Creation of Gene Database for Microarray Analysis of *Shewanella oneidensis* MR-1

Mary Alverson, Josh Kuroda, Ron Legaspi, Emily Simso

BIOL/CMSI 367 - Biological Databases
Loyola Marymount University
December 15, 2015

Overview

- Background of *Shewanella oneidensis*
- Results of Gene Database creation
    - Gene Database Schema Figure
    - Summary of how many OrderedLocusNames were found
    - Match command and PGAdmin III Query
    - Quantity and identity of gene IDs that weren't exported
    - Code was changed to handle ID's without underscores
- DNA Microarray paper
- Sanity check showed that F60 and C60 had most relevant results
- Results of DNA microarray analysis
    - Iron acquisition, anaerobic energy metabolism, protein degradation
- Non-redundant GO terms
- MAPPFinder Results of Increased and Decreased Criterion
- GenMAPP MAPP Relevant to Results

Overview

- **Background of *Shewanella oneidensis***
- Results of Gene Database creation
    - Gene Database Schema Figure
    - Summary of how many OrderedLocusNames were found
    - Match command and PGAdmin III Query
    - Quantity and identity of gene IDs that weren't exported
    - Code was changed to handle ID's without underscores
- DNA Microarray paper
- Sanity check showed that F60 and C60 had most relevant results
- Results of DNA microarray analysis
    - Iron acquisition, anaerobic energy metabolism, protein degradation
- Non-redundant GO terms
- MAPPFinder Results of Increased and Decreased Criterion
- GenMAPP MAPP Relevant to Results

Shewanella oneidensis is Useful for Bioremediation

- *S. oneidensis* might prove useful for direct bioremediation of both metal and organic pollutants under anaerobic conditions.

- *S. oneidensis* can reduce the levels of oxidized metals.

- *S. oneidensis* is infrequently a human pathogen.

- *Shewanella oneidensis* is an important model organism for bioremediation studies because of its diverse respiratory capabilities, conferred in part by multicomponent, branched electron transport systems.

Overview

- Background of *Shewanella oneidensis*
- **Results of Gene Database creation**
    - **Gene Database Schema Figure**
    - Summary of how many OrderedLocusNames were found
    - Match command and PGAdmin III Query
    - Quantity and identity of gene IDs that weren't exported
    - Code was changed to handle ID's without underscores
- DNA Microarray paper
- Sanity check showed that F60 and C60 had most relevant results
- Results of DNA microarray analysis
    - Iron acquisition, anaerobic energy metabolism, protein degradation
- Non-redundant GO terms
- MAPPFinder Results of Increased and Decreased Criterion
- GenMAPP MAPP Relevant to Results

