Computationally Intensive Oceanic Microbial Community High Throughput Proteomics

Angela D. Norbeck[1], Christopher Oehmen[1], Matthew E. Monroe[1], Carrie D. Nicora[1], Ashoka Polpitiya[1], Heather M. Brewer[1], Sarah M. Sowell[2], Lea Constan[3], Stephen J. Giovannoni[2], Steven J. Hallam[3], Ljiljana Paša-Tolić[1], Mary S. Lipton[1], Richard D. Smith[1]

[1]Pacific Northwest National Laboratory, Richland, WA; [2]Oregon State University, Corvallis, OR; [3]University of British Columbia, Vancouver B.C., Canada

Pacific Northwest NATIONAL LABORATORY

Overview

- Proteomics data analysis of mixed (complex) organism samples such as microbial communities is challenging, due to:
  - Increased ambiguity and complexity in the peptide-to-protein relationship due to the presence of multiple variants of similar proteins across different species
  - Long processing times, data storage of GB-sized files, and related data management
- We developed a protein sequence similarity and clustering approach to address these issues that utilizes high-performance computing hardware and parallel processing algorithms.

Introduction

Two sites, two projects

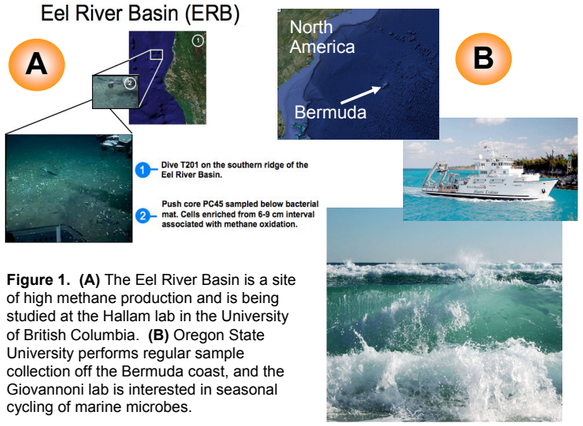


Figure 1. **(A)** The Eel River Basin is a site of high methane production and is being studied at the Hallam lab in the University of British Columbia. **(B)** Oregon State University performs regular sample collection off the Bermuda coast, and the Giovannoni lab is interested in seasonal cycling of marine microbes.

Methods

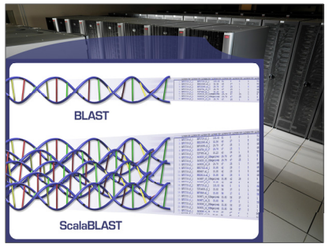
1) Microbial populations were extracted, lysed, enzymatically digested, and analyzed by LC-MS/MS
2) Community proteome files were generated from either the genome sequences from ocean sequencing efforts (GOS database at CAMERA) or a composite of shotgun genome sequences and fosmid reads for sediment communities
3) Proteins in proteome files were clustered into groups by either e value from ScalaBLAST[1] output or from assembly prediction algorithms
4) Peptides and proteins identified using SEQUEST or InsPecT[2] were mapped to the protein groups

Table 1. Proteome files used for matching MS/MS spectra to peptide sequences

Proteome	Protein Entries	Amino Acids
Open Ocean Microbial Community	6,115,750	1,226,044,980
Sea Sediment Microbial Community	171,973	28,386,466

Protein clustering

The sea sediment proteome file was condensed by combining entries with exact sequence duplication before being submitted for a ScalaBLAST[2] sequence comparison on a supercomputer of all proteins against all other proteins to determine similarity. Searches used 700+ cpu processors.



Open ocean community file

» clustered and assembled by JCVI[3]

» downloaded via ftp from CAMERA[4] (http://camera.calit2.net/)

Sea sediment community file

» clusters generated using BLAST e-value threshold and single linkage clustering

» performed using a VB.NET script

A – B
C – K Group 1 = A, B, E
B – E Group 2 = C, D, F, K
F – K
D – K

Single linkage clustering involved grouping member pairs sequentially

Results

Determining peptide confidence

SEQUEST identifications for the sediment community were modeled to determine confidence. InsPecT calculates a p-value for each identification.

Table 2. For both communities, the number of confident peptide and protein identifications fell into a small number of protein clusters. Two or more peptides were required for a protein identification.

