Environmental Proteomics of SAR11 and OM43 Clade Proteobacteria in Oregon Coastal Seawater

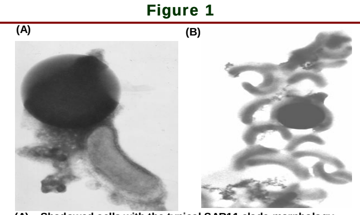
P. E. Abraham[1], S. M. Sowell[2], J. B. Kitner[2], M. G. Lefsrud[1], M. Shah[1], R. L. Hettich[1], N. C. VerBerkmoes[1], S. J. Giovannoni[2]

1) Oak Ridge National Laboratory, Oak Ridge, Tennessee 2) Oregon State University, Corvallis, Oregon

OVERVIEW

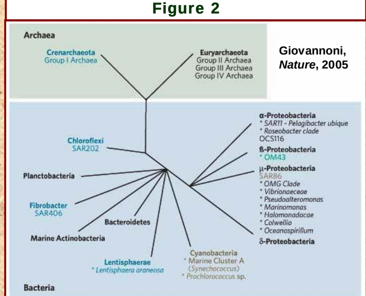
- Overlying the surface of the coastal seawater are some of the largest microbial communities on the planet. Accounting for the major biomass of the coastal sea, these communities play major roles in sustaining not only their own ecosystem but also the entire global ecosystem. Coastal phytoplankton account for nearly 50% of global carbon fixation, and half of the carbon fixed into organic matter is rapidly respired by heterotrophic microorganisms (Giovannoni, *Nature* 2005).
- There are specific interactions occurring between bacterioplankton and phytoplankton that are still unknown. Therefore, activity and cell counting measurements have been taken to link the interactions between these two communities.
- There are several dominant bacterioplankton members that are assembled over continental shelves; see position in phylogenetic tree in Figure 2. Among these dominant members are the SAR11 clade (alpha-proteobacteria) (Figure 1) located outside the bloom and the OM43 clade (beta-proteobacteria) highly contained within the bloom (contains MB2181).
- To understand the contribution of OM43 clade and SAR11 to coastal ecology, we used mass spectrometry based proteomics and the recently determined genomes of OM43 strain MB2181 and SAR11 to examine the metaproteomes of two samples 5 miles off the Oregon Coast along the Newport harbor hydroline.
- Our methodology for environmental proteome analyses involved sample collection (~100L of sea water), tangential flow filtration, centrifugation, single tube cell lyses, protein denaturation and digestion with trypsin followed by automated analysis with 2-dimensional nano-LC-MS/MS on a LTQ-Orbitrap (Figure 3).
- The goal of this study was to use MS-based proteomics to characterize the metabolic potential and dominant species occurring in Oregon coastal seawater.

Figure 1



(A) Shadowed cells with the typical SAR11 clade morphology.
(B) Negatively stained cell.
The latex beads in A and B have a diameter of 0.514 mm.
Rappé, *Nature*. 2002

Figure 2



Giovannoni, *Nature*, 2005

METHODS

- We applied 2D-LC-MS/MS for the characterization of two ocean microbial communities collected 5 miles off the coast of Newport, Oregon (44°39.1'N, 124°10.6'W) at depth 10m. 100 Liters of seawater containing ~7x10^5 cells/ml were pre-filtered through a filter removing larger, non-bacterial organisms.
- The flow-through, containing ~5x10^5 cells/ml, was concentrated using tangential flow filtration (Figure 3) followed by centrifugation (which resulted in 21mg and 47mg wet weight pellets) . The microbial pellets were lysed via a single tube small processing method, and the proteins were then denatured, reduced, and digested with trypsin. Samples were analyzed with technical duplicates via two-dimensional (2D) nano-LC MS/MS system with a split-phase column (RP-SCX-RP) on a hybrid linear ion trap-Orbitrap (Thermo Finnigan). Full scans were acquired at 30,000 resolution in the Orbitrap; MS/MS scans were acquired in the LTQ (Figure 4).
- A protein database was generated from *Pelgibacter ubique* genome (SAR11 clade), the MB 2181 genome (OM43 clade) as well as *R. palustris, S. oneidensis,* and the AMD biofilm (Tyson, Nature 2004) as a distracter database. Reverse database searches were run for a subset of the dataset to test for false positives.
- All MS/MS spectra were searched with the SEQUEST algorithm, against databases above, and filtered with DTASelect/Contrast at the peptide level with conservative filters [Xcorrs of at least 1.8 (+1), 2.5 (+2) 3.5 (+3)]. Only proteins identified with two fully tryptic peptides at conservative filters were considered for further biological study.
- For estimating false positive and false negative levels a subset of the data were re-filtered at liberal Xcorrs filters of at least 1.5 (+1), 2.0 (+2) 3.0 (+3)] and two unique peptides

