ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/231738882

Computational Prediction and Experimental Validation of Signal Peptide Cleavages in the Extracellular Proteome of a Natural Microbial Community

ARTICLE *in* JOURNAL OF PROTEOME RESEARCH · MARCH 2010

Impact Factor: 4.25 · DOI: 10.1021/pr900877a

CITATIONS
17

READS
18

8 AUTHORS, INCLUDING:

Nathan C Verberkmoes
Berg Pharma
128 PUBLICATIONS **4,075** CITATIONS

SEE PROFILE

Steven W Singer
Lawrence Berkeley National Laboratory
72 PUBLICATIONS **1,300** CITATIONS

SEE PROFILE

Michael P Thelen
Lawrence Livermore National Laboratory
75 PUBLICATIONS **2,730** CITATIONS

SEE PROFILE

Robert L Hettich
Oak Ridge National Laboratory
252 PUBLICATIONS **6,902** CITATIONS

SEE PROFILE

Available from: Michael P Thelen
Retrieved on: 25 January 2016

research articles Journal of proteome research

Computational Prediction and Experimental Validation of Signal Peptide Cleavages in the Extracellular Proteome of a Natural Microbial Community

Brian K. Erickson,[†,‡] Ryan S. Mueller,[⊥] Nathan C. VerBerkmoes,[‡] Manesh Shah,[§] Steven W. Singer,[||] Michael P. Thelen,[||] Jillian F. Banfield,[⊥] and Robert L. Hettich*[,‡]

Graduate School of Genome Science & Technology, University of Tennessee, Knoxville, Tennessee 37830, Chemical Sciences Division and Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, Lawrence Livermore National Laboratory, Livermore, California 94550, and University of California at Berkeley, Berkeley, California 94720

Received September 29, 2009

An integrated computational/experimental approach was used to predict and identify signal peptide cleavages among microbial proteins of environmental biofilm communities growing in acid mine drainage (AMD). SignalP-3.0 was employed to computationally query the AMD protein database of >16,000 proteins, which resulted in 1,480 predicted signal peptide cleaved proteins. LC−MS/MS analyses of extracellular (secretome) microbial preparations from different locations and developmental states empirically confirmed 531 of these signal peptide cleaved proteins. The majority of signal-cleavage proteins (58.4%) are annotated to have unknown functions; however, Pfam domain analysis revealed that many may be involved in extracellular functions expected within the AMD system. Examination of the abundances of signal-cleaved proteins across 28 proteomes from biofilms collected over a 4-year period demonstrated a strong correlation with the developmental state of the biofilm. For example, class I cytochromes are abundant in early growth states, whereas cytochrome oxidases from the same organism increase in abundance later in development. These results likely reflect shifts in metabolism that occur as biofilms thicken and communities diversify. In total, these results provide experimental confirmation of proteins that are designed to function in the extreme acidic extracellular environment and will serve as targets for future biochemical analysis.

Keywords: Mass spectrometry • multidimensional liquid chromatography • shotgun proteomics • signal peptides • microbial communities

Introduction

In order to understand how microorganisms cooperate and compete in natural environments, it is vital to be able to identify and fully characterize the expressed protein complement (i.e., whole community proteome, or metaproteome) for uncultivated as well as cultivated organisms.[1,2] Determination of the cellular location(s) of proteins provides important contextual information, which can support proposed functions. Of particular interest are proteins that mediate interactions between a microorganism and its environment and operate under external conditions that may differ substantially from conditions in the cytoplasm. These secreted proteins are critical for nutrient transport, as well as organismal communication and survival (i.e., defense).

A primary but not exclusive method of protein transport to the extracellular region, periplasmic space, or outer membrane of Gram-negative bacteria is through signal peptide mediated transport.[3] In this highly conserved process, trafficking of the protein is dependent on the presence of a specific sequence of amino acids, typically located within the first 50 amino acids of the N-terminus.[4,5] Targeting generally occurs through two pathways: one involves the signal recognition particle (SRP) and occurs cotranslationally, and the other involves SecB and occurs post-translationally. Following targeting, protein transport to the cytosolic membrane occurs through a complex of proteins known as the translocase, which includes membrane proteins as well as an ATPase.[6] The result of these activities is the directed transport of a protein and cleavage of the signal peptide. Additional models of protein secretion are utilized within Gram-negative bacteria but were not specifically probed within this study.

The ability to computationally predict signal peptides has advanced significantly in recent years.[7,8] Current prediction algorithms utilize machine learning techniques, such as neural

[*] To whom correspondence should be addressed. Chemical Sciences Division, Oak Ridge National Laboratory, P.O. Box 2008 MS-6131, Oak Ridge, TN 37831-6131. Ph: (865) 574-4968. Fax: (865) 576-8559. Email: hettichrl@ornl.gov.

[†] University of Tennessee.

[‡] Chemical Sciences Division, Oak Ridge National Laboratory.

[§] Life Sciences Division, Oak Ridge National Laboratory.

[||] Lawrence Livermore National Laboratory.

[⊥] University of California at Berkeley.

10.1021/pr900877a

© 2010 American Chemical Society

Published on Web 03/10/2010

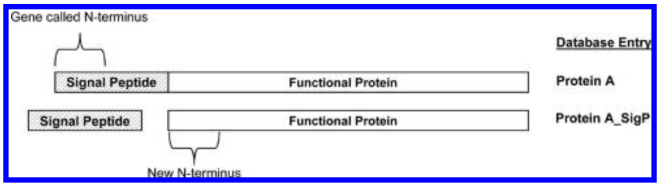


Figure 1. Representation of signal peptide and cleavage The original peptide of a protein was termed the "gene called N-terminus". Following removal of a signal peptide, a new N-terminus will be present on a protein. This is termed the "new N-terminus". The protein database contains both the preprocessed form of the protein, "Protein A", and the signal peptide cleaved sequence, "Protein A_SigP".

networks and hidden Markov models (HMM) to increase accuracy and precision.[9−11] These computational techniques identify patterns of amino acid composition in the N-terminal region of a protein in order to ascertain if a signal peptide is present. The pattern recognition has been optimized through the use of training sets and is specific for eukaryotes and Gram-negative and Gram-positive bacteria. The training sets are composed of hundreds to thousands of experimentally verified signal peptide sequences. SignalP-3.0 is one widely used and accepted algorithm that effectively identifies the presence of a signal peptide along with the probable cleavage site. The current version of SignalP (3.0) has been improved over previous iterations by utilizing expanded training sets containing additional experimentally verified signal peptides, as well as including HMM resulting in increased prediction accuracy. For Gram-negative bacteria, prediction of a signal peptide as well as identification of the cleavage site is proposed to be >90% accurate.[12]

Experimental data provides important confirmations of signal peptide predictions. N-Terminal sequencing techniques, such as Edman sequencing, can be used to verify the sequences of mature protein forms, but this is not an effective method for profiling the thousands of proteins present in microbial community proteomes. An alternative approach is to verify signal-cleaved peptides using shotgun multidimensional liquid chromatography tandem mass spectrometry (2D-LC−MS/MS). Since peptide assignments using shotgun proteomics depend on the presence of the exact predicted peptide sequences in databases, signal-cleaved peptides and noncleaved peptides can be readily distinguished (Figure 1).[13,14] Mass spectrometry is appropriate for signal peptide analysis in microbial community samples due to its unrivaled throughput, as well as its high dynamic range and mass accuracy.[15] Mass spectrometry also provides relative quantification of peptides and proteins, allowing for detection of trends in abundance patterns of exported proteins across samples.[16]

In this study, we evaluated the approach of integrating computational prediction of signal peptide-containing proteins with high-throughput mass spectrometry to validate signal peptide predictions for a diverse mixture of proteins from a natural microbial community. We focused on microbial biofilms with limited species richness from an acid mine drainage (AMD) system.[1,2,17] The biofilms grow in hot (40 °C), pH ∼1.0 solutions that contain near molar concentrations of metals (in particular Fe). Proteogenomic analyses, which combine proteomic measurements and metagenomic data, have been previously applied to these biofilms to catalogue and evaluate abundance patterns for thousands of proteins from the most abundant bacterial and archaeal populations.[18−22] As of yet, a specific identification and characterization of the secreted proteins present in the extremophilic AMD system has not been completed. The analysis presented in this study has broad implications for characterizing extremophilic microbial communities. High confidence identification of the secretomes will provide vital clues into microbial community interaction, function, and survival at the environmental and cellular interface. The combination of protein enrichment in the secretome and the presence of signal-cleaved peptides provide strong evidence for protein localization and clues to protein function. Characterization of the changes in the abundances of signal-cleaved proteins across microbial communities from biofilms growing in different geochemical environments and of different growth states permits a greater understanding of the roles of these proteins *in situ*. A subset of these secreted proteins may be critical for organismal survival in the highly acidic environment and should provide unprecedented insight into the global acid mine drainage phenomenon. Finally, the methodologies presented within can be readily applied to a variety of microbial systems for specific prediction/characterization of their secreted proteins.

Experimental Procedures

Signal Peptide Prediction. The experimental approach for this study consisted of parallel computational prediction and mass spectrometric identification of signal peptide cleaved proteins (Figure 2). SignalP-3.0[23] was downloaded (http://www.cbs.dtu.dk/cgi-bin/nph-sw_request?signalp) and locally installed. Each of the 16,171 proteins present in the AMD database (Biofilm_5wayCG_UBA_08052007.fasta) was individually submitted to SignalP-3.0 for analysis using a Perl script that iteratively selects and copies each FASTA formatted protein sequence from the sequence database, along with accompanying header information, to a separate temporary text file. This text file was submitted to SignalP-3.0, which was executed with the following parameters: organism set to Gram-negative bacteria, output short format, quiet analysis, and protein sequence truncation after the first 50 amino acids. Results of SignalP-3.0 were exported to a temporary file, and identification of signal peptides was accomplished by parsing the results of the hidden Markov model analysis conducted by SignalP-3.0. Following the prediction of a signal peptide, the position of the cleavage site was noted in order to generate the processed protein sequences. Sequences of proteins predicted to have a signal peptide were truncated at the predicted cleavage site and their protein names were appended with "SigP" in the new protein sequence database. This database was labeled "Biofilm_5wayCG_UBA_08052007_SigP_Removed.fasta" and contains both the original gene-called protein sequence and, if predicted to be present, a signal peptide cleaved protein sequence. The complete, SignalP-3.0 derived database can be found at "http://compbio.ornl.gov/biofilm_amd_extracellular_proteome".

