The Dynamic Range Challenges for Proteomics in Microbial Communities

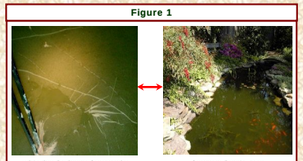
N. C. VerBerkmoes[1], R. J. Ram[2], M. Thompson[1], C. Shook[1], V. Denef[2], A. Borole[1], B. Raman[1], M. Shah[1], B. Davison[1], M. P. Thelen[3] J. F. Banfield[2], R. L. Hettich[1]

1) Oak Ridge National Laboratory, Oak Ridge, TN 2) University of California, Berkeley, CA 3) Lawrence Livermore National Laboratory, Livermore, CA

OVERVIEW

- Microbial communities play key roles in the Earth's biogeochemical cycles. Our knowledge of the structure and function of these communities is limited because analyses of microbial physiology and genetics have been largely confined to isolates grown in laboratories.
- Recent acquisition of genomic data directly from natural samples has begun to reveal the genetic potential of communities (Tyson, Nature 2004) and environments (Venter, Science 2004) thus opening the door for microbial community proteomics.
- The challenges of obtaining useful proteomic information from these samples are diverse and daunting. One of the most challenging aspects of these types of measurements is the dynamic range of the species and proteins they express in the sample. Microbial communities can vary from a few microbial species in extreme environments or hundreds to thousands in soil and water samples (Figure 1).
- The concentration range of individual species can range from 80-90% dominating species to less than 0.1% of the total sample.
- Our methodology for community proteome analyses involves cell lyses, protein fractionation, protein denaturation and digestion with trypsin followed by automated analysis with 2-dimensional nano-LC-MS/MS Level one (Figure 2)
- This study explores the challenges of these communities and how new mass spectrometry technology can further enable deeper more accurate measurements.

Figure 1



Acid Mine Drainage Community
Life at pH less than 1
4-5 community members
1-2 dominate members at a time

Common Pond Water
Life in organic and energy abundance
10,000 -1 Million Members

METHODS

Project Overview

- This research project is a combined proteogenomics approach with three main research areas:
  1) environmental sample collection, genetics, physiology and geochemistry,
  2) MS based shotgun proteomics and proteome informatics of enriched proteins, protein complexes and proteomes
  3) biochemistry and functional characterization of enriched proteins and protein complexes directly from the environment (Figure 3).

Samples and Sample Preparation

- The Acid Mine Drainage (AMD) biofilm samples used in this study were collected from the underground regions of the Richmond Mine at Iron Mountain near Redding, California (USA). These biofilms grew on the surface of acidic (pH ~0.8) sulfuric acid-rich, hot (~42°C), metal-contaminated solutions. All proteome samples were denatured and reduced, digested with trypsin, desalted, and concentrated prior to analyses.

LC/LC-MS/MS and Informatics

- All samples were analyzed via two-dimensional (2D) nano-LC MS/MS system with a split-phase column (RP-SCX-RP) on a linear ion trap (LTQ) or LTQ-Orbitrap (Thermo Finnigan) (Figure 2).
- All MS/MS spectra were searched with the SEQUEST algorithm and filtered with DTASelect/Contrast at the peptide level [Xcorrs of at least 1.8 (+1), 2.5 (+2) 3.5 (+3)]. Only proteins identified with *two fully tryptic peptides* were considered for further biological study.
- Databases included the predicted AMD proteome from the original genome sequencing project (AB End site Tyson, Nature 2004) and a recently available partial sequence from the UBA collection site (contained all of Lepto II predicted proteome).
- All data files from the initial AMD study, databases and resulting MS/MS identifications, can be downloaded from the AMD Proteome Website Analysis Page (http://compbio.ornl.gov/biofilm_amd/). *Linkable spectra for identified peptides are downloadable, a step towards open access proteome results* (Ram, Science, 2005).

