

# An Integrated Computational and Experimental Approach to Characterize the Extracellular Proteome From a Natural Extremophilic Microbial Community

Brian Erickson<sup>1</sup>; Nathan C. VerBerkmoes<sup>1</sup>; Manesh Shah<sup>1</sup>; Steven Singer<sup>2</sup>; Michael P. Thelen<sup>2</sup>; Jillian F. Banfield<sup>3</sup>; Robert Hettich<sup>1</sup>  
<sup>1</sup>Oak Ridge National Laboratory, Oak Ridge, TN; <sup>2</sup>Lawrence Livermore Natl. Lab, Livermore, CA; <sup>3</sup>University of California, Berkeley, CA



## OVERVIEW

- The extracellular region of the acid mine drainage (AMD) system is the interface between the harsh environment and the stable, thriving microbial community. Previous studies have identified thousands of proteins, many of which have unknown function.
- Technology integration is necessary in order to further characterize the unique proteins of the low species complexity AMD microbial community.
- Protein cellular location, frequently determined by signal peptide presence, can provide clues to protein function.
- Signal peptides are a highly conserved method of protein targeting and transportation to the extracellular region.
- Signal peptides are routinely predicted by computational methods, such as SignalP-3.0.
- Mass spectrometry allows for high-throughput, peptide level information.
- Coupling computational prediction and mass spectrometry analysis results in a thorough characterization of signal peptide cleavage.
- From this analysis we aim to identify the subset of proteins exhibiting signal peptide cleavage. The combination of cellular location, signal peptide prediction and Pfam analysis may further characterize proteins with previously unknown function.

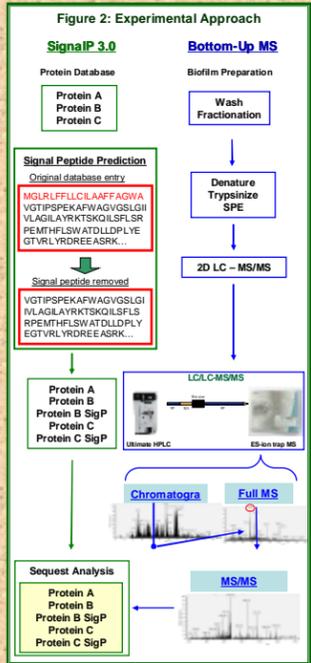
## INTRODUCTION

- Measurement of post-translational modifications often provides valuable clues to protein function. These modifications can also include targeting signals which direct proteins to various cellular locations, including the extracellular space.
- Computational based prediction of signal peptides can provide highly accurate prediction but when coupled with mass spectrometry, peptide level, non-ambiguous identifications of potentially cleaved proteins are possible.
- The results of the integrated computational and experimental characterization provides several valuable metrics including: protein identification, potential cellular location, relative abundance and ultimately potential function.
- An additional result is the relative comparison of protein expression across spatial separated sampling locations and differences in biofilm age.
- Acid Mine Drainage (AMD) - Figure 1  
 Iron Mountain Mine, CA (near Redding, CA)  
 Biofilm exists in pH < 1, -42°C, Molar/sub-molar concentrations of Fe, Zn, Cu, As  
 Genome sequence (Tyson et al., 2004, Nature)  
 Proteome Characterization (Ram et al., 2005, Science)



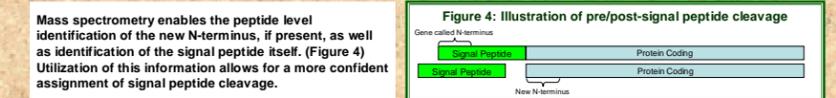
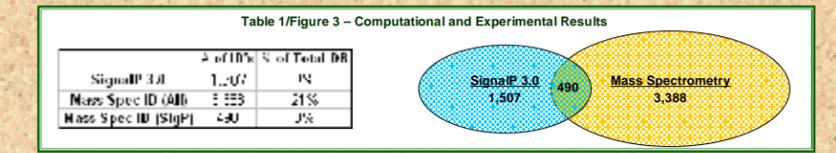
## EXPERIMENTAL