Shewanella oneidensis MR-1 GDB Schema Figure
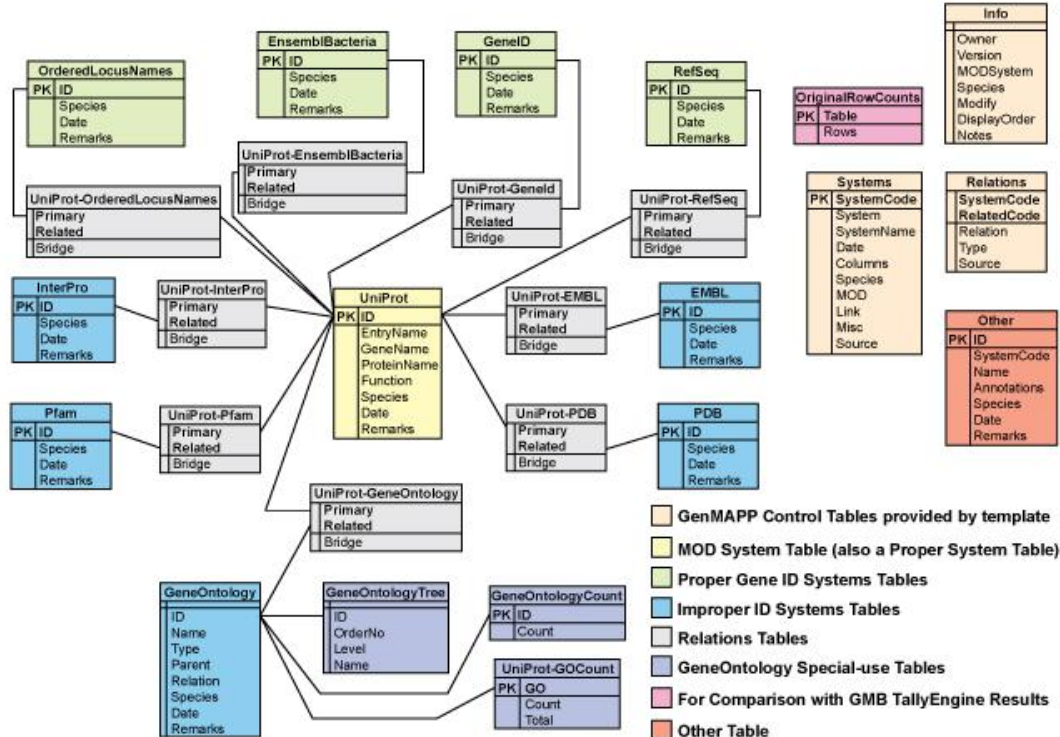
NOTE: Some Relations tables are not shown. All possible pairwise Relations tables exist between Proper ID systems and between Proper and Improper ID systems, but not between Improper ID systems (i.e., Proper-Proper, Proper-Improper, but NOT Improper-Improper).
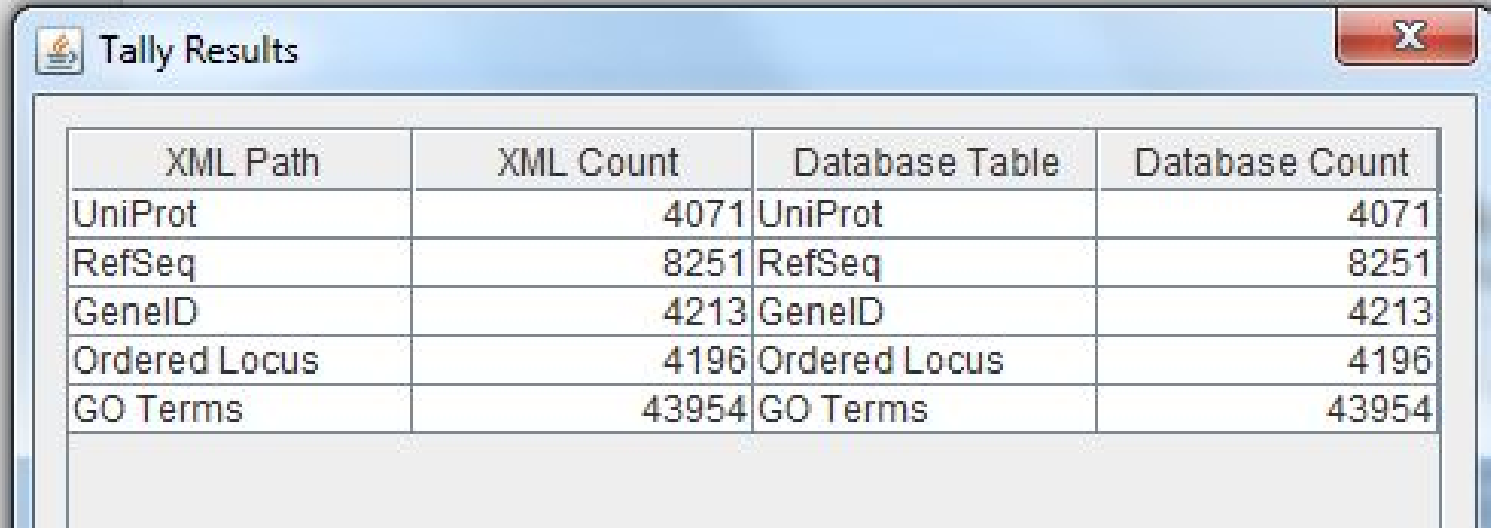
Overview