Figure 3

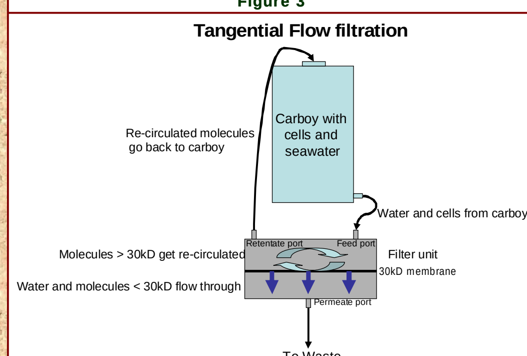
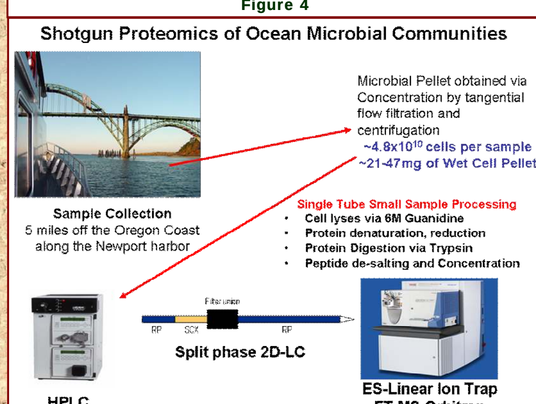


Figure 4



RESULTS AND DISCUSSION

Global Results

- In this study, we extracted microbial communities from 100L of sea water via tangential flow filtration (Figure 3) followed by centrifugation at two similar locations at different times, both of these locations were not during a diatom bloom.
- Two pellets were obtained (21 mg and 47 mg wet cell weights).
- The microbial pellets were lysed via a single tube lyses method and digested into peptides.
- The complex peptides mixtures were run by 2D-LC-MS/MS on both a LTQ and LTQ-Orbitrap.
- The chromatograms were very complex, with 100,000s of peptides.
- All MS/MS spectra were searched with SEQUEST against two abundant ocean isolate genomes (SAR11 and OM43). Two isolate microbes from soil and the acid mine drainage metagenome were used as distracters.
- In total, 654 proteins were identified with at least two unique peptides and conservative filters.
- Table 1 list the total proteins, unique spectra, and non-unique spectra from each database entry type.
- In the 21 mg sample, no single species dominated, but in the 47 mg sample, SAR11 clearly dominated with ~80% of the unique spectra from every run. In a recent study, the SAR11 was shown to be a dominating member outside the diatom blooms (Giovannoni, *Environ Microbiol*, 2006).
- Figure 5 illustrates the functional categories for all SAR11 proteins identified (412 total proteins).
- The MB2181 clade did not dominate in the dataset, which supports a recent study finding that the OM43 clade was a dominating lineage within diatom blooms but not outside of them.

Table 1

Ocean_NH5_10m_21mg_LTQ_Run 1					
Species	Proteins	Unique Spectra	% of Total	Non Unique Spectra	% of Total
MB2181	18	62	12.65%	47	13.99%
SAR11	41	162	33.06%	58	17.26%
AMD Biofilm	32	45	9.18%	54	16.07%
R. Palustris	44	139	28.37%	106	31.55%
Shewanella	28	82	16.73%	71	21.13%
Totals	163	490		336	

Ocean_NH5_10m_47mg_LTQ_Run 1					
Species	Proteins	Unique Spectra	% of Total	Non Unique Spectra	% of Total
MB2181	24	99	5.99%	75	15.53%
SAR11	176	1277	77.21%	91	18.84%
AMD Biofilm	50	58	3.51%	106	21.95%
R. Palustris	43	123	7.44%	132	27.33%
Shewanella	38	97	5.86%	79	16.36%
Totals	331	1654		483	

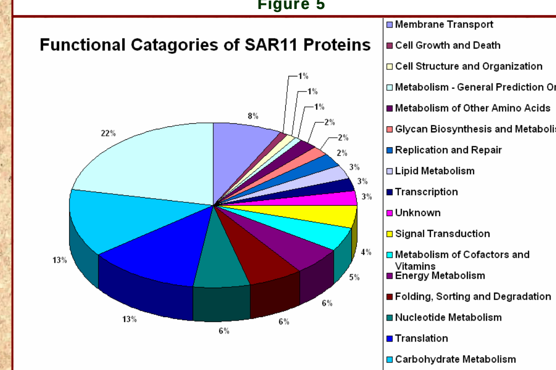
Ocean_NH5_10m_21mg_OrbiFull_30K_Run1					
Species	Proteins	Unique Spectra	% of Total	Non Unique Spectra	% of Total
MB2181	21	78	16.81%	38	10.33%
SAR11	29	152	32.76%	56	15.22%
AMD Biofilm	33	56	12.07%	71	19.29%
R. Palustris	38	108	23.28%	144	39.13%
Shewanella	25	70	15.09%	59	16.03%
Totals	146	464		368	

Ocean_NH5_10m_47mg_OrbiFull_30K_Run1					
Species	Proteins	Unique Spectra	% of Total	Non Unique Spectra	% of Total
MB2181	26	130	5.43%	78	10.94%
SAR11	199	1931	80.59%	150	21.04%
AMD Biofilm	42	46	1.92%	139	19.50%
R. Palustris	46	149	6.22%	240	33.66%
Shewanella	41	140	5.84%	106	14.87%
Totals	354	2396		713	

Ocean_NH5_10m_21mg_OrbiFull_30K_Run2					
Species	Proteins	Unique Spectra	% of Total	Non Unique Spectra	% of Total
MB2181	19	89	20.65%	43	11.94%
SAR11	26	161	37.35%	55	15.28%
AMD Biofilm	24	32	7.42%	76	21.11%
R. Palustris	29	81	18.79%	130	36.11%
Shewanella	20	68	15.78%	56	15.56%
Totals	118	431		360	

Ocean_NH5_10m_47mg_OrbiFull_30K_Run2					
Species	Proteins	Unique Spectra	% of Total	Non Unique Spectra	% of Total
MB2181	22	139	5.68%	90	11.36%
SAR11	170	1920	78.40%	157	19.82%
AMD Biofilm	47	66	2.69%	167	21.09%
R. Palustris	43	173	7.06%	245	30.93%
Shewanella	40	151	6.17%	133	16.79%
Totals	322	2449		792	

Figure 5



Abundant SAR11 Proteins

- We organized the identified SAR11 proteins by total spectral counts.
- Table 2 illustrates the top 25 SAR11 proteins from one run on the LTQ-Orbitrap (arranged by total spectral counts).
- Many known abundant proteins such as GroEL, ATP synthase, elongation factors (EF-Tu and EF-G), ABC transport systems, and ribosomal proteins were found as abundant proteins.
- Interestingly, in the top 25 proteins, three abundant proteins were identified as hypothetical proteins with no known function.
- Figure 6 illustrates the sequence coverage and two identified spectra for a 35 kDa hypothetical protein found with 44% sequence coverage, 13 peptides, and 47 spectral count.