Figure 3 Program Organization

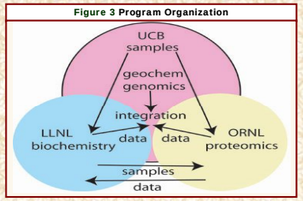
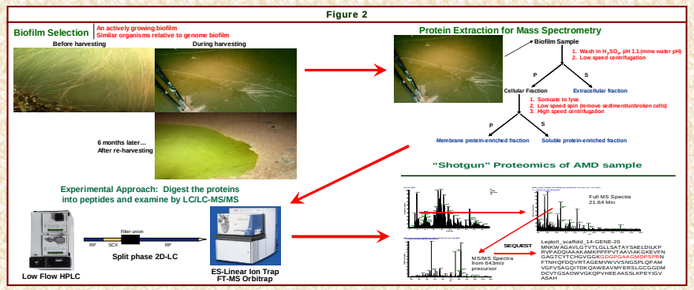


Figure 2



RESULTS

Initial AMD Proteome Characterization

- A combined proteogenomic characterization was applied to a section of AMD biofilm collected from AB drift of the Richmond Mine (Ram et al, Science, 2005).
- We have used the same approach for the characterization of the AMD proteomes from a variety of sites from the mine (Figure 3) (Table 1), species abundance was compared with fluorescent *in-situ* hybridization results.
- The UBA site is of great interest since shotgun DNA sequencing and annotation of the exact same site is almost finished allowing for the *first direct comparison of genome and proteome from the same environmental sample.*

Figure 3

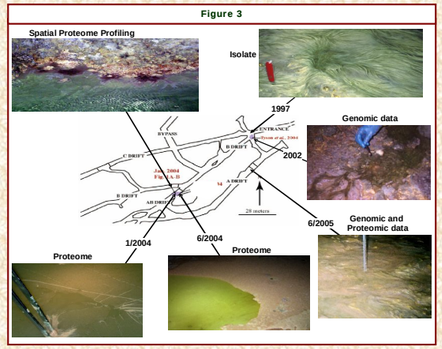


Table 1						
		Species Percentages based on Protein IDs*				Agreed with
Sample	Total Protein IDs	Fer1	Fer2	GP1	LeptoII	LeptoIII	FISH Abunance
AB End	2036	3.7	4.4	5.9	72.9	12.9	Yes
AB Front Sample 1	2020	14.3	11.9	34.1	29.3	10.3	No
AB Front Sample 2	1902	8.6	7.2	23.6	44.5	15.8	Yes
UBA	2107	2.8	4.3	8.2	58.6	25.9	Yes
C Pink	1378	5.7	6.1	19.1	50.9	18	Yes

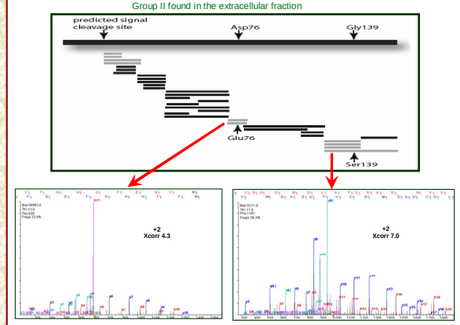
* Also calculated by spectral counts giving similar results

Abundant Biofilm Proteins and SNPs

- The most abundant proteins are likely to be critical to the biofilm community. Determining relative or absolute abundance of individual proteins is a recognized challenge in MS-based proteomics. % sequence coverage, # of unique peptide hits, and MS/MS spectral counts are all potential indicators for protein abundance.
- By looking at proteins with high spectral counts but not complete sequence coverage it is possible to hypothesis about areas of proteins modified by N-terminal cleavages, post-translational modifications and amino acid variants.
- Of the proteins highly enriched in the extracellular fraction was found with very high spectral count but average sequence coverage. Further analyses by PCR and N-terminal cleavage prediction predicted an N-terminal cleaved peptide and two SNPs. Figure 4 describes how these were all verified by MS-proteomics.
- The variant was first characterized in the AB end sample but has also now been found in the other community samples.

