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

**Brian Erickson[1]; Nathan C. VerBerkmoes[1]; Manesh Shah[1]; Steven Singer[2]; Michael P. Thelen[2]; Jillian F. Banfield[3]; Robert Hettich[1]**

*[1]Oak Ridge National Laboratory, Oak Ridge, TN; [2]Lawerence Livermore Natl. Lab, Livermore, CA; [3]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

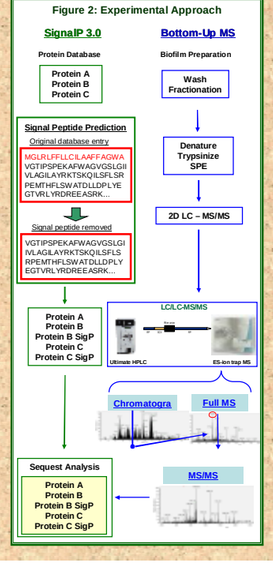
- Measurement of post-translational modifications often provides valuable clues to protein function. These modifications can also include targeting signals which direct proteins to various cellular locations, including the extracellular space.
- Computational based predication of signal peptides can 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*)

**Figure 1:** (A) Acid Mine Drainage (B) Biofilm Collection (C) Location of AMD (Images: Dr. Jillian Banfield)

## EXPERIMENTAL

- Signal Peptide Prediction – Figure 2
  - Signal-P 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/SCX)-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)

**Figure 2: Experimental Approach**



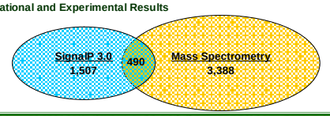

## COMPARISON OF COMPUTATIONAL VS. EXPERIMENTAL

**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 archael signal peptides due to a lack of suitable training sets.

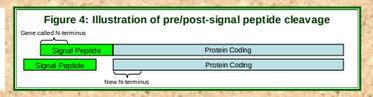
**Table 1/Figure 3 – Computational and Experimental Results**

| | # of IDs | % of Total IDs |
|---|---|---|
| SignalP 3.0 | 1,507 | 9% |
| Mass Spec ID (All) | 5,353 | 21% |
| Mass Spec ID (SigP) | 490 | 3% |




Mass spectrometry enables the peptide level identification of the new N-terminus, if present, as well as identification of the signal peptide itself. (Figure 4) Utilization of this information allows for a more confident assignment of signal peptide cleavage.

**Figure 4: Illustration of pre/post-signal peptide cleavage**




**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 *Leptospirilium* 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.

**Figure 5: (Left) FISH Analysis – Yellow: *Leptospirillum* group II Red: *Sulfobacillus* (Right) Species Distribution (Tyson et al., 2004)**

Four abundant members
- *Leptospirillum* II (most abundant)
- *Leptospirillum* III
- *Sulfobacillus*
- *Ferroplasma* (archaea)

**Table 2: Species Distribution**

| | | | | | |
|---|---|---|---|---|---|
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| Total | [illegible] | | | | |

**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**

| Name | [illegible] | [illegible] | Total Spectra Count | [illegible] | [illegible] |
|---|---|---|---|---|---|
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |

## 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.

**Figure 6: Sampling Locations**

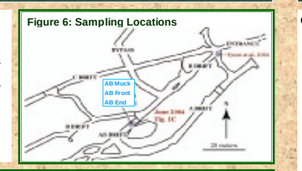



**Table 3: Highly Conserved with Confirming N-Terminus spectra**

| Name | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
|---|---|---|---|---|---|
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |