Abstract
E. Kolker, BIATECH
“Interdisciplinary Study of Shewanella oneidensis MR-1’s Metabolism & Metal Reduction”

Since our project became part of the Shewanella Federation, we focused our work mostly on analysis of different types of data produced by global high-throughput technologies to characterize gene and protein expression as well as getting a better understanding of the cellular metabolism. Specifically, first year activities include development of:

- new labeling technique for quantitative proteomics, so called methyl esterification labeling approach, complementary to currently available methods;
- new algorithm for de novo protein sequencing;
- new statistical model for spectral analysis of arbitrary shape data;
- one of the first analyses of the transcriptome of the entire microorganism;
- new approach to predict operon structures and transcripts within untranslated regions;
- the first protein experimental mixtures with known physico-chemical characteristics for high-throughput proteomics experiments;
- the first statistical models for peptide and protein identifications for high-throughput proteomics analysis;
- Shewanella metabolic capability experiments with minimal media on aerobically & anaerobically grown cells and transformation experiments.

Several collaborations have been established within the Shewanella Federation with PNNL, USC, ORNL, and MSU. The first year of this project, supported by DOE’s Offices of Biological and Environmental Research and Advanced Scientific Computing Research, also resulted in 6 published papers.

This is a joint work of A. Keller, A. Nesvizhskii, A. Picone, B. Tjaden, D. Goodlett, S. Purvine, S. Stolyar, and T. Cherny done at BIATECH and ISB.

Eugene Kolker, PhD
President & Director              Editor-in-Chief
BIATECH                            OMICS A Journal of
nonprofit research center          Integrative Biology
BIATECH (Kolker et al.)

- Sequence and data analysis
- Statistical models
- Quality assessments for HT analyses
Proteins

Sample

Proteins

Peptides

Peptide Probabilities

Peptide 1 0.999
Peptide 2 0.500
Peptide 3 0.750
Peptide 4 0.001

Trypsin

Step 1

Step 2

Step 3

Step 4

Step 5

Step 6

Good Spectra

Bad Spectra

MS/MS Database Search Results

High Confidence Protein Identifications

BIATECH
Protein-Protein Interaction Maps

high-throughput mass spectrometric protein complex identification approach

940,000 MS/MS spectra
35,000 peptide identifications
8,118 potential interactions


No attempt to estimate confidence levels of protein identifications

BIATECH
MS/MS Data Analysis

- Thousands of spectra from each experiment, but much of the data are of low quality
- Correct peptide identification or false positive? Requires decision from a human analyst

Quality Assessment of MS is needed
Quality Assessment of MS/MS Spectra

Good Spectrum

Bad Spectrum
Spectrum Quality Clustering

Discriminant Function 1

Discriminant Function 2

Quality
- group centroids
- 4 - very good
- 3 - good
- 2 - bad
- 1 - very bad

Training set (manually assigned quality): ~ 1,000 spectra, HI, LC/MS/MS
How to Mimic Complex Samples & Develop Statistical Models of Peptide & Protein Identifications?

- **20** selected, purified proteins
- Different concentrations
- **1,000:1** dynamic range
- Different database searches

*To Build Statistical Models*
<table>
<thead>
<tr>
<th>Protein Name</th>
<th>MW(Daltons)</th>
<th>Conc.(nM)</th>
<th>CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Bovine beta-casein</td>
<td>25,107</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>2. Bovine carbonic anhydrase</td>
<td>28,980</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>3. Bovine cytochrome c</td>
<td>11,572</td>
<td>40</td>
<td>63</td>
</tr>
<tr>
<td>4. Bovine beta-lactoglobulin</td>
<td>19,883</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td>5. Bovine alpha-lactalbumin</td>
<td>16,246</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>6. Bovine serum albumin</td>
<td>69,293</td>
<td>40</td>
<td>100</td>
</tr>
<tr>
<td>7. Chicken ovalbumin</td>
<td>42,750</td>
<td>0.4</td>
<td>0</td>
</tr>
<tr>
<td>8. Bovine serotransferrin</td>
<td>77,753</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>9. Rabbit GAPDH</td>
<td>35,688</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>10. Rabbit glycogen phosphorylase</td>
<td>97,158</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>11. EC beta-galactosidase</td>
<td>116,351</td>
<td>0.4</td>
<td>18</td>
</tr>
<tr>
<td>12. Bovine gamma-actin</td>
<td>41,661</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td>13. Bovine catalase</td>
<td>57,585</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>14. Rabbit myosin</td>
<td>241,852</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td>15. EC alkaline phosphatase</td>
<td>49,438</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td>16. Horse myoglobin</td>
<td>16,951</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>17. B. lich. alpha amylase</td>
<td>66,924</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>18. S. cer. mannose-6-phosphate isomerase</td>
<td>48,057</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Filtered and Unfiltered Distributions

DB search score

number of spectra

[M+2H]^{2+}

incorrect

unfiltered

filtered

correct

x 10

BIATECH
EM Learns Search Score Distributions

Incorrect Peptide Assignments

Correct Peptide Assignments

No. of spectra

Database search score

BIATECH
**EM Iteration 8**

Incorrect Peptide Assignments

Correct Peptide Assignments

Database search score
Accuracy of Probability Model: Test Set

Test data: SEQUEST results of known validity; ~36,000 MS/MS spectra generated from the control protein mixture (22 LC/MS/MS runs)
Discriminating Power of Computed Probabilities: Test Data Set

Sensitivity: fraction of all correct results passing filter.

Error: fraction of all results passing filter that are incorrect

Ideal Spot
Proteins

Peptide Probabilities

Peptide 1 0.999
Peptide 2 0.500
Peptide 3 0.750
Peptide 4 0.001

Sample

Proteins

trypsin

Step 1

Peptides

Step 6

High Confidence Protein Identifications

Bad Spectra

Step 2

Good Spectra

Step 3

Step 4

Step 5

MS/MS Database Search Results

BIATECH
Future Needs - SF

- Data Integration
- Modeling
- WGA

- Sequencing of Shewanella strains
- Controls, standards, and quality assessments for sample preps, HT and data analyses
- GtL coordination/updates

ATTENTION:

OMICS J Integr Biol: Integrative Microbiology, 2003 (GtL) issue
and ASM, May 2003, DC: Systems Microbiology
**Shewanella oneidensis**

*oneidensis* is the abbreviated name of the bacterium *Shewanella oneidensis*, which according to the definitive text, which categorizes bacteria *Bergey's Manual*, belongs to the gram negative gamma-subgroup (as *E. coli* and *H. influenzae*) Alteromonadales, genus XII *Shewanella*.

The name *oneidensis* comes from the name of Lake Oneida where from our collaborator and friend Ken Nealson first isolated and characterized *S. oneidensis* fifteen years ago. *S. oneidensis* is at the very top of the priority list of the US Department of Energy, because of its unique ability to reduce heavy metals like uranium, degrade organic wastes, and sequester a range of toxic metals. Environments in places like Hanford or Chernobyl can be significantly improved if we would understand *Shewanella* better.

We are still not there...

**More Information on *S. oneidensis***:
- *DOE's information on *S. oneidensis*
- *DOE’s Genomes to Life*
- *Shewanella Federation Web site*
- *Shewanella Genome Annotation (02/03/03)*

**PIs of Shewanella Federation**:
Eugene Kolker, BIATECH
James Fredrickson, Pacific Northwest National Laboratory
James Tiedje, Michigan State University
Jizhong Zhou, Oak Ridge National Laboratory
Kenneath Nealson, University of Southern California