenum is limited in the environment. Prior to this research, the cyanobacterium *Anabaena variabilis* (which also contains a V-nitrogenase) was the only organism known to contain a defined high-affinity vanadate

transport system. In *R. palustris*, genes RPA1381-1386 are annotated as components of a vanadate nitrogen fixation system based on homology to other similar proteins. However, in *R. palustris*, homology search approaches failed to identify the high-affinity vanadate transport system. Our ligand mapping approach identifies the RPA1385 protein as the vanadate SBP gene for this ABC transport system. This finding not only identifies a key component of the vanadate nitrogenase fixation pathway for this organism, but may also confirm a proposed hypothesis that the presence of this system in *R. palustris* suggests vanadate transport systems have evolved at least twice from dissimilar ancestral genes.

The functional assignments in conjunction with gene expression profiles and transcription factor DNA binding sites enable the identification of the cellular regulatory and metabolic components that enable the use of lignin degradation products for cell growth. This approach is being applied to other sequenced strains of *R. palustris* to provide evolutionary insight for the number and substrate specificity of this family of ABC type transporters. These capabilities will enable the identification and characterization of metabolic and regulatory pathways that are associated with a specific environmental niche.

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. Brookhaven National Laboratory is operated under Contract No. DE-AC02-98CH10886.

227

Phylogenomics-Guided Validation of Function for Conserved Unknown Genes

Valérie de Crécy-Lagard^{1*} (vcrecy@ufl.edu), Basma El Yacoubi,¹ Marc Bailly,¹ Ian K. Blaby,¹ Anne Pribat,² Aurora Lara-Núñez,² and Andrew D. Hanson²

¹Dept. of Microbiology and Cell Science and ²Dept. of Horticultural Sciences, University of Florida, Gainesville

Project Goals: Our overall goal is to establish an innovative integrative approach to predict and experimentally verify the in-vivo function of genes that lack homologs of known function ('unknown' gene families) and that are highly conserved among prokaryotes and plants. By implementing this approach we will predict, and experi-

mentally validate for a chosen subset, the function of ~1500 unknown genes.

Identifying the function of every gene in all sequenced organisms is a central challenge of the post-genomic era. We are submerged in genomic, transcriptomic, and proteomic data but the functions of about half (range 20 to 60%) of the genes in any given organism are still unknown. Our goal is to predict and experimentally verify the *in-vivo* function of proteins that lack homologs of known function ('unknown' protein families) and that are highly conserved between prokaryotes and plants. Our approach combines the extensive post-genomic resources of the plant field with the use of comparative genomic tools made possible by the availability of thousands of sequenced microbial genomes. This is an integrative approach to predict gene function whose early phase is computer-assisted, and whose later phases incorporate intellectual input from expert plant and microbial biochemists. It allows bridging of the gap between automated homology-driven annotations and the classical gene discovery efforts driven mainly by experimentalists. Our goal is to predict and experimentally validated the function of 15 "unknown" protein families". We have already validated predictions for seven families (in orange in Table 1 and we present the other eight most mature predictions that currently being tested (in yellow in Table 1). Two examples of this second list will be presented in more detail to emphasize the synergistic aspects of plant-microbe comparative genomics.

COG0799 proteins occur in plants, in nearly all bacteria, and in animals and fungi. Plants have two isoforms, one apparently chloroplastic, the other mitochondrial. The archetypal member of the family is the plant Iojap protein; *iojap* mutants of maize lack functional chloroplast ribosomes. In bacteria, COG0799 genes cluster strongly with the NAD synthesis gene *nadD* (nicotinate mononucleotide adenyltransferase) and sometimes the two genes are fused. COG0799 genes also cluster with genes encoding the ribosomal biogenesis protein ObgE and ribosomal proteins L21 and L27, making a connection with the ribosome lesion in the maize *iojap* mutant. Furthermore, transcriptomic data from *Arabidopsis* show co-expression of *iojap* with various chloroplast ribosome protein genes. NadD mediates a reaction in the *de novo* synthesis of NAD and potentially in salvage of nicotinamide mononucleotide (NMN). We therefore predict that Iojap catalyzes a process in ribosome biogenesis that releases NMN from NAD, and that the NMN is recycled by NadD. Possible Iojap reactions include a NAD-dependent DNA ligase-like reaction or an ADP-ribosyltransferase.