Table 2

Gene #	Peptides	Spectral Count	Sequence Coverage	MW	pI	Function
SAR11|gi|71083646|ref|YP_266366.1|	40	181	78.60%	37685	6.3	ABC transporter
SAR11|gi|71082971|ref|YP_265690.1|	26	88	57.40%	42158	9.2	TRAP dicarboxylate transporter - DctP subunit
SAR11|gi|71082868|ref|YP_265587.1|	35	84	48.20%	58421	5.4	GroEL 60 kDa chaperonin
SAR11|gi|71083509|ref|YP_266228.1|	24	84	50.80%	33123	5.4	taurine transport system periplasmic protein
SAR11|gi|71082935|ref|YP_265654.1|	28	82	57.40%	50728	5	F1-ATP Synthase Beta Chain
SAR11|gi|71084024|ref|YP_266744.1|	24	73	49.30%	40582	5.2	spermidine/putrescine-binding periplasmic protein
SAR11|gi|71083499|ref|YP_266218.1|	13	47	44.10%	35140	6.6	hypothetical protein
SAR11|gi|71083894|ref|YP_266614.1|	15	46	53.70%	34506	9.5	putative tricarboxylic transport TctC
SAR11|gi|71083074|ref|YP_265793.1|	27	42	39.60%	70040	5.1	Chaperone protein
SAR11|gi|71083810|ref|YP_266530.1|	26	41	19.20%	152477	5.6	DNA-directed RNA polymerase
SAR11|gi|71083989|ref|YP_266709.1|	23	41	64.00%	33805	6.2	Substrate-binding region of ABC-type glycine betaine transport system
SAR11|gi|71083564|ref|YP_266283.1|	12	40	35.30%	39723	7.2	hypothetical protein
SAR11|gi|71083471|ref|YP_266190.1|	19	38	52.70%	44902	8.5	probable binding protein component of ABC sugar transporter
SAR11|gi|71083809|ref|YP_266529.1|	32	38	28.90%	154508	8.6	DNA-directed RNA polymerase beta prime chain
SAR11|gi|71083817|ref|YP_266537.1|	24	35	71.50%	43195	5.7	translation elongation factor EF-Tu
SAR11|gi|71082976|ref|YP_265695.1|	11	32	35.00%	57075	6.5	putative ABC transporter solute-binding protein
SAR11|gi|71083806|ref|YP_266526.1|	17	31	36.30%	76606	5.3	translation elongation factor EF-G
SAR11|gi|71084033|ref|YP_266753.1|	20	31	57.60%	39990	5.5	hypothetical protein
SAR11|gi|71083097|ref|YP_265816.1|	14	25	36.70%	75679	5.4	polyribonucleotide nucleotidyltransferase
SAR11|gi|71084048|ref|YP_266768.1|	16	24	36.10%	44786	8.8	Leu/Ile/Val-binding protein precursor
SAR11|gi|71083649|ref|YP_266369.1|	10	23	32.00%	27549	5.7	general L-amino acid transport ATP-binding protein
SAR11|gi|71082937|ref|YP_265656.1|	13	22	25.60%	55569	5.7	ATP synthase subunit A
SAR11|gi|71083811|ref|YP_266531.1|	9	22	62.80%	12603	5.1	probable 50S ribosomal protein L31
SAR11|gi|71082867|ref|YP_265586.1|	13	21	83.30%	10330	6.1	10 kDa chaperonin
SAR11|gi|71083359|ref|YP_266078.1|	14	21	47.10%	45054	8.4	probable Leu/Ile/Val-binding protein

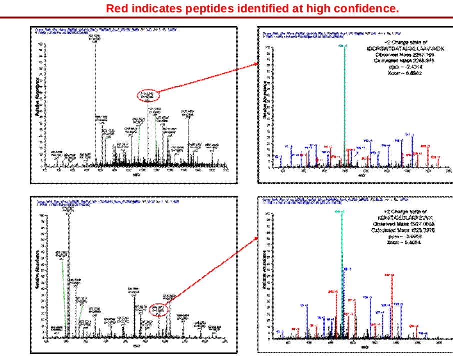
Figure 6

SAR11|gi|71083499|ref|YP_266218.1

Hypothetical protein SAR11_0797 [Candidatus Pelagibacter ubique HTCC1062]
PI: 6.23
MW: 35140

MKIIKTTLVAGLISLFATFSTFAEKVKIGDPGWTGATAIANLLAAVVNDKMGGEAELVPG
NNTAIYAAIDRSKGEIEVHPDIWLPNQQAYTNDLVPKGTLKLSSKPYEGNQGYCVSQQFA
KKMNITAIEDLARPEVVKAMDSDGNGKGEFWIGADGWASANVNQVKLRDYGLYDAGIEAI
RAAEAVKNARVLDSIKKGEGYAFYCYKPHAIWGMADVVMLTEPKFDEAKYKMVQPKEDAD
WYKKSYVASKDALKQIQIGWGTSLEAKSPAIVDFFKNFQLSADDVSLMAFQISVEKKDPA
DVARTWMKNNKSKVDGWLGL

Red indicates peptides identified at high confidence.



High Mass Accuracy for Environmental Proteomics with LTQ -ORBITRAP

- There are a number of grand technical challenges facing the full development of shotgun proteomics for the analyses of complex microbial communities from natural environments.
- One of the biggest challenges is the lack of full metagenomic data for any given sample. Very often datasets will be compared to reference metagenomes. Thus, there is the chance for a large number of measured peptides of high quality that will not match database entries. This can increase the rate of false positives.
- The use of the high resolution Orbitrap provides high mass accuracy MS and MS/MS scans. By just taking high resolution full MS scans (Figure 6), the duty cycle is not dramatically effected (the same number of proteins can be identified, as illustrated in the LTQ Table 1).
- Figure 7 illustrates the distribution of peptides Xcorrs from SEQUEST vs. PPM error (data is filtered at conservative filters). Clearly, most peptides align between 10 and -10 PPM.
- We compared the number of forward hits and reverse hits for two LTQ-Orbitrap runs (47mg) at both liberal and conservative Xcorr filters and at three different PPM categories. The PPM categories were broken down into between 10 and -10 PPM, over 10 and -10 PPM, and undefined parent mass (Orbitrap full scan was not resolved but good MS/MS spectra were obtained that was matched by SEQUEST (Table 3 and Table 4).