The lack of archaeal training data sets limits the effectiveness of SignalP-3.0 in predicting archaeal signal peptides. For this reason, the predictions and identifications of signal peptide cleaved proteins in this study were generally found for the abundant Gram-negative bacterial populations in AMD biofilms.

AMD Biofilm Sample Preparation. Biofilm samples collected from the AB-End, AB-Front, AB-Muck (Friable), AB-Muck (GSII), and UBA locations[21] of the mine each contained approximately 1 × 10^{10} cells. AB-End was an earlier growth state biofilm than AB-Front and AB-Muck, which were designated as Growth Stage II (GSII). AB-Muck (Friable) exhibited a unique

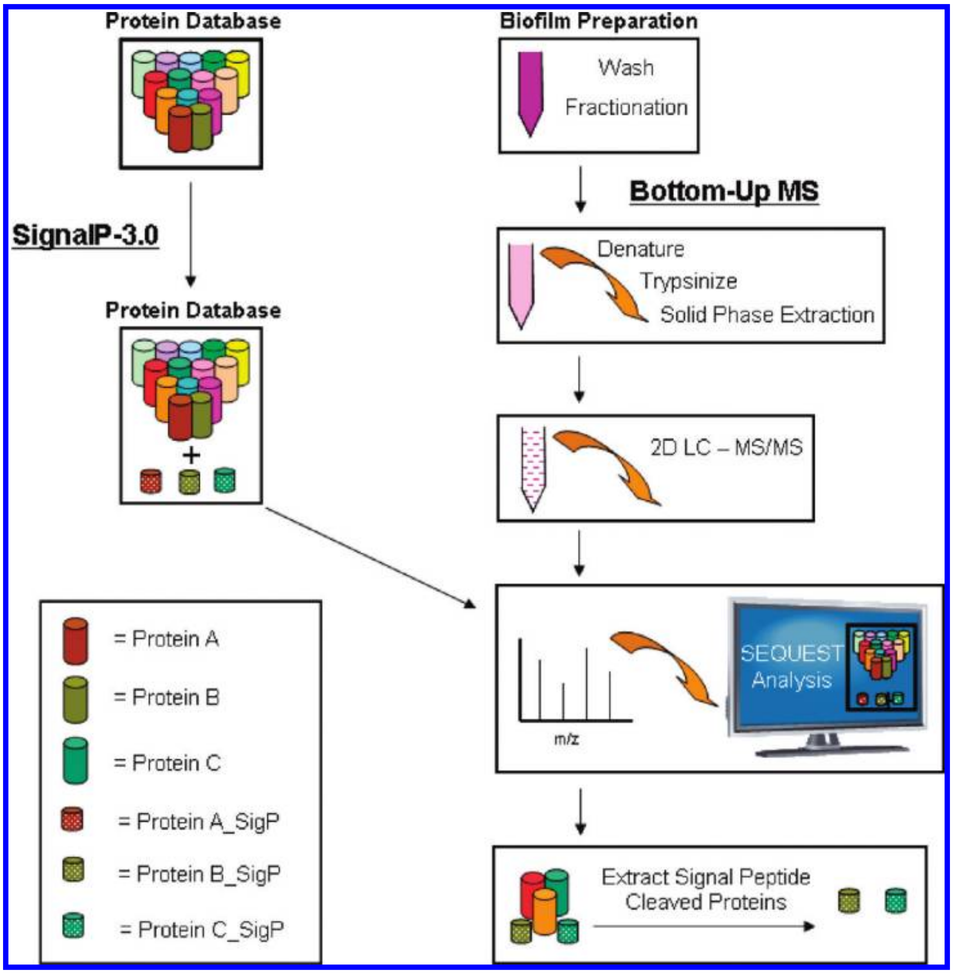


Figure 2. Methodology. The crude biofilm was fractionated and digested into peptides for LC−MS/MS. The acid mine drainage protein database was subjected to signal peptide cleavage prediction with SignalP-3.0. The proteins with predicted signal peptide cleavage were appended to the database with the signal peptide sequence removed and noted with a "SigP". The preprocessed protein sequence was retained in the database. Following MS/MS analysis the spectra were searched with SEQUEST utilizing the signal peptide appended database and parsed for signal peptide cleavage identifications.

shift in the dominant microbial species, and UBA exists only in the A-drift region of the mine. The five frozen samples were thawed and processed, as follows, at 4 °C. Cells were suspended in 3 volumes of H_2SO_4 (pH 1.1), washed by rotation for 30 min, and recovered by centrifugation at 12,000*g* for 20 min. This wash was repeated once by resuspending the cell pellet in the same volume of sulfuric acid solution, and the two reddish-yellow supernatants were combined to form the extracellular fraction. The whole cell fractions were further processed to membrane and soluble fractions and used in other studies.[19−22] Since the extracellular fraction was collected after treatment of the biofilm by cold osmotic shock, it is likely enriched in both periplasmic and secreted proteins. Proteins within the extracellular fraction were precipitated with ice-cold 10% trichloroacetic acid, and the pellet was rinsed with cold methanol and air-dried. Each of the extracellular pellets (~1−2 mg protein material) were resuspended, denatured, and reduced in 2 mL of 6 M guanidine-HCl, 10 mM DTT, at 60 °C for 1 h. Samples were diluted 6-fold in 50 mM Tris-HCl/10 mM $CaCl_2$ (pH 7.8), sequencing-grade trypsin (Promega, Madison, WI) was added at ~1:100 (w/w), and digestions were performed with gentle rocking at 37 °C for 18 h. This was followed by a second addition of trypsin at 1:100 and an additional 5-h incubation. The samples were then treated with 20 mM DTT for 1 h at 37 °C as a final reduction step and immediately desalted with Sep-Pak Plus C18 (Waters, Milford, MA). All samples were concentrated and solvent exchanged into 0.1% formic acid in water by centrifugal evaporation to ~1 mg/mL starting material, filtered, aliquoted, and frozen at −80 °C until LC−MS/MS analysis.

Mass Spectrometric Characterization. Mass spectrometric measurements of five biofilm extracellular fractions were done in triplicate on an LTQ mass spectrometer (Thermo Fisher Scientific, San Jose, CA). The mass spectrometer was coupled online with an Ultimate HPLC (LC Packings, a division of Dionex, San Francisco, CA). The system utilized a 2D nano-LC tandem mass spectrometry (2D-LC−MS/MS) setup. The flow rate from the pump was maintained at ~100 μL/min, which was then split precolumn to provide an approximate flow of ~200−300 nL/min at the nanospray tip. The split-phase columns were prepared in-house and consisted of strong cation-exchange material (Luna SCX 5 μ 100 Å Phenomenex, Torrance, CA) and C18 reverse phase (RP) material (Aqua C18 5 μ 125 Å Phenomenex). For all samples, ~200−500 μg of protein material was loaded off-line onto the back of the multiphase column. The loaded RP-SCX column was then positioned on the instrument behind a ~15 cm C18 RP column (Aqua C18 5 μ 125 Å Phenomenex) also packed via a pressure cell into a Pico Frit tip (100 μm with 15 μm tip New Objective, Woburn, MA). All samples were analyzed via a 24-h 12-step 2D analysis. During the entire chromatographic process, the LTQ mass spectrometer was operated in a data-dependent MS/MS mode detailed below. The chromatographic methods and HPLC columns were virtually identical for all analyses. The LC-MS system was fully automated and under direct control of the XCalibur software system (Thermo Fisher Scientific). The LTQ mass spectrometer was operated with the following parameters: nanospray voltage (2.4 kV), heated capillary temp 200 °C, full scan m/z range (400–1700). The LTQ data-dependent MS/MS mode was set up with the following parameters: 5 MS/MS spectra for every full scan, 2 microscans averaged for both full scans and MS/MS scans, 3 m/z isolation widths for MS/MS isolations, and 35% collision energy for collision-induced dissociation. To prevent repetitive analysis of the same abundant peptides, dynamic exclusion was enabled with a repeat count of 1 and an exclusion duration of 3 min on the LTQ.

Proteome Bioinformatics. The "Biofilm_5wayCG_UBA_08052007_SigP_Removed" database contains annotated proteins from the abundant microbial members of AMD biofilms.[20] The protein database also includes common contaminants (trypsin, keratin, etc.). Protein assignment of the MS/MS spectra was accomplished with the SEQUEST algorithm[24] and was run using the following parameters: enzyme type, trypsin; Parent Mass Tolerance, 3.0 Da.; Fragment Ion Tolerance, 0.5 Da.; up to 4 missed cleavages allowed, and fully tryptic peptides only. Resulting output files were sorted and filtered using DTASelect with the following parameters: tryptic peptides only, deltaCN value of at least 0.08, and Xcorr values of at least 1.8 (+1), 2.5 (+2), 3.5 (+3) with a two peptide minimum. Cross-comparison among DTASelect outputs was accomplished with Contrast[25] and an in-house script that provides similar functions. Rapid filtering of the signal peptide cleaved proteins identified in the DTASelect output was accomplished using a Perl script that extracted protein identifications containing the "SigP" designation to an additional table. Accompanying information regarding protein sequence coverage, number of peptide identifications, and spectral counts were also recorded. In order to support the identification of a signal peptide cleavage, the DTASelect output was parsed for the predicted, preprocessed signal peptide. Identifications of a preprocessed signal peptide were noted along with accompanying spectral counts. A false positive rate of <1% was calculated on the basis of forward-reverse database searching according to Elias et al.[26] All databases, peptide and protein results, MS/MS spectra, and supplementary tables for all database searches are archived and made available as open access via http://compbio.ornl.gov/biofilm_amd_extracellular_proteome. All MS ".raw" files or other extracted formats are available upon request.