Case	Hypothesis	TAIR ID	COG, gene name	Subsystem in SEED	Experimental verification status	PubMed ID
1	Pterin carbinolamine dehydratase with role in Moco metabolism	At1g29810 At5g51110	COG2154, phhB	Pterin_carbinolamine_dehydratase	Validated in in 7 eukaryotes and 8 prokaryotes	18245455
2	t6A biosynthesis	At5g60590	COG0009, YrdC	YrdC-YciO	Validated in Yeast, Archaea and two bacteria;	19287007
3	PTPS family protein replacing the FolB step in folate synthesis	-	COG0720	Experimental-PTPS	Validated in 1 eukaryote and 8 prokaryotes	19395485, 18805734
4	Metal chaperone-Zinc homeostasis	At1g15730, At1g26520, At1g80480	COG0523	COG0523	Validated in several bacteria	19822009
5	Folate-dependent Fe/S cluster synthesis or repair protein	At4g12130 At1g60990	COG0354, ygfZ	YgfZ-Fe-S	Validated in <i>E. coli</i> , <i>Haloferax volcanii</i> , <i>Arabidopsis</i> , <i>Leishmania</i> , yeast, mouse	Submitted
6	Alternative route for 5-formyltetrahydrofolate disposal	At2g20830	COG3643	Experimental_Histidine_Degradation	Verified in 5 prokaryotes	Manuscript in prep
7	t6A biosynthesis	At2g45270, At4g22720	COG0533, YgiD	YrdC-YciO	Validated in yeast	Manuscript in prep
8	NAD-dependent nucleic acid AMP ligase	At3g12930, At1g67620	COG0799, alr4169	Iojap	In progress <i>E. coli</i>	
9	5-Formyltetrahydrofolate cycloligase paralog	At1g76730	COG0212	5-FCL-like_protein	Predicted role in thiamine recycling	
10	Hydroxyproline-galactosyl hydrolase	At5g12950, At5g12960	COG3533, SAV1144	COG3533	In progress in <i>X. campestris</i>	
11	m6A in small rRNA	At4g28830	COG2263	rRNA_modification_Archaea	Mutant analysis in <i>H. volcanii</i> in progress	
12	Choline transporter	NiaP homolog At1g13050	MFS superfamily	Choline transport and metabolism	In progress <i>R. solanacearum</i> and <i>B. xenovorans</i>	
13	Ribosome assembly/translation termination	At1g09150	COG2016	rRNA_modification_Archaea	In progress in yeast and <i>H. volcanii</i>	
14	Phytol phosphate kinase	At1g78620	COG1836, alr1612	COG1836	In progress in Synechocystis	
15	Pyridoxal phosphate enzyme in amino acid metabolism, most likely in the Glu-Pro area	At4g26860, At1g11930	COG0325, yggS	PROSC	In progress in <i>E. coli</i>	

Table 1. Status of most advanced fifteen families

COG3533 genes are found in all plants and occur sporadically in plant pathogens (bacteria and fungi) and in human pathogens. The corresponding proteins are similar to glycosyl hydrolase but the specific substrates are not known. The *Arabidopsis* COG3533 genes (At5g12950 and At5g12960) are expressed highly in pollen. Bacterial COG3533 genes are physically clustered with genes for hydroxyproline degradation, arabinose catabolism, and peptidases. We therefore propose that COG3533 proteins are glycosylhydrolases that cleave the hydroxyproline-linked galactosides found in plant cell wall proteins or in collagen. Such a hydrolase has been predicted to have a role in pollen growth and would allow plant pathogens to utilize plant cell wall components as carbon sources.