Figure 7

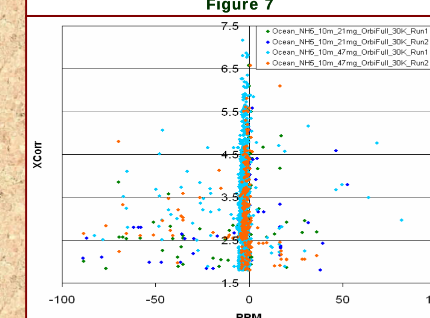


Table 3

Ocean_NH5_10m_21mg_OrbiFull_30K_Run1		
PPM Categories	# of Peptides	% of Whole
Between 10 and -10	340	67.6
Above 10 and -10	127	25.2
Undefined Parent Mass	36	7.2
Totals	503	

Ocean_NH5_10m_47mg_OrbiFull_30K_Run1		
PPM Categories	# of Peptides	% of Whole
Between 10 and -10	1446	79.2
Above 10 and -10	219	12.0
Undefined Parent Mass	160	8.8
Totals	1825	

Ocean_NH5_10m_21mg_OrbiFull_30K_Run2		
PPM Categories	# of Peptides	% of Whole
Between 10 and -10	347	74.0
Above 10 and -10	88	18.8
Undefined Parent Mass	34	7.2
Totals	469	

Ocean_NH5_10m_47mg_OrbiFull_30K_Run2		
PPM Categories	# of Peptides	% of Whole
Between 10 and -10	1258	76.8
Above 10 and -10	221	13.5
Undefined Parent Mass	160	9.8
Totals	1639	

Table 4

Conservative Filters 47mg Run1	Forward Hits	Reverse Hits	% of False Positive Peptides
# of hits Between 10 and -10 PPM	1333	5	0.75
# of hits Above 10 and -10 PPM	167	50	46.08
# of hits with Undefined Parent Mass	143	5	6.76
Conservative Filters 47mg Run2	**Forward Hits**	**Reverse Hits**	**% of False Positive Peptides**
# of hits Between 10 and -10 PPM	1239	1	0.16
# of hits Above 10 and -10 PPM	178	32	30.48
# of hits with Undefined Parent Mass	152	4	5.13
Liberal Filters 47mg Run1	**Forward Hits**	**Reverse Hits**	**% of False Positive Peptides**
# of hits Between 10 and -10 PPM	1983	77	7.48
# of hits Above 10 and -10 PPM	1202	870	83.98
# of hits with Undefined Parent Mass	459	194	59.42
Liberal Filters 47mg Run2	**Forward Hits**	**Reverse Hits**	**% of False Positive Peptides**
# of hits Between 10 and -10 PPM	1710	55	6.23
# of hits Above 10 and -10 PPM	1057	805	86.47
# of hits with Undefined Parent Mass	431	214	66.36

CONCLUSIONS

- We have shown that "shotgun" proteomics can be applied to the complex microbial communities in ocean water.
- When a dominant species is present with a reference genome, moderate depth in proteome coverage can be achieved.
- Large numbers of quality MS/MS spectra are acquired for which no correct database entries exist. Clearly, metagenome data from the same samples would greatly enhance the depth of proteome coverage at high confidence.
- The use of the LTQ-Orbitrap with high resolution and high mass accuracy is essential for these types of sample sets. The acquisition of high mass accuracy full scan spectra greatly improves the confidence of peptide identifications (lowers false positives) and allows for the use of more liberal filter levels (lowering false negatives).
- Future studies should focus on ocean samples collected during diatom blooms which be primarily dominated by the OM43 clade with the MB2181 genus.

ACKNOWLEDGMENTS

- This research sponsored by U.S. Department of Energy under contract DE-AC05-00OR22725 with Oak Ridge National Laboratory, managed and operated by UT-Battelle, LLC.
- 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. F. Larimer and M. Land of the ORNL Genome Analysis and System Modeling Group for computational resources for proteomic analysis.