- Signal Peptide Prediction – Figure 2**
  - SignalP-3.0 was locally installed and a perl script was used to automate the submission of >16,000 predicted AMD protein sequences.
  - The proteins predicted to contain a signal peptide were identified with a "SigP" tag and the mature, signal peptide cleaved sequence was appended to the search database.
- Mass Spectrometry**
  - 5 samples from varying locations and biofilm growth state were denatured, reduced, and proteolytically digested into peptides with trypsin.
  - Peptides were measured, in triplicate, via a 2D (RP/SiX)-LC-MS/MS pipeline utilizing a linear ion trap (Thermo Finnigan LTQ).
- Proteome Bioinformatics**
  - Bottom-up spectra were searched with Sequest, utilizing the "SigP" appended database, and filtered with DTASelect. (Tabb et al., Analytical Chemistry, 2005)

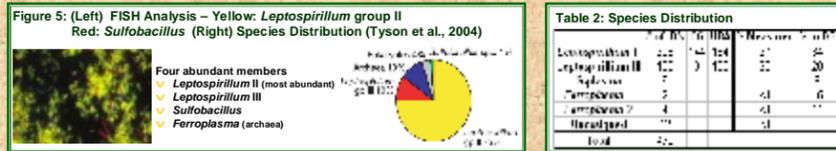


## COMPARISON OF COMPUTATIONAL VS. EXPERIMENTAL

- Mechanisms of SignalP-3.0**
  - Bacterial signal peptides do not typically contain a consensus sequence, which necessitates more complex algorithms for identification.
  - SignalP-3.0 implemented hidden Markov model (HMM) abilities for signal peptide identification and has resulted in >90% effectiveness in identifying the presence of a signal peptide and the cleavage site.
  - SignalP-3.0 is well suited for bacterial signal peptide identification, but lacks confident scoring for archaeal signal peptides due to a lack of suitable training sets.



- Computational and Experimental Results**
  - The SignalP-3.0 search resulted in the prediction of 1,507 signal peptide cleaved proteins. (Table 1/Figure 3)
    - The 2D LC-MS/MS of the extracellular samples resulted in the identification of 3,388 total proteins.
    - From this, 490 proteins were predicted to contain a signal peptide and were identified in the mass spectrometry.
  - Species Distribution (Table 2)**
    - Previous experimentation has identified 5 dominant members, with *Leptospirillum* Group II, the most abundant. (Figure 5)
    - The proteins with predicted signal peptide cleavage approximately followed the distribution of the species in the biofilm, with *Lepto.* II derived proteins dominating the signal peptide results.



- Identification of Gene-called N-termini (Table 3)**
  - Proteins that were predicted to contain a signal peptide cleavage but were measured with a genome predicted N-terminus were termed "collisions".
  - 5 proteins out of 490 were identified as collisions.
  - The low spectral counts of the gene-called N-terminus weakly support the identification, while the higher spectral counts of the new N-terminus support signal peptide cleavage.
  - These collisions may be the result of a false positive identification or the presence of a non-processed form of the protein.

Table 3: Signal Peptide Cleaved Proteins with Colliding Gene-Called N-terminus

Gene	Signal Peptide	Gene-Called N-terminus	Protein	Function
UBA1	MLKLTAL...TALVLLGASL...SGPAPAS	MLKLTAL...TALVLLGASL...SGPAPAS	UBA1	Ubiquitin-protein ligase
UBA2	LTLRLGGQY...CYSFLDQVTKALMGVPGV	LTLRLGGQY...CYSFLDQVTKALMGVPGV	UBA2	Ubiquitin-protein ligase
UBA3	VSNKTDQV...TVGGNGDKLKE	VSNKTDQV...TVGGNGDKLKE	UBA3	Ubiquitin-protein ligase
UBA4	SAVIRANGIKOS...SWYCEAHRMN	SAVIRANGIKOS...SWYCEAHRMN	UBA4	Ubiquitin-protein ligase
UBA5	GVRHVSYNTQ...DQVTVGGNGDKLKE	GVRHVSYNTQ...DQVTVGGNGDKLKE	UBA5	Ubiquitin-protein ligase