Highly expressed signal-cleaved proteins with confirming new N-terminus spectra were submitted in batch form to Pfam for protein family and domain analysis.[27,28] The parameters for the search included a merged global and local strategy and an E-value cutoff of 1.0 The resulting Pfam hits were further filtered with an E-value cutoff $<1 \times 10^{-3}$, exceeding the stringency outlined in Altschul et al.[29]

N-Terminal Sequencing. Complementary experimental verification of signal peptide cleavage was accomplished on selected secretome proteins by Edman degradation. The extracellular fraction extracted from 50 mL of biofilm from the C-drift location collected in November, 2005 was fractionated by column chromatography on a SP-Sepharose FF column, as previously described.[30] After elution of Cyt_{579}, a 0–2 M NaCl gradient was applied at pH 5.0 (30 mL, 3 mL fractions). Greater than 95% of the proteins recovered from the NaCl gradient were present in the 1.4–2.0 M NaCl fractions. These fractions were precipitated with 10% tricholoroacetic acid in an ice bath and redissolved in 10 μL of SDS-PAGE sample buffer. The final protein weight dissolved in the sample buffer was 20–40 μg. The samples were visualized by SDS-PAGE (15% polyacrylamide precast gel, Bio-Rad), transferred to a polyvinylidene fluoride (PVDF) membrane, and the bands excised for sequencing. N-Terminal sequencing of the visualized proteins was performed as previously described.[19]

Hierarchical Cluster Analysis of Signal Peptide Cleavage in 28 AMD Proteomes. The abundance patterns of computationally predicted and experimentally verified signal peptide cleaved proteins were examined across a distinct set of 28 biofilm samples collected over a period of 4 years from 6 different locations within the Richmond Mine, from a different study.[21] The abundances of individual proteins were calculated using normalized spectral abundance factors (NSAF), which are based on the spectral counts of peptides for a given protein.[31,32] Resulting NSAF values were ASIN-transformed and used to cluster proteins and samples using Cluster version 3.0.[33] Clustering of mean-centered and scaled NSAF values was performed using an uncentered Pearson correlation metric, and groups were defined using average linkage clustering. Heat maps were visualized with TreeView.[34] To determine whether correlations exist between protein abundances and developmental state of the biofilm, samples were labeled as either a high- or low-developmental stage on the basis of the numbers of archaea detected within each community, as previously determined (Mueller et al., submitted for publication). Low developmental stage samples are highlighted in green, and high developmental stage biofilms are highlighted in blue for each heatmap presented. Detection of differentially expressed signal peptide-containing proteins between developmental stages was achieved using the significance analysis of microarrays technique.[35] Two class unpaired Wilcoxan tests were performed using 500 permutations. Significant genes were assessed at a <10% false discovery rate (FDR).

Results

Shotgun proteomics via 2D-LC–MS/MS provides the critical cataloging of proteolytic peptides, thereby enabling the discovery and validation of signal peptide cleavage events. The complete AMD protein database was interrogated for signal peptide prediction along with concurrent LC–MS/MS measurements of community biofilm samples in order to ascertain (1) whether the protein was expressed and detected, and (2) if so, did it reveal a new N-terminal sequence that would be representative of the processed, mature form of the protein?

SignalP-3.0 Prediction Results. The computational prediction of signal peptides resulted in 1,480 signal-cleaved proteins out of 16,171 proteins (9%) from the AMD database (Table 1). Approximately 18% of the proteins from Gram-negative organisms were predicted to contain a signal peptide and more than half of the signal peptide predictions were from the dominant organisms, two strains of *Leptospirillum* group II (395 from the CG strain and 397 from the UBA strain). *Leptospirillum* group III contained 304 predicted signal peptide containing proteins (11.1% of its total annotated proteome), representing >20% of the signal peptide database (Table 2).

Table 1. Summary of Computational Prediction and MS Identification of Signal Peptide Cleaved Proteins from Five Distinct Extracellular AMD Samples

	no. of IDs	% of total DB
SignalP-3.0 prediction	1,480	9
measured protein IDs (all)	3,388	21
measured protein IDs (SigP)	531	3

Table 2. Distribution of Predicted Signal Peptide Cleaved Proteins for the Dominant Microbes in the AMD Microbial Community

	no. of proteins in SigP database	% of SigP database
Leptospirillum II	792	53.5
Leptospirillum III	304	20.5
G-plasma	66	4.5
Ferroplasma 1	105	7.1
Ferroplasma 2	136	9.2
unassigned	77	5.2
total	1,480	100.0

Experimental MS Results. Extracellular fractions from five distinct biofilms from different locations in the Richmond Mine were analyzed in triplicate by 2D-LC−MS/MS. Detailed information on each sample is provided in the six supplemental tables. Overall, the MS analysis resulted in the identification of 3,388 total proteins; 531 proteins with predicted signal peptides were identified in at least one of the five sample sets (Figure 3). After removal of orthologous proteins, 377 nonredundant signal peptide cleaved proteins were identified. From these results, 115 nonredundant proteins were measured and identified as signal peptide cleaved proteins in *all samples and technical replicates*, and 46 of the 531 proteins were determined to have signal peptide cleavages with high confidence on the basis of the *presence of a least one spectra corresponding to the new, processed N-terminus and MS identification in all samples and replicates*. Also, 125 total proteins were identified in at least one sample with spectra matching to the new N-terminus generated by signal peptide cleavage. Although the identification of the new, signal peptide cleaved N-terminus provides strong support for the classification of that protein as signal peptide cleaved, the absence of a new N-terminal peptide identification *does not necessarily* indicate that the protein does not contain a signal peptide. For example, there are numerous proteins predicted to contain signal peptides for which no N-terminal peptides were experimentally identified with the current methods employed. The identification of a new N-terminus is dependent on the predicted signal peptide sequence and resulting MS peptide identification and would not be confused with simple tryptic cleavage to result in a new N-terminal identification. Computationally, the new N-terminus is designated in such a way that it is distinguishable from N-terminal tryptic peptides. Therefore, any new N-terminal identification is the result of a specific signal peptide computational prediction and corresponding experimental verification. Among the secretome results, several proteins not predicted to contain a signal peptide by SignalP-3.0 were still identified in the MS analysis. Examples include highly abundant proteins such as GroEL, numerous ribosomal subunits, and various transcription factors. Cell lysis or incomplete fractionation could account for these abundant proteins, which are frequently identified in proteomic analyses of the AMD microbial community.

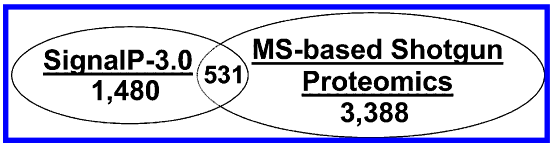


Figure 3. Venn diagram of predicted and measured signal peptides. Computational analysis of the acid mine drainage protein database with SignalP-3.0 resulted in the prediction of 1,480 proteins with a signal peptide. Following MS/MS analysis a total of 3,382 proteins were confidently identified. Among these, 531 proteins were predicted to contain a signal peptide cleavage and were ultimately identified through mass spectrometry.

Table 3. Results of Edman N-Terminal Sequencing of Select Secretome Proteins from the AMD Microbial Community[a]

band number	gene product	observed N-terminal sequence[b]
1	UBA_LeptoII_8241_114	ASTTKGWVFR*
2	UBA_LeptoII_8524_128	**SDVVGVVDVL***
3	UBA_LeptoII_8692_12	AGTPSEKLIQ[c]
4	UBA_LeptoII_8241_693	**ASNITI***
5	UBA_LeptoII_8049_366	**(A)DAYKTGH***
6	UBA_LeptoII_8524_180	**DQAAPAAPA***
7	UBA_LeptoII_8524_180	AAKKKPAKKA
8	UBA_LeptoII_8524_180	GKAKPSMFV MGKAKKPSMF
9	UBA_LeptoII_8524_180	KKAAKKPMKK
10	UBA_LeptoII_8241_349	**(EA)HMDHHRMMMR***

[a] Bolded sequences correspond with SignalP-3.0 prediction. [b] The asterisk (*) indicates mass spectrometric confirmation. [c] UBA_8692_12 and its CG homologue have predicted sequence AGDPSEKLIQ.

Clearly our experimental approach will be most successful for identifying soluble secreted proteins. We recognize that predicted signal-peptide proteins designed for membrane insertion would likely be under-represented in our data sets. We used the transmembrane predictor tool TMHMM[36] to interrogate the entire set of SignalP predicted proteins (1,480) and found that about 30% of them contain one or more transmembrane helix predictions. As expected, our experimentally identified signal peptide cleaved proteins did not contain any transmembrane predictions. For the signal-peptide proteins *not identified* in this current study, we propose the following possible scenarios: (1) they were not expressed, (2) they are expressed at levels too low to detect, or (3) they are membrane proteins and thus escape detection by this method. While algorithms such as TMHMM can predict the last category, these cannot definitively define why other proteins went undetected.

Signal Peptide Prediction Disparities. In five cases, peptides predicted by SignalP-3.0 to be cleaved from the mature protein were identified in the uncleaved form by 2D-LC−MS/MS. These five proteins, represented by only 15 spectra, are derived from *Leptospirillum* Group II (4) and Group III (1). The four from *Leptospirillum* Group II are conserved proteins of unknown function, whereas the *Leptospirillum* Group III protein has no known function. Of special note, we identified alternative forms of four of the *Leptospirillum* Group II proteins that had the predicted cleaved N-terminus. This could indicate that proteins are incompletely processed, that there were lysed cells with unprocessed protein in the extracellular fraction, or the identifications could be wrong (due to false positive spectral assignments). However, it is important to note that these five