- Background of *Shewanella oneidensis*
- **Results of Gene Database creation**
  - Gene Database Schema Figure
  - **Summary of how many OrderedLocusNames were found**
  - Match command and PGAdmin III Query
  - Quantity and identity of gene IDs that weren't exported
  - Code was changed to handle ID's without underscores
- DNA Microarray paper
- Sanity check showed that F60 and C60 had most relevant results
- Results of DNA microarray analysis
  - Iron acquisition, anaerobic energy metabolism, protein degradation
- Non-redundant GO terms
- MAPPFinder Results of Increased and Decreased Criterion
- GenMAPP MAPP Relevant to Results

The Expected Number of IDs Were Found

Source	Number of OrderedLocusNames IDs
by XMLPipeDB match in UniProt XML File	**4207**
by TallyEngine in the UniProt XML File	**4196**
by TallyEngine in PostgreSQL database	**4196**
total OrderedLocusNames in the gdb	**8392**
Coding Genes in the MOD	**4196**

Tally Engine Results Show That The Counts Match

XML Path	XML Count	Database Table	Database Count
UniProt	4071	UniProt	4071
RefSeq	8251	RefSeq	8251
GeneID	4213	GeneID	4213
Ordered Locus	4196	Ordered Locus	4196
GO Terms	43954	GO Terms	43954

Overview

- Background of *Shewanella oneidensis*
- **Results of Gene Database creation**
  - Gene Database Schema Figure
  - Summary of how many OrderedLocusNames were found
  - **Match command and PGAdmin III Query**
  - Quantity and identity of gene IDs that weren't exported
  - Code was changed to handle ID's without underscores
- DNA Microarray paper
- Sanity check showed that F60 and C60 had most relevant results
- Results of DNA microarray analysis
  - Iron acquisition, anaerobic energy metabolism, protein degradation
- Non-redundant GO terms
- MAPPFinder Results of Increased and Decreased Criterion
- GenMAPP MAPP Relevant to Results

Match Command & PGAdmin III Query

Match command:

```
java -jar xmlpipedb-match-1.1.1.jar "SO_A?[0-9][0-9][0-9][0-9]" <
    SOneidensisUNIPROT
```

PGAdmin III query:

```
select count(*) from genenametype where type = 'ordered locus' and
    value ~ 'SO_A?[0-9][0-9][0-9][0-9]';
```

Overview

- Background of *Shewanella oneidensis*
- **Results of Gene Database creation**
    - Gene Database Schema Figure
    - Summary of how many OrderedLocusNames were found
    - Match command and PGAdmin III Query
    - **Quantity and identity of gene IDs that weren't exported**
    - Code was changed to handle ID's without underscores
- DNA Microarray paper
- Sanity check showed that F60 and C60 had most relevant results
- Results of DNA microarray analysis
    - Iron acquisition, anaerobic energy metabolism, protein degradation
- Non-redundant GO terms
- MAPPFinder Results of Increased and Decreased Criterion
- GenMAPP MAPP Relevant to Results

ID's That Weren't Exported Are Unnecessary

S0_3699

S0_1312

S0_4269

S0_2875

S0_4532

S0_4580

S0_2662

S0_4423

S0_3156

S0_2967

S0_2024

- There were 11 OrderedLocusNames IDs that were nominally (no gene tag) not in the XML source.

- These 11 OrderedLocusNames IDs were not imported into Postgres because they are all only part of "protein-protein interaction" databases and can be safely left out.

- There were no OrderedLocusNames IDs in PostgreSQL that were not exported to the GenMAPP Gene Database.