Community	ClusterCount	ProteinCount	PeptideSequenceCount
Sea Sediments	1757	2946	2813
Open Ocean	16,446	297,364	290,437

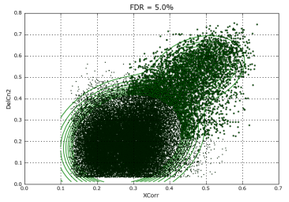


Figure 2. The false discovery of each identified peptide was assessed by applying a multivariate Gaussian distribution model to the 2-dimensional representation of SEQUEST scores[5]. Subsequent modeling allowed FDR thresholds to be set between 1 and 5%.

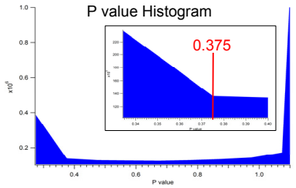


Figure 3. A maximum p-value of 0.375 resulted in 1 reversed hit from an InsPecT search using a combined forward and reversed protein database, indicating an FDR less than 0.01%.

Function group identification

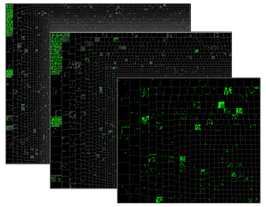


Figure 4. Organizational mapping of protein clusters by peptide coverage highlighted pathways that shared common proteins. The dominant pathway expressed by the sea sediment microbial community was methanogenic processing, and high coverage was observed confidently across those protein clusters.

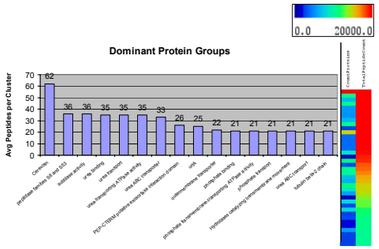


Figure 5. The dominant processes expressed by open ocean microbial populations were energy metabolism and active membrane transport, and these were confined to a small number of protein clusters, even though the proteome file used for searching contained over 6 million entries.

Conclusions

- Methods were developed to enable the processing of large metaproteomics datasets
  - Utilized high performance computing and reduced time of analysis from several weeks to hours
  - Allowed for characterization of dominant functional pathways from ocean surface waters and sea sediment samples
- Statistical relevance was determined for peptide and protein identification from large composite protein databases (over 6 million ORFs)
- Evidence indicates that the dominant protein pathways expressed in marine sediments involved methanogeneis function
- Protein pathways dominant in open ocean water were involved in the active transport of nutrients across membranes and energy metabolism
- Biological conclusions consistent with previous studies[6]

Acknowledgements

Samples were analyzed using capabilities developed under the support of the NIH National Center for Research Resources (RR18522) and the U.S. Department of Energy Biological and Environmental Research (DOE/BER). Significant portions of the work were performed in the Environmental Molecular Science Laboratory, a DOE/BER national scientific user facility at Pacific Northwest National Laboratory (PNNL) in Richland, Washington. PNNL is operated for the DOE by Battelle under contract DE-AC05-76RLO-1830.

References

1. Oehmen et al, "ScalaBLAST: A Scalable Implementation of BLAST for High-Performance Data-Intensive Bioinformatics Analysis," IEEE: Transactions on Parallel and Distributed Systems 2006,17:740-749.
2. Tanner et al., "InsPecT: Fast and accurate identification of post-translationally modified peptides from tandem mass spectra." Anal Chem 2005, 77:4626-4639.
3. Yooseph et al, "The Sorcerer II Global Ocean Sampling expedition: Expanding the universe of protein families," PLoS Biol 2007, 5:432-466.
4. CAMERA: Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis at California Institute for Telecommunications and Information Technology (CalIT) http://camera.calit2.net/
5. López-Ferrer et al. "Statistical Model for Large-Scale Peptide Identification in Databases from Tandem Mass Spectra Using SEQUEST," Anal Chem 2004, 76:6853-6860.
6. Sowell et al, "Proteomic Analysis of Stationary Phase in the Marine Bacterium Candidatus Pelagibacter ubique", Appl Environ Microb 2008, 74: 4091-4100.

CONTACT: Angela D. Norbeck
Biological Sciences Division, K8-98
Pacific Northwest National Laboratory
P.O. Box 999, Richland, WA 99352
E-mail: angela.norbeck@pnl.gov