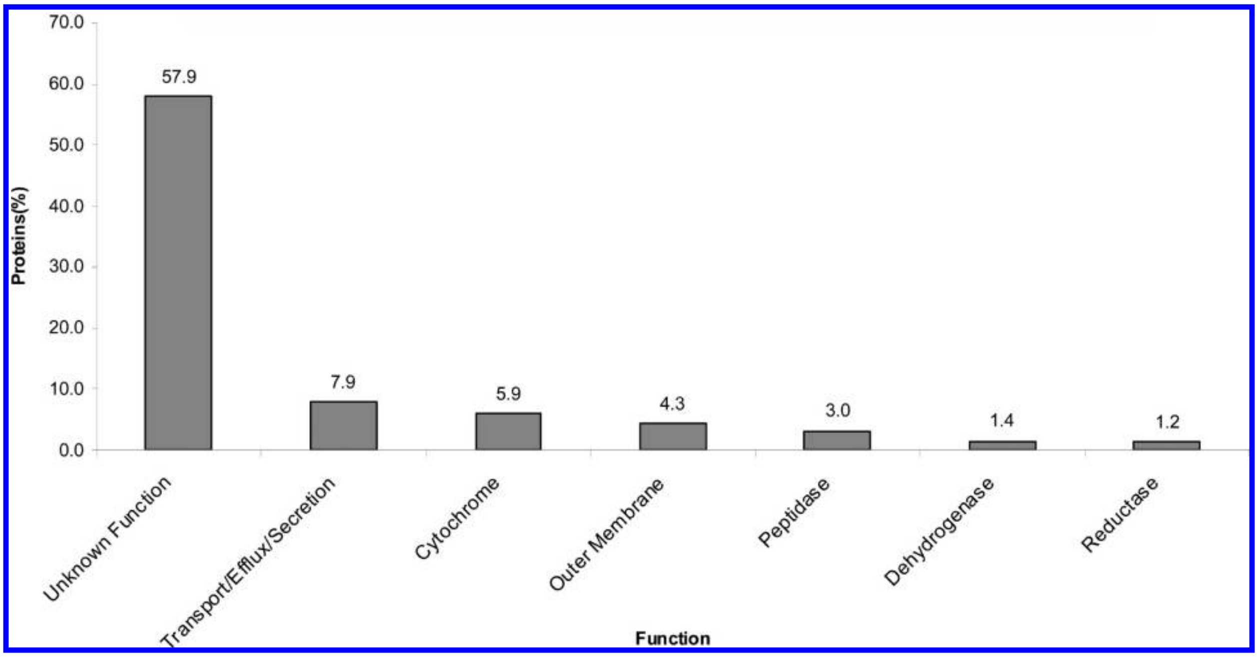


Figure 4. Functional distribution of identified signal peptide cleaved proteins. Functional analysis of the 377 nonredundant proteins predicted to contain a signal peptide and identified through mass spectrometry. Proteins annotated as hypothetical or with an unknown function are widely present and comprise over 57% of the identified proteins. Expected extracellular proteins such as cytochromes, reductases, and peptidases were also identified.

cases represent a minority of the signal cleaved proteins detected and verified.

Validation of N-Terminal Protein Sequences. Edman degradation sequencing was used to confirm some of the N-termini of proteins in the extracellular fraction predicted by SignalP and identified by MS. The N-termini of seven *Leptospirillum* group II gene products were determined, and all of these correlated to the predicted N-termini determined through this study, except for the protein encoded by *Leptospirillum* group II UBA scaffold 8692 gene 12. This protein has an amino acid variation (measured AG*T*PSEKLIQ, predicted AG*D*PSEKLIQ), accounting for the discrepancy (Table 3). Two of the most abundant secreted proteins are encoded by *Leptospirillum* group II UBA scaffold 8524 gene 128 and *Leptospirillum* group II UBA scaffold 8524 gene 180. The predicted N-terminus was observed for *Leptospirillum* II UBA scaffold 8524 gene 128, which is annotated to be a putative outer membrane protein (OmpH); however, a second form of the protein with the same N-terminus was present at higher molecular weight. This second form may represent a post-translational modification of the protein. For the protein product of *Leptospirillum* II UBA scaffold 8524 gene 180, five additional N-termini were identified in addition to the predicted N-terminus, which suggests that this protein is highly susceptible to protease cleavage.

Discussion

This integrated computational/experimental study revealed a large complement of proteins that are actively transported beyond the cytosol in the dominant bacterial AMD community members. Given that the periplasms and outer membranes of cells are exposed to the very acidic, metal-rich environment, proteins localized there, including those involved in Fe^{2+} oxidation and electron transport,[37] must be adapted to these environmental challenges.

Figure 4 summarizes the functional grouping of signal peptide cleaved proteins. Several proteins identified as transported across the cytoplasmic membrane were annotated as efflux/protein transporters (8%), cytochromes (~6%), dehydrogenases, proteases, and reductases, as described in Goltsman et al.[22] This finding correlates well with an experimental investigation of the secretome of *Bacillus subtilis,* a Gram-positive bacteria, where many proteases, dehydrogenases, and metal binding proteins were also highly abundant.[38] Over 58% of the identified signal peptide cleaved proteins are currently annotated as having an unknown function. Two novel cytochromes, cytochrome 579 (Cyt_{579}) and cytochrome 572 (Cyt_{572}) are highly abundant within the AMD biofilms. In particular, Cyt_{579}, thought to function as an electron transfer protein,[30] was identified in all 15 MS experiments (270 spectra corresponding to the predicted new N-terminus of Cyt_{579} were identified).

The proteins with the highest confidence signal peptide cleavage are those that contain spectra matching to peptides representing the new N-terminus. Table 4 lists 46 proteins for which signal peptide cleaved peptides were identified in all 15 extracellular samples (triplicate analysis of 5 AMD biofilms). These highest confidence cases include functionally distinct proteins from *Leptospirillum* group II (CG and UBA strains) and *Leptospirillum* group III. In addition to proteins of unknown function, cytochromes, isomerases, and outer membrane proteins were also identified. Several proteins of unknown function exhibited high spectral counts, suggesting that they are metabolically critical. For example, we identified 531 spectra for the cleaved N-terminus of the protein encoded by *Leptospirillum* group II UBA scaffold 8049 gene 83 and its ortholog, CG scaffold 11390 gene 17. Another example is the *Leptospirillum* group II protein encoded by UBA scaffold 8524 gene 180. This ~9.6 kDa signal peptide cleaved protein contains a C-terminal region with a high scoring peptidoglycan-binding

Table 4. Highly Conserved and Replicated Signal Peptide Cleaved Proteins with Confirming N-Terminus Spectra[a]

name	N-terminal spectra	function
UBA LeptoII Scaffold 8049 GENE 83 SigP	531	protein of unknown function
5wayCG LeptoII Contig 11390 GENE 17 SigP	531	protein of unknown function
UBA LeptoII Scaffold 8241 GENE 693 SigP	529	periplasmic phosphate binding protein
UBA LeptoII Scaffold 8524 GENE 180 SigP	490	protein of unknown function
UBA LeptoII Scaffold 8062 GENE 372 SigP	290	cytochrome 579 variant 1
UBA LeptoII Scaffold 8062 GENE 147 SigP	290	cytochrome 579 variant 2
UBA LeptoII Scaffold 8135 GENE 9 SigP	277	conserved protein of unknown function
5wayCG LeptoII Contig 11233 GENE 46 SigP	277	conserved protein of unknown function
UBA LeptoII Scaffold 8241 GENE 153 SigP	236	protein of unknown function
UBA LeptoII Scaffold 8524 GENE 128 SigP	198	putative outer membrane protein (OmpH)
UBA LeptoII Scaffold 8241 GENE 297 SigP	100	protein of unknown function
5wayCG LeptoII Contig 11184 GENE 47 SigP	93	protein of unknown function
UBA LeptoII Scaffold 7931 GENE 111 SigP	65	putative cytochrome
5wayCG LeptoII Contig 11238 GENE 99 SigP	65	putative cytochrome
UBA LeptoII Scaffold 7931 GENE 73 SigP	54	peptidyl-prolyl cis−trans isomerase (EC 5.2.1.8)
5wayCG LeptoII Contig 11238 GENE 58 SigP	54	peptidyl-prolyl cis−trans isomerase (EC 5.2.1.8)
UBA LeptoII Scaffold 8241 GENE 348 SigP	35	putative outer membrane protein
UBA LeptoII Scaffold 7931 GENE 365 SigP	35	protein of unknown function
UBA LeptoII Scaffold 7931 GENE 101 SigP	33	protein of unknown function
5wayCG LeptoII Contig 11238 GENE 88 SigP	33	protein of unknown function
UBA LeptoII Scaffold 8062 GENE 53 SigP	32	protein of unknown function
5wayCG LeptoII Contig 11216 GENE 10 SigP	32	hypothetical protein
UBA LeptoII Scaffold 8135 GENE 71 SigP	26	secretion protein HlyD
UBA LeptoII Scaffold 8062 GENE 151 SigP	25	protein of unknown function
UBA LeptoIII Contig 7952 GENE 72 SigP	22	YceI family protein
UBA LeptoII Scaffold 8049 GENE 48 SigP	21	protein of unknown function
UBA LeptoII Scaffold 8049 GENE 366 SigP	20	conserved protein of unknown function
UBA LeptoII Scaffold 8062 GENE 173 SigP	20	cytochrome C peroxidase (EC 1.11.1.5)
UBA LeptoII Scaffold 8062 GENE 32 SigP	17	protein of unknown function
UBA LeptoII Scaffold 7931 GENE 352 SigP	12	protein of unknown function
5wayCG LeptoII Contig 10608 GENE 3 SigP	11	hypothetical protein
5wayCG LeptoII Contig 10961 GENE 20 SigP	10	protein of unknown function
UBA LeptoIII Contig 9432 GENE 53 SigP	9	hypothetical protein
UBA LeptoII Scaffold 8524 GENE 248 SigP	9	conserved protein of unknown function
5wayCG LeptoII Contig 11111 GENE 93 SigP	5	protein of unknown function
UBA LeptoII Scaffold 8241 GENE 81 SigP	5	protein of unknown function
5wayCG LeptoII Contig 11277 GENE 93 SigP	5	protein of unknown function
5wayCG LeptoII Contig 11391 GENE 14 SigP	4	conserved protein of unknown function
UBA LeptoII Scaffold 8241 GENE 238 SigP	3	protein of unknown function
UBA LeptoII Scaffold 8241 GENE 573 SigP	3	protein of unknown function
UBA LeptoII Scaffold 8524 GENE 269 SigP	3	protein of unknown function
UBA LeptoII Scaffold 8241 GENE 522 SigP	2	putative peptidyl-prolyl cis−trans isomerase
UBA LeptoII Scaffold 8524 GENE 127 SigP	2	putative bacterial surface antigen (D15)
UBA LeptoII Scaffold 8241 GENE 298 SigP	1	putative outer membrane protein
UBA LeptoII Scaffold 8524 GENE 249 SigP	1	putative OmpA family protein
UBA LeptoII Scaffold 7931 GENE 338 SigP	1	Conserved protein of unknown function