Figure 4

Characterization of Scaff_14_GENE_20
An unknown abundant protein from *Leptospirillum* Group II found in the extracellular fraction



LeptoII_scaff_14_GENE_20	
Protein Region Genome Database	**Spectral Counts**
. . . HQFDQVR…	3
. . . DWVGKQF…	5
Protein Region SNP Database	
. . . HQFEQVR…	162
. . . DWVSKQF…	274

The Effects of SNPs on Protein Identification

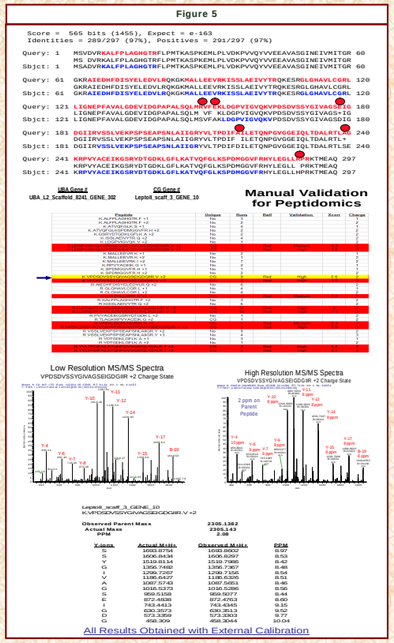
- Protein identification from community samples will almost always rely on a composite genome or reference genome related to the environmental sample being analyzed.
- Thus quite often, as shown in Figure 4, amino acids substitutions will occur at some random rate among proteins from the various species.
- Table 2 illustrates the dramatic effect of random substitutions at various percentages would have on the identification of peptides from the original AMD dataset.

Table 2					
Random Substitutions					
		Detected Proteins at Filter Level			
Mutated Database %Substitutions	Detected Peptides	1	2	3	
0	23121	5090	2875	1934	
0.1	22955	5087	2868	1927	
0.5	21209	4926	2743	1819	
1	19418	4708	2592	1719	
2	16378	4368	2304	1552	
5	9878	3388	1674	1087	
10	4580	2195	894	492	
20	1015	758	170	58	
50	9	9	0	0	

Validating Recombination Points with High Resolution Peptidomics

- Recently we have obtained a second genome sequence from an AMD biofilm community (UBA collection site). The assembly and annotation of the *Leptospirillum* group II genome from this sample is near complete.
- We appended this new database to the original database from the CG site and re-searched the entire AB end proteome.
- Large numbers of SNPs have been found in large groups moving from CG type to UBA type suggesting the possibility of a new recombinant form of *Leptospirillum* group II in the AB end biofilm.
- Figure 5 illustrates an single example of a protein with the CG type variant and how High resolution peptidomics can be used to validate individual peptide identifications thus mapping across variant regions with high accuracy.

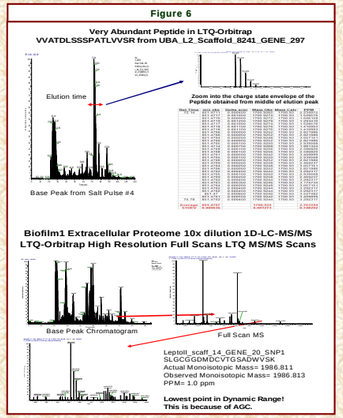
Figure 5



Dynamic Range Challenges in the AMD Proteome

- Due to the high abundance of some species and some proteins in the AMD fractions, dynamic range is a major challenge.
- One question is how well mass accuracy will hold for extremely abundant peptides in the LTQ-Orbitrap.
- Figure 6 top illustrates the high mass accuracy obtained across the entire peak area of a high abundance peptide with external calibration.
- Another concern is how well mass accuracy holds across low abundant peptides where only a single full scan and MS/MS scan are obtained. Figure 6 bottom illustrates this situation.

Figure 6



CONCLUSION

- We have established that current MS based proteome technologies can characterize simple microbial communities, even with reference genomes.
- There are clearly many challenges that need to be overcome.
- We are currently developing new methodologies in mass spectrometry and proteome informatics to tackle these challenges and further advance proteogenomics approaches for analyzing natural microbial communities.

ACKNOWLEDGMENTS

- This research was supported by grants from the DOE Microbial Genome Program, NSF Biocomplexity Program, DOE Energy Biosciences Program, and DOE Genomes to Life Program.
- We thank Dr. David Tabb and the Yates Proteomics Laboratory at Scripps Research Institute for DTASelect/Contrast software, the Institute for Systems Biology for proteome bioinformatics tools used in analysis of the MS data, and Dr. Frank Larimer at ORNL for providing computational resources.