228

Functional Annotation of Putative Enzymes in *Methanosarcina acetivorans*

Ethel Apolinario,¹ Libuse Brachova,³ Yihong Chen,² Zvi Kelman,² Zhuo Li,² Basil J. Nikolau,³ Lucas Showman,³ Kevin Sowers,¹ and **John Orban*** (orban@umbi.umd.edu)

¹Center of Marine Biotechnology, University of Maryland Biotechnology Institute, Baltimore; ²Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville; ³W. M. Keck Metabolomics Research Laboratory, Iowa State University, Ames

Project Goals: The goal of the project is to develop rapid experimental approaches for accurate annotation of putative enzymatic functions. Targets of interest range from those with tentatively assigned function to hypotheticals.

Methane-producing organisms provide an efficient and cost-effective biofuel which is self-harvesting and can be distributed readily using infrastructure that is already in place. As with other genomes, however, accurate functional

annotation in methanogens lags significantly behind the large body of sequence data, representing a sizable gap in our understanding of biology in these organisms. We are using the methanogenic archaeon, *Methanosarcina acetivorans* (MA), as a model system for developing experimental tools for rapid and reliable annotation and validation of function. The target genes are putative enzymes in MA with detectable *in vivo* expression.

Our experimental approach utilizes a combination of methods for rapid function assignment. NMR spectroscopy is used to screen for putative substrates, products, or their structural analogs. Where possible, we have followed up on function assignments by checking to see if the MA gene can complement the corresponding *E. coli* knockout. We have used this approach to both validate and correct functional assignments in MA target genes, as will be illustrated with examples. Further, insights into the functional annotation of “hypotheticals” are being obtained by integrating mass spectrometry based metabolite profiles of gene knockouts with NMR-based approaches and these will also be discussed.

229

Robust Prediction of Protein Localization Via Integration of Multiple Data Types

Margaret Romine^{1*} (Margie.romine@pnl.gov), Lee Ann McCue,¹ Gretta Serres,² Tatiana Karpinets,³ Mustafa Syed,³ Sam Purvine,¹ Michael Lueze,³ Guruprasad Kora,³ Denise Schmoyer,³ Ed Uberbacher,³ Jim Fredrickson,¹ and Mary Lipton¹

¹Pacific Northwest National Laboratory, Richland, Wash.;

²Marine Biological Laboratory, Woods Hole, Mass.; and

³Oak Ridge National Laboratory, Oak Ridge, Tenn.

Project Goals: The primary research emphasis will be on associations between autotrophic and heterotrophic microorganisms with the additional objective of obtaining a predictive understanding of how interactions impart stability and resistance to stress, environmental fitness, and functional efficiency.

Genome annotations play a central role in omics-based characterization of cellular behavior and consequently it is important that they are as accurate and functionally descriptive as possible. Currently, domain content is the primary type of functional evidence used for automated functional annotation of protein-coding genes (CDSs) deduced from genome sequences. While domain content can sometimes suggest a precise function or at least provide a general functional categorization (e.g. TonB-dependent receptor), they are more often only useful for establishing that proteins having the same domain (s) are somehow functionally related. Protein localization prediction is a form of evidence that is generally under-utilized in automated annotation pipelines but has the potential to provide very useful clues regarding CDS function. As part of our efforts to manually improve the annotation of the currently available *Shewanella* genome

sequences, we developed a strategy for more accurately predicting subcellular protein localization through integration of proteome data, the output of several different localization prediction tools, ortholog analysis, and domain analyses.