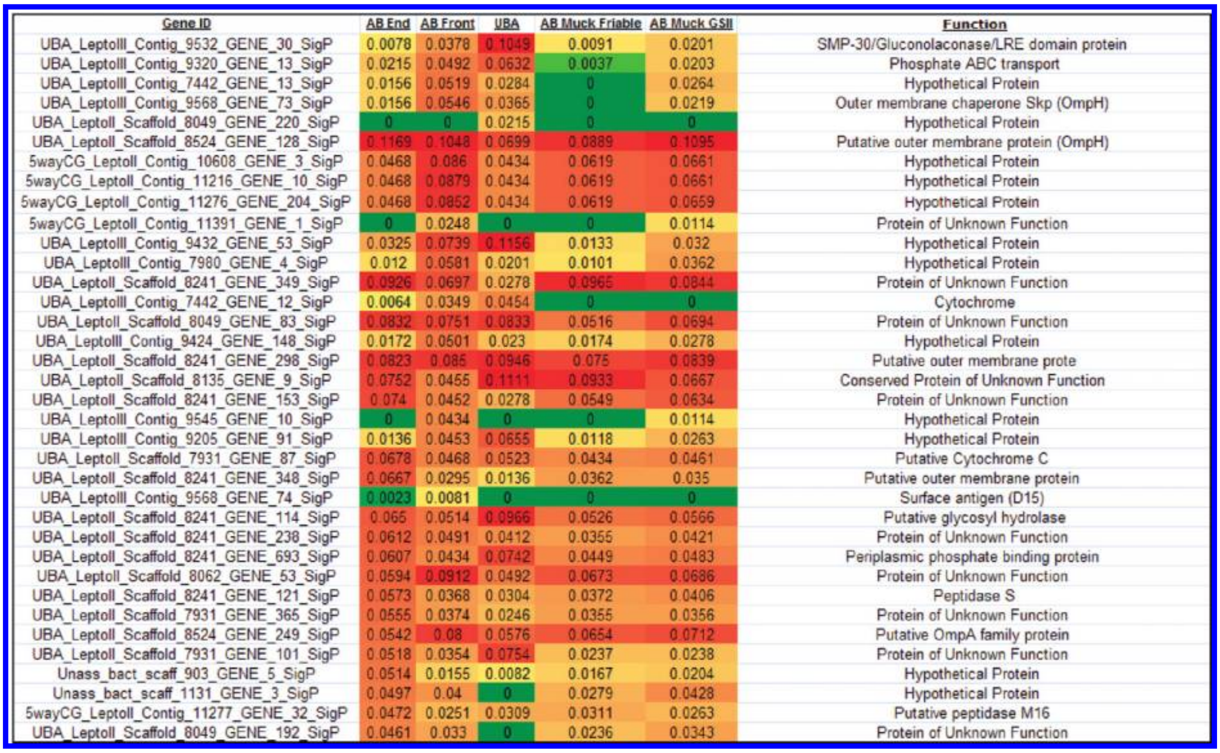
[a] The redundancy in spectral counts for the cleaved N-terminal peptides can be attributed to the overlap of protein expression among the CG and UBA strains present in the community.

domain. This domain has been previously implicated in metalloprotease functionality.[39] Edman sequencing has also identified additional N-terminal cleavages of this protein, suggesting alternate functions that may include signal transduction or peptidic defense. By identifying signal-cleaved proteins that are constitutively and highly expressed across all samples, this study has identified a conserved pool of target proteins that are strong candidates for further in-depth functional analyses.

Table 5 lists some secretome signal peptide cleaved proteins whose relative abundance *differs* according to sampling location or biofilm growth state, based on calculated NSAF values.[31] These may reflect responses to differences in the surrounding physiochemical environment as the result of changing growth state and sampling location. Subtle changes in the pH, temperature, or concentrations of heavy metals could induce changes in expression of specific proteins, such as cytochromes, solute transporters, and cofactors, as well as dehydrogenases, thioredoxins, cytochromes, and quinones. As expected, many of the proteins that exhibit the largest changes in abundances are currently annotated with an unknown function.

In some cases, the differences in expression are quite dramatic, as in a *Leptospirillum* group III protein from scaffold 9532 gene 30, which exhibits nearly a 100-fold increase in expression in the UBA location relative to the AB-Front or AB-End samples. However, it must be noted that the reduced expression of this protein could be partly accounted for by the lower abundance of this organism in the AB-drift biofilms. The protein exhibits high BLAST sequence similarity (*E*-value $<9^{-59}$) to numerous proteins containing a NHL repeat. This feature has been shown to confer catalytic activity in monooxygenases

Table 5. NSAF Comparison of Select, Highly Differential Signal Peptide Cleaved Proteins



Gene ID	AB End	AB Front	UBA	AB Muck Friable	AB Muck GSII	Function
UBA_LeptoIII_Contig_9532_GENE_30_SigP	0.0078	0.0378	0.1049	0.0091	0.0201	SMP-30/Gluconolaconase/LRE domain protein
UBA_LeptoIII_Contig_9320_GENE_13_SigP	0.0215	0.0492	0.0632	0.0037	0.0203	Phosphate ABC transport
UBA_LeptoIII_Contig_7442_GENE_13_SigP	0.0156	0.0519	0.0284	0	0.0264	Hypothetical Protein
UBA_LeptoIII_Contig_9568_GENE_73_SigP	0.0156	0.0546	0.0365	0	0.0219	Outer membrane chaperone Skp (OmpH)
UBA_LeptoII_Scaffold_8049_GENE_220_SigP	0	0	0.0215	0	0	Hypothetical Protein
UBA_LeptoII_Scaffold_8524_GENE_128_SigP	0.1169	0.1048	0.0699	0.0889	0.1095	Putative outer membrane protein (OmpH)
5wayCG_LeptoII_Contig_10608_GENE_3_SigP	0.0468	0.086	0.0434	0.0619	0.0661	Hypothetical Protein
5wayCG_LeptoII_Contig_11216_GENE_10_SigP	0.0468	0.0879	0.0434	0.0619	0.0661	Hypothetical Protein
5wayCG_LeptoII_Contig_11276_GENE_204_SigP	0.0468	0.0852	0.0434	0.0619	0.0659	Hypothetical Protein
5wayCG_LeptoII_Contig_11391_GENE_1_SigP	0	0.0248	0	0	0.0114	Protein of Unknown Function
UBA_LeptoIII_Contig_9432_GENE_53_SigP	0.0325	0.0739	0.1156	0.0133	0.032	Hypothetical Protein
UBA_LeptoIII_Contig_7980_GENE_4_SigP	0.012	0.0581	0.0201	0.0101	0.0362	Hypothetical Protein
UBA_LeptoII_Scaffold_8241_GENE_349_SigP	0.0926	0.0697	0.0278	0.0965	0.0844	Protein of Unknown Function
UBA_LeptoIII_Contig_7442_GENE_12_SigP	0.0064	0.0349	0.0454	0	0	Cytochrome
UBA_LeptoII_Scaffold_8049_GENE_83_SigP	0.0832	0.0751	0.0833	0.0516	0.0694	Protein of Unknown Function
UBA_LeptoIII_Contig_9424_GENE_148_SigP	0.0172	0.0501	0.023	0.0174	0.0278	Hypothetical Protein
UBA_LeptoII_Scaffold_8241_GENE_298_SigP	0.0823	0.085	0.0946	0.075	0.0839	Putative outer membrane prote
UBA_LeptoII_Scaffold_8135_GENE_9_SigP	0.0752	0.0455	0.1111	0.0933	0.0667	Conserved Protein of Unknown Function
UBA_LeptoII_Scaffold_8241_GENE_153_SigP	0.074	0.0452	0.0278	0.0549	0.0634	Protein of Unknown Function
UBA_LeptoIII_Contig_9545_GENE_10_SigP	0	0.0434	0	0	0.0114	Hypothetical Protein
UBA_LeptoIII_Contig_9205_GENE_91_SigP	0.0136	0.0453	0.0655	0.0118	0.0263	Hypothetical Protein
UBA_LeptoII_Scaffold_7931_GENE_87_SigP	0.0678	0.0468	0.0523	0.0434	0.0461	Putative Cytochrome C
UBA_LeptoII_Scaffold_8241_GENE_348_SigP	0.0667	0.0295	0.0136	0.0362	0.035	Putative outer membrane protein
UBA_LeptoIII_Contig_9568_GENE_74_SigP	0.0023	0.0081	0	0	0	Surface antigen (D15)
UBA_LeptoII_Scaffold_8241_GENE_114_SigP	0.065	0.0514	0.0966	0.0526	0.0566	Putative glycosyl hydrolase
UBA_LeptoII_Scaffold_8241_GENE_238_SigP	0.0612	0.0491	0.0412	0.0355	0.0421	Protein of Unknown Function
UBA_LeptoII_Scaffold_8241_GENE_693_SigP	0.0607	0.0434	0.0742	0.0449	0.0483	Periplasmic phosphate binding protein
UBA_LeptoII_Scaffold_8062_GENE_53_SigP	0.0594	0.0912	0.0492	0.0673	0.0686	Protein of Unknown Function
UBA_LeptoII_Scaffold_8241_GENE_121_SigP	0.0573	0.0368	0.0304	0.0372	0.0406	Peptidase S
UBA_LeptoII_Scaffold_7931_GENE_365_SigP	0.0555	0.0374	0.0246	0.0355	0.0356	Protein of Unknown Function
UBA_LeptoII_Scaffold_8524_GENE_249_SigP	0.0542	0.08	0.0576	0.0654	0.0712	Putative OmpA family protein
UBA_LeptoII_Scaffold_7931_GENE_101_SigP	0.0518	0.0354	0.0754	0.0237	0.0238	Protein of Unknown Function
Unass_bact_scaff_903_GENE_5_SigP	0.0514	0.0155	0.0082	0.0167	0.0204	Hypothetical Protein
Unass_bact_scaff_1131_GENE_3_SigP	0.0497	0.04	0	0.0279	0.0428	Hypothetical Protein
5wayCG_LeptoII_Contig_11277_GENE_32_SigP	0.0472	0.0251	0.0309	0.0311	0.0263	Putative peptidase M16
UBA_LeptoII_Scaffold_8049_GENE_192_SigP	0.0461	0.033	0	0.0236	0.0343	Protein of Unknown Function

and serine/threonine kinases.[40] Additionally, several high scoring BLAST hits correspond to SMP-30/gluconolaconase/LRE domain-containing proteins. This annotation describes a region of sequence similarity observed in a variety of bacterial and archaeal enzymes. A putative ABC Transporter, 5wayCG *Leptospirillum* group III contig 9320 gene 13, was also inferred to show variation in abundance levels among samples. Finally, an annotated cytochrome encoded by UBA *Leptospirillum* III scaffold 7442 gene 12 is identified in relatively high abundance in the AB-End, UBA, and AB-Front samples but is not identified in the AB-Muck samples. This is in stark contrast to the previously mentioned Cyt_{579}, which is ubiquitously identified in all samples. These results suggest that protein expression patterns reflect varying responses to local environmental conditions or biofilm age.

We conducted Pfam domain analysis on the 46 proteins identified with a signal peptide cleaved N-terminus. Nine proteins contain domains currently annotated in the Pfam database (Table 6), including cytochromes, outer membrane folds, catalytic sites from metabolic enzymes, and multiple Pfam domains. These domains correspond well with the predicted cellular extracytosolic location of the proteins. Additional domains include those involved in lipid binding, proteolytic digestion, and protein folding. Pyrrolo-quinoline quinone (PQQ) illustrates a common repeat that results in a characteristic β-propeller tertiary structure found within quinones, which are integral members of electron transport chains.[41] Within our analysis, the PQQ repeat was identified in a *Leptospirillum* group II protein, encoded by scaffold 8241 gene 348, which is currently annotated as an outer membrane protein. The *Leptospirillum* group II protein, from scaffold 8062 gene 173, displays a high scoring (9.7×10^{-79}) Pfam identification to a cytochrome *c* peroxidase domain (CCP_MauG). CCP_MauG proteins have been found within the periplasmic space of Gram-negative bacteria and are known to use two heme groups to reduce hydrogen peroxide without the formation of free radicals.[42] Another prevalent domain was the NHL tandem repeat (described above), which was identified multiple times within two proteins currently annotated as having unknown functions (encoded by *Leptospirillum* group II CG contig 11233 gene 46 and *Leptospirillum* group II UBA scaffold 8135 gene 9).[40] A YceI-like domain was also found with high confidence (3.90×10^{-53}) in a protein of unknown function from *Leptospirillum* group II (encoded by scaffold 8049 gene 366). This domain is characterized by a β-barrel motif and functions in lipid binding. A previous study of *E. coli* resulted in the identification YceI as one of three proteins currently annotated with an unknown function but showed a marked response to pH.[43] Domain prediction is not conclusive evidence for protein function, but it does provide valuable insight when coupled with the determination of extracellular location and signal peptide cleavage. The previous high-scoring domain identifications highlight the diversity of the extracellular fraction as well as the need for continued study.

The abundances of the signal peptide cleaved proteins identified in this study were examined using a more extensive and previously published data set for 28 biofilm samples[21] to more comprehensively define changes in protein abundances across the AMD environment. Samples have been classified as low or high developmental stage biofilms based on their observed maturity (see Experimental Procedures).

Of the 377 nonredundant signal peptide cleaved proteins identified in this study, 174 were also found in the 28 biofilm proteomes. The previous study focused on the whole cellular proteome, and thus the extracellular fractions of these samples were not implicitly retained and analyzed separately. Thus, this captures the composite total of all proteins identified, whether or not they are specifically exported to the extracellular region. The lower rate of identification of signal peptide cleaved proteins in this case is consistent with their inferred periplasmic

Table 6. Pfam Domain Analysis of Conserved and High Confidence Signal Peptide Cleaved Proteins

name	start no.	end no.	Pfam accession no.	*E* value	Pfam ID
5wayCG LeptoII Contig 11233 GENE 46 SigP	86	115	PF08450.3	2.60×10^{-6}	SGL
5wayCG LeptoII Contig 11233 GENE 46 SigP	137	165	PF01436.12	5.90×10^{-6}	NHL
5wayCG LeptoII Contig 11233 GENE 46 SigP	25	52	PF01436.12	7.50×10^{-5}	NHL
5wayCG LeptoII Contig 11233 GENE 46 SigP	195	223	PF01436.12	8.40×10^{-3}	NHL
5wayCG LeptoII Contig 11238 GENE 58 SigP	13	177	PF00160.12	9.00×10^{-60}	Pro_isomerase
5wayCG LeptoII Contig 11238 GENE 99 SigP	150	233	PF00034.12	1.20×10^{-3}	Cytochrom_C
5wayCG LeptoII Contig 11391 GENE 14 SigP	13	178	PF04264.4	5.90×10^{-54}	YceI
UBA LeptoII Scaffold 7931 GENE 111 SigP	150	233	PF00034.12	1.20×10^{-3}	Cytochrom_C
UBA LeptoII Scaffold 7931 GENE 338 SigP	69	104	PF08238.3	9.60×10^{-10}	Sel1
UBA LeptoII Scaffold 7931 GENE 338 SigP	177	212	PF08238.3	2.90×10^{-8}	Sel1
UBA LeptoII Scaffold 7931 GENE 338 SigP	141	176	PF08238.3	9.30×10^{-8}	Sel1
UBA LeptoII Scaffold 7931 GENE 338 SigP	105	140	PF08238.3	6.80×10^{-6}	Sel1
UBA LeptoII Scaffold 7931 GENE 338 SigP	33	68	PF08238.3	1.80×10^{-4}	Sel1
UBA LeptoII Scaffold 7931 GENE 338 SigP	213	248	PF08238.3	2.00×10^{-3}	Sel1
UBA LeptoII Scaffold 7931 GENE 73 SigP	13	177	PF00160.12	9.00×10^{-60}	Pro_isomerase
UBA LeptoII Scaffold 8049 GENE 366 SigP	14	179	PF04264.4	3.90×10^{-53}	YceI
UBA LeptoII Scaffold 8062 GENE 173 SigP	1	171	PF03150.5	9.70×10^{-79}	CCP_MauG
UBA LeptoII Scaffold 8135 GENE 9 SigP	101	130	PF08450.3	8.20×10^{-7}	SGL
UBA LeptoII Scaffold 8135 GENE 9 SigP	152	180	PF01436.12	5.90×10^{-6}	NHL
UBA LeptoII Scaffold 8135 GENE 9 SigP	25	52	PF01436.12	6.40×10^{-5}	NHL
UBA LeptoII Scaffold 8135 GENE 9 SigP	210	238	PF01436.12	8.40×10^{-3}	NHL
UBA LeptoII Scaffold 8241 GENE 298 SigP	40	139	PF00691.11	1.30×10^{-24}	OmpA
UBA LeptoII Scaffold 8241 GENE 348 SigP	254	292	PF01011.12	7.40×10^{-5}	PQQ
UBA LeptoII Scaffold 8241 GENE 348 SigP	107	144	PF01011.12	1.20×10^{-4}	PQQ
UBA LeptoII Scaffold 8241 GENE 348 SigP	213	250	PF01011.12	1.40×10^{-4}	PQQ
UBA LeptoII Scaffold 8241 GENE 348 SigP	399	435	PF01011.12	1.30×10^{-2}	PQQ
UBA LeptoII Scaffold 8241 GENE 522 SigP	66	145	PF09312.2	3.90×10^{-16}	SurA_N
UBA LeptoII Scaffold 8241 GENE 522 SigP	163	256	PF00639.12	2.00×10^{-8}	Rotamase
UBA LeptoII Scaffold 8241 GENE 522 SigP	5	26	PF09312.2	1.40×10^{-3}	SurA_N
UBA LeptoII Scaffold 8241 GENE 693 SigP	1	156	PF01547.16	7.00×10^{-4}	SBP_bac_1
UBA LeptoII Scaffold 8524 GENE 127 SigP	432	748	PF01103.14	8.70×10^{-36}	Bac_surface_Ag
UBA LeptoII Scaffold 8524 GENE 127 SigP	252	330	PF07244.6	4.20×10^{-22}	Surf_Ag_VNR
UBA LeptoII Scaffold 8524 GENE 127 SigP	333	405	PF07244.6	5.20×10^{-18}	Surf_Ag_VNR
UBA LeptoII Scaffold 8524 GENE 127 SigP	8	79	PF07244.6	1.10×10^{-14}	Surf_Ag_VNR
UBA LeptoII Scaffold 8524 GENE 127 SigP	160	249	PF07244.6	7.10×10^{-14}	Surf_Ag_VNR
UBA LeptoII Scaffold 8524 GENE 127 SigP	80	157	PF07244.6	2.00×10^{-12}	Surf_Ag_VNR
UBA LeptoII Scaffold 8524 GENE 128 SigP	1	159	PF03938.5	6.80×10^{-18}	OmpH
UBA LeptoII Scaffold 8524 GENE 180 SigP	39	90	PF01471.9	5.30×10^{-12}	PG_binding_1
UBA LeptoII Scaffold 8524 GENE 249 SigP	129	224	PF00691.11	1.10×10^{-42}	OmpA
UBA LeptoIII Contig 7952 GENE 72 SigP	13	178	PF04264.4	1.00×10^{-58}	YceI

or extracellular location. Clustering of the NSAF values for these proteins revealed distinct trends in the protein abundances with respect to developmental stage (Figure 5A). Each row in Figure 5 represents one of the 174 identified signal peptide cleaved proteins, with yellow indicating high expression (MS detection) and blue indicating low expression (MS detection). Based on the clustering of samples (across the *x*-axis), it is evident that the abundances of signal peptide cleaved proteins correlates significantly with biofilm growth state. Specifically, samples representing low developmental stage biofilms (green highlights) generally cluster tightly, but separately from a cluster of samples representing high developmental stage biofilms (blue highlights). When the clustering of proteins based on their abundances across samples is examined (down the *y*-axis), it is noted that there is a subset of predicted signal peptide cleaved proteins that exhibit high abundances in low developmental stage biofilms, but low or no detectable expression in high developmental stage biofilms. Similarly, another subset displays no or low expression in low developmental stages and increased expression in high biofilm developmental stages.