## PROTEIN CHARACTERIZATION

- CONSERVED VS. DIVERGENT SIGNAL PEPTIDE CLEAVED PROTEINS (Table 3/Table 4)**
  - The 5 samples consisted of 3 different sampling locations in the mine, as well as 2 different biofilm growth states. Each sample was measured in triplicate resulting in 15 total data sets. (Figure 6)
  - Table 3 represents a subset of signal peptide cleaved proteins which were highly conserved across all 15 MS analyses and had identifying new N-terminus spectra.
  - These 46 proteins represent critical proteins for microbial survival and proliferation due to their degree of conservation.
  - Conversely, Table 4 contains spectral counts for signal peptide cleaved proteins which are highly divergent in their expression.
  - These proteins represent those that are potentially highly specialized based on local environment differences and may represent biofilm tuning by differential protein expression.

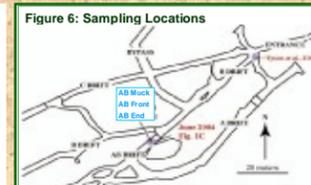


Table 3: Highly Conserved with Confirming N-Terminus spectra

Gene	Signal Peptide	Protein	Function
UBA1	MLKLTAL...TALVLLGASL...SGPAPAS	UBA1	Ubiquitin-protein ligase
UBA2	LTLRLGGQY...CYSFLDQVTKALMGVPGV	UBA2	Ubiquitin-protein ligase
UBA3	VSNKTDQV...TVGGNGDKLKE	UBA3	Ubiquitin-protein ligase
UBA4	SAVIRANGIKOS...SWYCEAHRMN	UBA4	Ubiquitin-protein ligase
UBA5	GVRHVSYNTQ...DQVTVGGNGDKLKE	UBA5	Ubiquitin-protein ligase

Table 4: Spectral Counts of Divergent Signal Peptide Cleaved Proteins

Gene	Signal Peptide	Protein	Function
UBA1	MLKLTAL...TALVLLGASL...SGPAPAS	UBA1	Ubiquitin-protein ligase
UBA2	LTLRLGGQY...CYSFLDQVTKALMGVPGV	UBA2	Ubiquitin-protein ligase
UBA3	VSNKTDQV...TVGGNGDKLKE	UBA3	Ubiquitin-protein ligase
UBA4	SAVIRANGIKOS...SWYCEAHRMN	UBA4	Ubiquitin-protein ligase
UBA5	GVRHVSYNTQ...DQVTVGGNGDKLKE	UBA5	Ubiquitin-protein ligase

- Identified Signal Peptide Cleaved Protein Function (Figure 7)**
  - The majority of identifications (>58%) correspond to proteins which are annotated as hypothetical or with an unknown function. The remaining proteins are annotated with functions consistent with an extracellular location.

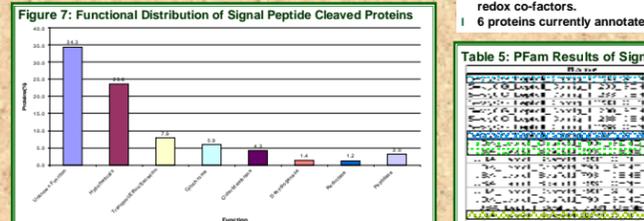


Figure 8: Top Down Validation  
 UBA1 Leptol Scaffold 7931 GENE 101 SigP  
 With Signal Peptide (11,020,837 Da):  
 MLKLTAL...TALVLLGASL...SGPAPAS  
 LTLRLGGQY...CYSFLDQVTKALMGVPGV  
 VSNKTDQV...TVGGNGDKLKE  
 SAVIRANGIKOS...SWYCEAHRMN  
 Signal Peptide Cleaved (6,591,706 Da):  
 APEKLT...LRLGGQY...CYSFLDQVTKALMGVPGV  
 GVRHVSYNTQ...DQVTVGGNGDKLKE  
 RAINGVKSSWY...CEAHRMN  
 Potential disulfide bond (-2 Da)