At the outset of this exercise we recognized that one of the major limitations of tools that computationally predict protein localization is that an accurate gene model is needed. Since many of the commonly used tools search for the occurrence of characteristic N-terminal targeting peptides, they will be unable to detect secretion substrates that are encoded by genes with start codon inaccuracies, gene sequencing mistakes, or genomic mutations that result in displacement or loss of sequences that encode the N-terminal targeting peptide. In order to address issues with the accuracy of the gene models we first mined available MS-MS proteome data from 12 sequenced shewanellae genomes for partial tryptic peptides that could be mapped to the mature termini of proteins deduced from the original or subsequently adjusted gene models. These analyses included searches for peptides that map to N-termini produced by cellular proteolytic processing by signal peptidase I, methionine aminopeptidase, or proline aminopeptidase. We identified such peptides for 1290 proteins (~30% of the total predicted) in the extensively studied model organism *S. oneidensis* MR-1 and between 299 and 661 proteins (~10% of the total predicted) in 11 other shewanellae for which proteome data was available but was derived from only a single sample. The positions of mobile elements (insertion elements, MITES, phage, and other integrative elements) were mapped to facilitate detection of gene fragments encoding targeting peptides that were displaced by gene interruption. This analysis resulted in an increase in pseudo-gene count from 735 to 1499. Ortholog tables comprised of proteins from all 19 strains were constructed so that we could compare, within each ortholog group, the output of several localization and domain predictors with the expectation that inconsistencies in predictions would most often arise due to errors in either the gene model or predicted ortholog grouping. Ortholog groups with inconsistencies in predicted domain content, function, or location prediction or for which members were missing in a genome were then manually evaluated for inaccuracies in gene models or ortholog grouping. These analyses lead to the addition of 769 new genes and removal of 1554 genes from the gene models of these 19 shewanellae. Taking into account only changes made to the gene models of intact genes, we adjusted the start position positions in 2466 genes thereby achieving a greater consistency in predictions of localization or domain content within each ortholog group.

Since Gram negative bacteria like *Shewanella* sp. have a complex cell envelope consisting of inner and outer membranes that are separated by a periplasmic space, they employ specialized systems to mediate translocation of proteins across one or both membranes, insertion of proteins into one membrane or the other, or to tether them to one side of a membrane. The sorting signals recognized by these systems differ from one another and thus no single algorithm is optimal for predicting the subcellular locations of all proteins. This need to apply more complex

logic for predicting protein location became evident when we discovered that representatives of all six specialized protein translocation systems (T1SS-T6SS) known to occur in gram negative bacteria were present in at least one sequenced *Shewanella*. We developed a series of rules to identify substrates of specialized secretion systems as either bioinformatics tools were not available to identify their substrates or their predictions were not particularly robust. For example, combining domain information and proteomics data for the NiFe hydrogenase orthologs allowed us to identify these proteins as substrates of the twin arginine translocation (TAT) system. Substrates of the TAT secretion system are expected to include proteins that possess metallic redox active centers and therefore all proteins having such domains, including the NiFe hydrogenases, were carefully evaluated for the presence of N-terminal targeting peptide recognized by this secretion pathway. In *Shewanella*, the NiFe hydrogenases have an unusually long targeting peptide that was validated by proteome analysis (68 amino acids) but routinely missed by both TatP and Tatfind algorithms. The identification of outer membrane proteins was also not very accurate using a single computational tool. The Bomp beta barrel prediction tool, for example, inconsistently detected outer membrane proteins within ortholog groups even after gene model adjustment. Therefore, we supplemented these analyses by searching for a C-terminal outer membrane targeting consensus motif. Since it is known that some outer membrane proteins do not encode this domain at the C-terminus (e.g. OmpA family proteins, secretins) we also used location-informative domains to assist in identification of outer membrane proteins. Other systems, such as the type II secretion system (T2SS) that translocate periplasmic proteins across the outer membrane have no universally recognized targeting motif, but are instead believed to be recognize targeting signals that are species-specific. In *Shewanella* it is known that at least three lipoproteins are substrates of this system. A comparative analysis of these lipoproteins with other proteins deduced from the genome sequence revealed a putative targeting motif similar to those described for extracellular proteins in other bacteria, providing us a means to expand the number of predicted T2SS substrates in this Genus.

We estimate that approximately 40% of the predicted proteome for each strain of *Shewanella* is translocated out of the cytoplasm. These extracytoplasmic proteins play a central role in modulating the interactions of members of this genus with their external environments and in generating the energy and accessing the nutrient necessary to support growth and metabolism. As part of PNNL's new Foundational Science Focus area on Biological Systems Interactions we intend to employ this general strategy to identify secreted proteins in new model organisms and microbial communities to facilitate future studies directed at developing a broader understanding of microbial interactions.

submitted post-press

Genome-Scale Phylogenetic Function Annotation of Large and Diverse Protein Families

Barbara E. Engelhardt,^{1,4} Michael I. Jordan,² Susanna Repo,³ and Steven E. Brenner³ (brenner@compbio.berkeley.edu)

¹EECS Dept., ²Dept. of Statistics, and ³Plant and Microbial Biology Dept., University of California, Berkeley; and ⁴Computer Science Dept., University of Chicago, Ill.