An interesting result of these analyses was the NSAF-based abundance trends of numerous signal peptide cleaved cytochromes (Figure 5B). In general, we detect early expression of class I cytochromes, whereas cytochrome oxidases appear to be abundant later in development. These results most likely denote shifts in metabolism, which occur as biofilms age. These results are consistent with the increased abundance of Cyt_{579} and *c*-type cytochromes in early development stage biofilms.[44]

Other significant differences in the abundances of proteins between the two developmental stages were also defined (Figure 5C), with many currently annotated with no known function. As identified in the analysis of the five biofilms, we noted that the low developmental stage displays numerous cytochromes that are not identified in high developmental stages. Conversely, it was found that two chemotaxis sensory proteins were in greater abundance in high developmental stages. An increase in chemotaxis protein expression may result from the depletion of nutrients that may occur as biofilms age and more organisms colonize the environment. Proteomic adaptation, through dynamic expression of signal peptide cleaved proteins, may assist these microbes in identifying regions of the biofilm where nutrients are not limiting.

Finally, the potential proteomic adaptation of secreted proteins to the highly acidic AMD environment was probed by utilizing the pool of predicted and identified signal peptide cleaved proteins as a representation of the extracellular fraction.

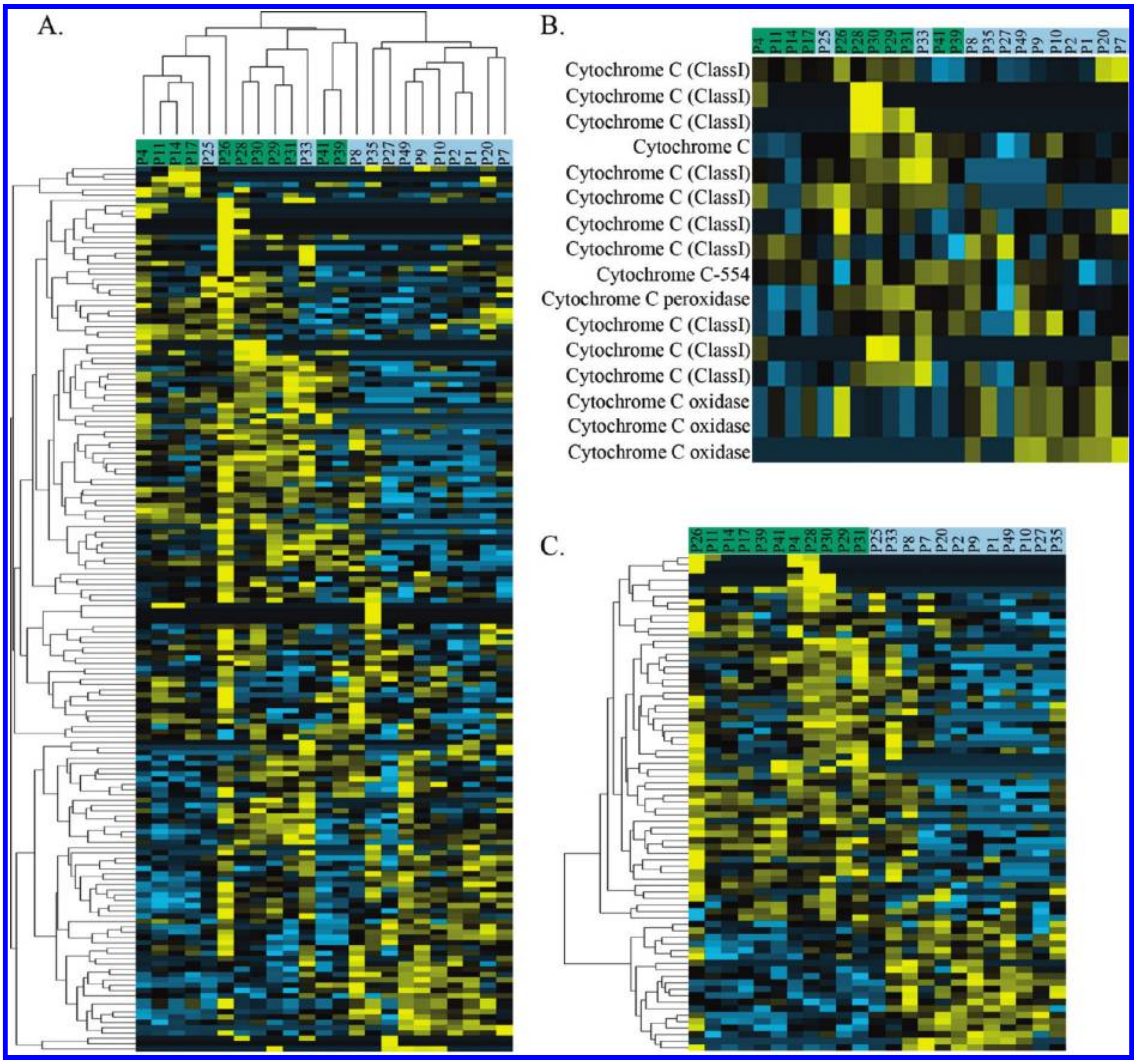


Figure 5. NSAF cluster analysis of signal peptide cleaved proteins identified in 28 biofilm samples. (A) Low developmental stage (green) and high developmental stage (blue) of the 28 biofilm samples and the clustering based on proteomic expression. Rows represent individual signal peptide cleaved proteins, with yellow indicating increased expression and blue indicating low expression. (B) Growth state expression dynamics of AMD cytochromes. (C) Subset of the proteins identified in panel A that exhibit dramatic expression changes as a function of growth state. Among these, numerous cytochromes are present in the early growth states, and several chemotaxis proteins are present in late growth stages.

Protein adaptation to acidic environments has been examined in two previous studies that have compared the calculated isoelectric points (p*I*) of proteins from organisms that tolerate and/or grow within highly acidic conditions and those from more mesophilic organisms.[45,46] The results of these two studies are in conflict, with one finding significant differences and the other not. One reason for this conflict may be that both studies include the complete genomes of each organism to calculate median p*I*'s. Including all *predicted proteins* in these analyses, even those that are not exposed to highly acidic extracellular environments, can introduce unintended biases. In an earlier study, we examined the predicted p*I*'s for proteins of the extracellular fraction of the AB-End biofilm and determined that the distribution exhibited a bimodal appearance with the largest proportion of proteins falling between 9−11.[19] However, this study did not explicitly resolve proteins with signal peptides. In this current work, the median p*I* of proteins predicted to have signal peptides from *Leptospirillum* Group II was compared to the remaining pool of proteins from this organism and a significant difference was observed (median p*I* of SigP proteins = 9.10, median p*I* of remaining proteins = 6.95; *t* test; *p*-value $<1 \times 10^{-6}$). This distribution of p*I*'s closely followed the previous study, with a significant proportion falling between 9−9.9. In this study, the protein sampling size was over 10 times larger than the previous study, providing increased confidence in the pI determination. A potential caveat of this methodology is the inability to include secondary or tertiary protein structure. For example, it has previously been found that a maltose-binding protein from a thermoacidophilic bacteria has a calculated p*I* of 6.5 and a measured p*I* of 10, and this discrepancy is due to the large number of basic residues constituting the solvent exposed face of the protein.[47] Therefore, future studies examining protein adaptation to various environments will need to account for perceived differences in amino acid sequence and p*I* within the context of protein localization and the three-dimensional structure of a given protein. Given that the identified signal peptide cleaved proteins are known to be functional outside of the cytosol, they would serve as excellent candidates for detailed biochemical analysis of protein adaptation to extreme environmental conditions of the AMD environment.

Conclusions

We have integrated computational prediction with experimental verification as a methodology for validating, characterizing, and comparing signal peptide cleavage from an acidophilic microbial consortium. The ability to validate computational prediction of signal peptide cleavage by mass spectrometry at the peptide level has enabled refinement of the secretome. Analysis of the AMD protein database resulted in the prediction of over 1000 potential signal peptide cleaved proteins. Without experimental verification, the validity and confidence of the assignments is uncertain. By combining the prediction with high-throughput LC−MS/MS techniques, we were able to confidently identify hundreds of signal peptide cleaved proteins. No marked differences in signal peptide cleaved protein identification were observed relative to the distribution of species in the biofilm, as expected. What is evident though is the degree of conservation and divergence of exported signal peptide cleaved proteins from varying sampling locations. This supports the inference that the proteome is dynamic, depending on local environmental conditions or biofilm age. These results are also supported by examining the expression patterns of the proteins identified in this study within a larger sample set (28 samples) representing 4 years of sample collection. Here, distinct sets of signal peptide cleaved proteins were associated with both low and high developmental stage biofilms. This study also highlights the predominance of proteins that are annotated as either hypothetical or with an unknown function in the expressed proteomes, since the majority of identified signal peptide cleavage proteins fall within these two categories. By combining the results of Pfam analysis with the newly obtained information of potential cellular location and signal peptide cleavage, it is possible to at least partially decipher the role of some of the putative unknown proteins. The integrated prediction and identification of proteins that are specifically targeted to extra-cytosolic locations and the characterization of their expression patterns in this study have identified numerous proteins that are essential for many key functions within the AMD system.

Acknowledgment. The authors acknowledge Dr. Vincent Denef for his experimental acquisition of the 28 biofilm samples and corresponding data sets. B.K.E. acknowledges graduate stipend support from the Genome Science and Technology Graduate School at UT-Knoxville. Research funding was provided by the DOE Genomics:GTL Program. Research sponsored by the Division of Chemical Sciences of Biological Environmental Research, U.S. Department of Energy. Oak Ridge National Laboratory is managed by University of Tennessee-Battelle LLC for the Department of Energy under contract DOE-AC05-00OR22725.

Supporting Information Available: This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Wilmes, P.; Bond, P. L. Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends Microbiol.* **2006**, *14* (2), 92–7.

(2) VerBerkmoes, N. C.; Denef, V. J.; Hettich, R. L.; Banfield, J. F. Systems biology: Functional analysis of natural microbial consortia using community proteomics. *Nat. Rev. Microbiol.* **2009**, *7* (3), 196–205.

(3) Rapoport, T. A. Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature* **2007**, *450* (7170), 663–9.

(4) Gierasch, L. M. Signal sequences. *Biochemistry* **1989**, *28* (3), 923–30.