- Pfam Analysis of Highly Conserved, High Confident ID's (Table 5)**
  - The Pfam results correlate well with proteins located in the extracellular space. Domain identifications include cytochromes, cell-cell signaling, lipid binding and redox co-factors.
  - 6 proteins currently annotated with an unknown function matched to Pfam domains

Table 5: Pfam Results of Signal Peptide Cleaved Proteins

Gene	Signal Peptide	Protein	Function
UBA1	MLKLTAL...TALVLLGASL...SGPAPAS	UBA1	Ubiquitin-protein ligase
UBA2	LTLRLGGQY...CYSFLDQVTKALMGVPGV	UBA2	Ubiquitin-protein ligase
UBA3	VSNKTDQV...TVGGNGDKLKE	UBA3	Ubiquitin-protein ligase
UBA4	SAVIRANGIKOS...SWYCEAHRMN	UBA4	Ubiquitin-protein ligase
UBA5	GVRHVSYNTQ...DQVTVGGNGDKLKE	UBA5	Ubiquitin-protein ligase

## DISCUSSION

- Conserved & Divergent**
  - The unique and extreme environment of the AMD system necessitates that colonizing microbes express a selective and conserved proteome. It follows then, that a significant proportion of proteins will be expressed throughout the mine and across temporal differences in biofilm age. In this study, 46 proteins were identified across all locations with probable signal peptide cleavage. It is highly likely that these proteins represent the critical protein components for microbial survival. On the other hand, a large number of proteins were identified that displayed significantly varying spectral counts, indicating probable differences in protein expression. These proteins may represent the subset of proteins that are finely tuned to subtle environment changes or differences in biofilm age. This subset of variably expressed proteins are of significant interest and are targets for further study.
- Pfam**
  - Identification of functional domains can help provide clues of protein function. This analysis provides further support for protein location, by correlating functional domains with our analysis.
- Top-Down Characterization**
  - A distinct extracellular fraction from the cDfRt location of the mine was fractionated and measured by LC-MS/MS on a 9.47 FT-ICR for intact protein measurement and verification of signal peptide cleavage. One example in Figure 8 displays the deconvoluted mass of Gene 101 after signal peptide removal as measured in the ICR. The 2 Da difference may be a disulfide bond in the intact protein. Top-down mass spectrometry provides a 3<sup>rd</sup> level of signal peptide cleavage validation.

## CONCLUSIONS

- Computational prediction coupled with mass spectrometric verification is a rapid and robust method of confidently identifying signal peptide cleavage.
- 490 proteins from the AMD microbial community were predicted and identified by mass spectrometry.
- <1% of the proteins resulted in collisions, with predicted signal peptide cleavage and identified, low spectral count gene called N-termini peptides.
- 46 proteins were identified across all replicates and sample locations as having a high probability of signal peptide cleavage.
- Pfam analysis has confirmed previous annotations and present new domains for hypothetical and unknown proteins. When coupled with the potential cellular location as well as identifying signal peptide cleavage, it is possible to postulate putative function.

## ACKNOWLEDGMENTS

- B. Erickson acknowledges financial support from the UTK-ORNL GST program.
- This research sponsored by U.S. Department of Energy, Genomics: Genomes-to-Life Program under contract DE-AC05-00OR22725 with Oak Ridge National Laboratory, managed and operated by UT-Battelle, LLC.