Project Goals: The goal of the project is to enhance the algorithms and statistical models of SIFTER, our protein function prediction method. We will also extend SIFTER's applicability by including additional sources of function evidence. With these improvements, SIFTER will become applicable to a broader range of protein families, including large, and functionally diverse families, and to work on genome-wide scale. In addition, we will adapt SIFTER on metagenomic data.

It is now easier to discover thousands of protein sequences in a new microbial genome than it is to biochemically characterize the specific activity of a single protein of unknown function. Through metagenomic analysis, next-generation sequencing heralds unprecedented opportunities for understanding the environmental microbiota. A single experiment alone, the Global Ocean Sampling study, more than doubled the number of known protein sequence entries. However, despite this large body of new sequence information, functional annotation remains a major challenge. Molecular functions of proteins in the novel genomes continue to be discovered, in large part by homology to those experimentally characterized in model organisms.

Typically, protein function annotation involves finding homologs of a protein sequence, followed by database queries and computational techniques to predict function from the annotated homologs. These methods rely on the principle that proteins from a common ancestor may share a similar function. However, most protein families have sets of proteins with different functions and therefore traditional bioinformatics approaches are unable to reliably assign the appropriate function to unannotated proteins. Currently, protein function databases have a large proportion of erroneously annotated proteins, where the incorrect annotations were either derived using an imprecise computational technique or inferred using another incorrect annotation¹⁻⁴.

We have proposed integrating available functional data using the evolutionary relationships of a protein family, and we implemented this method in the program SIFTER (Statistical Inference of Function Through Evolutionary Relationships). The SIFTER methodology uses a statistical graphical model to compute the probabilities of molecular functions for unannotated proteins. Currently, SIFTER takes as input a reconciled phylogeny and a set of annota-

tions for some of the proteins in the protein family. We incorporate known information about function by computing the probability of each of the candidate functions for the proteins in the tree with available functional evidence from the GOA database. The candidate molecular functions are represented as a boolean vector, where initially the probability associated with each candidate function is a function of the set of annotations for that protein and their corresponding evidence types (e.g., experimental, electronic). From this reconciled phylogeny with sparse observations, SIFTER computes the posterior probability of each molecular function for all proteins in the family using a simple statistical model of protein function evolution.

We tested the performance of SIFTER on three different protein families: AMP/adenosine deaminases, sulfotransferases and Nudix hydrolases with cross-validation experiments. SIFTER's performance was compared with three other function prediction algorithms: BLAST, GOtcha and Orthostrapper, and SIFTER was shown to outperform the other methods. In addition, on a genome-wide scale we used SIFTER to annotate the experimentally characterized proteins from *Schizosaccharomyces pombe*, based on the annotations from 26 other fungal genomes. The newest version of SIFTER implements a faster method for calculating the posterior probabilities, and this improvement, together with a more general evolutionary model make SIFTER applicable on large and functionally diverse protein families and on genome-scale function annotation.

The development of SIFTER is an ongoing project and a new version of the program is now available (manuscript under review). We are currently testing SIFTER for metagenomic sequences with the acid mine drainage datasets from Jill Banfield. In the near future, we are planning to expand our analysis to other metagenomic datasets, such as the termite gut datasets from the JGI. We also use SIFTER to annotate enzymes from chlorite dismutase and perchlorate reductase families, in order to identify species that are capable of perchlorate reduction. Furthermore, we are validating SIFTER predictions experimentally using the large and extremely diverse Nudix family of hydrolases as a test bed.

This project has been funded with DOE grant number BER KP 110201.

References

1. Brenner SE 1999 *Trends Genet.* **15** 132-3
2. Galperin MY and Koonin EV 1998 *In Silico Biol.* **1** 55-67
3. Jones CE, Brown AL and Baumann U 2007 *BMC Bioinformatics.* **8** 170
4. Schnoes AM, Brown SD, Dodevski I and Babbit PC 2009 *PLoS Comput Biol.* **5** e1000605