(5) von Heijne, G. The signal peptide. *J. Membr. Biol.* **1990**, *115* (3), 195–201.

(6) Driessen, A. J.; Manting, E. H.; van der Does, C. The structural basis of protein targeting and translocation in bacteria. *Nat. Struct. Biol.* **2001**, *8* (6), 492–8.

(7) Nair, R.; Rost, B. Sequence conserved for subcellular localization. *Protein Sci.* **2002**, *11* (12), 2836–47.

(8) McGeoch, D. J. On the predictive recognition of signal peptide sequences. *Virus Res.* **1985**, *3* (3), 271–86.

(9) Reinhardt, A.; Hubbard, T. Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.* **1998**, *26* (9), 2230–6.

(10) Zhang, Z.; Wood, W. I. A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics* **2003**, *19* (2), 307–8.

(11) Nielsen, H.; Krogh, A. Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1998**, *6*, 122–30.

(12) Bendtsen, J. D.; Nielsen, H.; von Heijne, G.; Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **2004**, *340* (4), 783–95.

(13) Yates, J. R., 3rd; Eng, J. K.; McCormack, A. L. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal. Chem.* **1995**, *67* (18), 3202–10.

(14) Sadygov, R. G.; Cociorva, D.; Yates, J. R. 3rd, Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat. Methods* **2004**, *1* (3), 195–202.

(15) Cravatt, B. F.; Simon, G. M.; Yates, J. R. 3rd, The biological impact of mass-spectrometry-based proteomics. *Nature* **2007**, *450* (7172), 991–1000.

(16) Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* **2007**, *389* (4), 1017–31.

(17) Allen, E. E.; Banfield, J. F. Community genomics in microbial ecology and evolution. *Nat. Rev. Microbiol.* **2005**, *3* (6), 489–98.

(18) Tyson, G. W.; Chapman, J.; Hugenholtz, P.; Allen, E. E.; Ram, R. J.; Richardson, P. M.; Solovyev, V. V.; Rubin, E. M.; Rokhsar, D. S.; Banfield, J. F. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **2004**, *428* (6978), 37–43.

(19) Ram, R. J.; Verberkmoes, N. C.; Thelen, M. P.; Tyson, G. W.; Baker, B. J.; Blake, R. C., 2nd; Shah, M.; Hettich, R. L.; Banfield, J. F. Community proteomics of a natural microbial biofilm. *Science* **2005**, *308* (5730), 1915–20.

(20) Lo, I.; Denef, V. J.; Verberkmoes, N. C.; Shah, M. B.; Goltsman, D.; DiBartolo, G.; Tyson, G. W.; Allen, E. E.; Ram, R. J.; Detter, J. C.; Richardson, P.; Thelen, M. P.; Hettich, R. L.; Banfield, J. F. Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* **2007**, *446* (7135), 537–41.

(21) Denef, V. J.; VerBerkmoes, N. C.; Shah, M. B.; Abraham, P.; Lefsrud, M.; Hettich, R. L.; Banfield, J. F. Proteomics-inferred genome typing (PIGT) demonstrates inter-population recombination as a strategy for environmental adaptation. *Environ. Microbiol.* **2009**, *11* (2), 313–325.

(22) Goltsman, D. S.; Denef, V. J.; Singer, S. W.; Verberkmoes, N. C.; Lefsrud, M.; Mueller, R.; Dick, G. J.; Sun, C.; Wheeler, K.; Zemla, A.; Baker, B. J.; Hauser, L.; Land, M.; Shah, M. B.; Thelen, M. P.; Hettich, R. L.; Banfield, J. F. Community genomic and proteomic analysis of chemoautotrophic, iron-oxidizing "Leptospirillum rubarum" (Group II) and Leptospirillum ferrodiazotrophum (Group III) in acid mine drainage biofilms. *Appl. Environ. Microbiol.* **2009**, *75*, 4599–615.

(23) Nielsen, H.; Engelbrecht, J.; Brunak, S.; von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **1997**, *10* (1), 1–6.

(24) Yates, J. R., 3rd; Eng, J. K.; McCormack, A. L.; Schieltz, D. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.* **1995**, *67* (8), 1426–36.

(25) Tabb, D. L.; McDonald, W. H.; Yates, J. R. 3rd, DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **2002**, *1* (1), 21–6.

(26) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4* (3), 207–14.

(27) Bateman, A.; Coin, L.; Durbin, R.; Finn, R. D.; Hollich, V.; Griffiths-Jones, S.; Khanna, A.; Marshall, M.; Moxon, S.; Sonnhammer, E. L.; Studholme, D. J.; Yeats, C.; Eddy, S. R. The Pfam protein families database. *Nucleic Acids Res.* **2004**, *32* (Database issue), D138–41.

(28) Finn, R. D.; Mistry, J.; Schuster-Bockler, B.; Griffiths-Jones, S.; Hollich, V.; Lassmann, T.; Moxon, S.; Marshall, M.; Khanna, A.; Durbin, R.; Eddy, S. R.; Sonnhammer, E. L.; Bateman, A. Pfam: clans, web tools and services. *Nucleic Acids Res.* **2006**, *34* (Database issue), D247–51.

(29) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215* (3), 403–10.

(30) Singer, S. W.; Chan, C. S.; Zemla, A.; Verberkmoes, N. C.; Hwang, M.; Hettich, R. L.; Banfield, J. F.; Thelen, M. P. Characterization of Cytochrome 579, an unusual cytochrome isolated from an iron-oxidizing microbial community. *Appl. Environ. Microbiol.* **2008**, *74*, 4454–62.

(31) Zybailov, B.; Mosley, A. L.; Sardiu, M. E.; Coleman, M. K.; Florens, L.; Washburn, M. P. Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J. Proteome Res.* **2006**, *5* (9), 2339–47.

(32) Florens, L.; Carozza, M. J.; Swanson, S. K.; Fournier, M.; Coleman, M. K.; Workman, J. L.; Washburn, M. P. Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* **2006**, *40* (4), 303–11.

(33) de Hoon, M. J.; Imoto, S.; Nolan, J.; Miyano, S. Open source clustering software. *Bioinformatics* **2004**, *20* (9), 1453–4.

(34) Saldanha, A. J. Java Treeview−extensible visualization of microarray data. *Bioinformatics* **2004**, *20* (17), 3246–8.

(35) Tusher, V. G.; Tibshirani, R.; Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98* (9), 5116–21.

(36) Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **2001**, *305* (3), 567–80.

(37) Druschel, G. K.; Baker, B. J.; Gihring, T. M.; Banfield, J. F. Acid mine drainage biogeochemistry at Iron Mountain, California. *Geochem. Trans.* **2004**, *5* (2), 13–32.

(38) Tjalsma, H.; Antelmann, H.; Jongbloed, J. D.; Braun, P. G.; Darmon, E.; Dorenbos, R.; Dubois, J. Y.; Westers, H.; Zanen, G.; Quax, W. J.; Kuipers, O. P.; Bron, S.; Hecker, M.; van Dijl, J. M. Proteomics of protein secretion by Bacillus subtilis: separating the “secrets” of the secretome. *Microbiol. Mol. Biol. Rev.* **2004**, *68* (2), 207–33.

(39) Seiki, M. Membrane-type matrix metalloproteinases. *APMIS* **1999**, *107* (1), 137–43.

(40) Husten, E. J.; Eipper, B. A. The membrane-bound bifunctional peptidylglycine alpha-amidating monooxygenase protein. Exploration of its domain structure through limited proteolysis. *J. Biol. Chem.* **1991**, *266* (26), 17004–10.

(41) Xia, Z.; Dai, W.; Zhang, Y.; White, S. A.; Boyd, G. D.; Mathews, F. S. Determination of the gene sequence and the three-dimensional structure at 2.4 angstroms resolution of methanol dehydrogenase from Methylophilus W3A1. *J. Mol. Biol.* **1996**, *259* (3), 480–501.

(42) Fulop, V.; Ridout, C. J.; Greenwood, C.; Hajdu, J. Crystal structure of the di-haem cytochrome c peroxidase from Pseudomonas aeruginosa. *Structure* **1995**, *3* (11), 1225–33.

(43) Stancik, L. M.; Stancik, D. M.; Schmidt, B.; Barnhart, D. M.; Yoncheva, Y. N.; Slonczewski, J. L. pH-dependent expression of periplasmic proteins and amino acid catabolism in Escherichia coli. *J. Bacteriol.* **2002**, *184* (15), 4246–58.

(44) Singer, S. W.; Erickson, B. K.; VerBerkmoes, N. C.; Hwang, M.; Shah, M. B.; Hettich, R. L.; Banfield, J. F.; Thelen, M. P. Post-translational modification and sequence variation of redox-active proteins correlate with biofilm lifecycle in a natural microbial community. *ISME J.* **2010**, accepted for publication.

(45) Hou, S. B.; Makarova, K. S.; Saw, J. H. W.; Senin, P.; Ly, B. V.; Zhou, Z. M.; Ren, Y.; Wang, J. M.; Galperin, M. Y.; Omelchenko, M. V.; Wolf, Y. I.; Yutin, N.; Koonin, E. V.; Stott, M. B.; Mountain, B. W.; Crowe, M. A.; Smirnova, A. V.; Dunfield, P. F.; Feng, L.; Wang, L.; Alam, M. Complete genome sequence of the extremely acidophilic methanotroph isolate V4, Methylacidiphilum infernorum, a representative of the bacterial phylum Verrucomicrobia. *Biol. Direct* **2008**, 3.

(46) Futterer, O.; Angelov, A.; Liesegang, H.; Gottschalk, G.; Schleper, C.; Schepers, B.; Dock, C.; Antranikian, G.; Liebl, W. Genome sequence of Picrophilus torridus and its implications for life around pH 0. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (24), 9091–9096.

(47) Schafer, K.; Magnusson, U.; Scheffel, F.; Schiefner, A.; Sandgren, M. O. J.; Diederichs, K.; Welte, W.; Hulsmann, A.; Schneider, E.; Mowbray, S. L. X-ray structures of the maltose-maltodextrin-binding protein of the thermoacidophilic bacterium Alicyclobacillus acidocaldarius provide insight into acid stability of proteins. *J. Mol. Biol.* **2004**, *335* (1), 261–274.

PR900877A