Overview

- Background of *Shewanella oneidensis*
- **Results of Gene Database creation**
    - Gene Database Schema Figure
    - Summary of how many OrderedLocusNames were found
    - Match command and PGAdmin III Query
    - Quantity and identity of gene IDs that weren't exported
    - **Code was changed to handle ID's without underscores**
- DNA Microarray paper
- Sanity check showed that F60 and C60 had most relevant results
- Results of DNA microarray analysis
    - Iron acquisition, anaerobic energy metabolism, protein degradation
- Non-redundant GO terms
- MAPPFinder Results of Increased and Decreased Criterion
- GenMAPP MAPP Relevant to Results

# The Code Was Changed to Handle ID's Without Underscores

```java
String newId = null;
for (int i = 0; i < substrings.length; i++) {

    newId = "SO" + substrings[i].substring(3,substrings[i].length());
```

- For every OrderedLocusNames Entry, the entry is duplicated and the underscore is removed.

- This results in the total OrderedLocusNames in the gdb to be twice the number of OrderedLocusNames found elswhere.
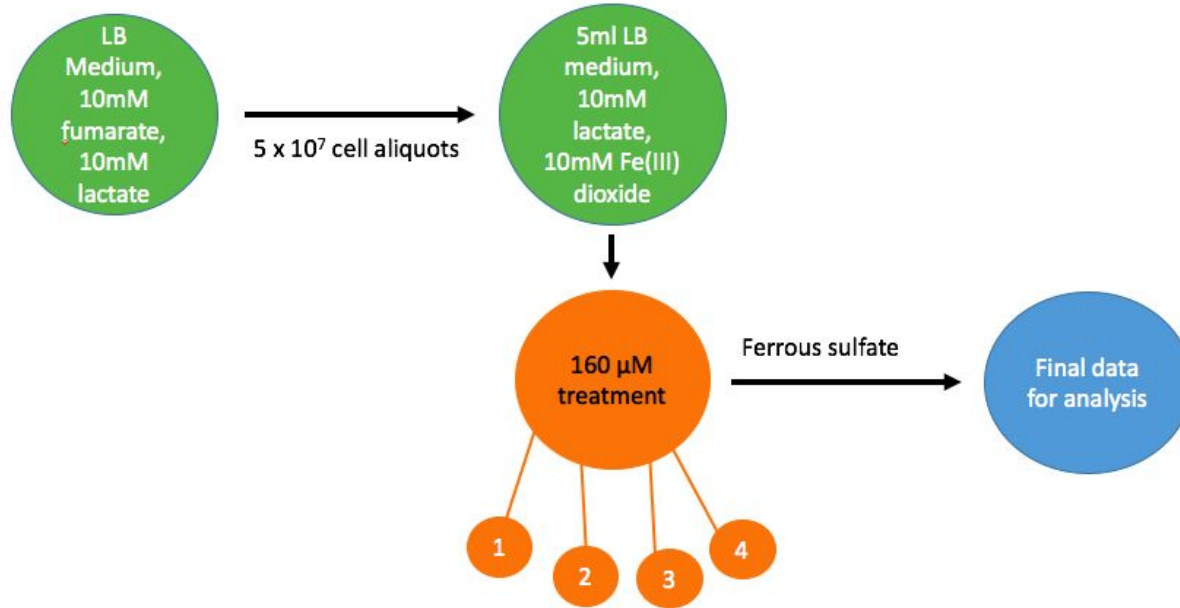
Overview

- Background of *Shewanella oneidensis*
- Results of Gene Database creation
    - Gene Database Schema Figure
    - Summary of how many OrderedLocusNames were found
    - Match command and PGAdmin III Query
    - Quantity and identity of gene IDs that weren't exported
    - Code was changed to handle ID's without underscores
- **DNA Microarray paper**
- Sanity check showed that F60 and C60 had most relevant results
- Results of DNA microarray analysis
    - Iron acquisition, anaerobic energy metabolism, protein degradation
- Non-redundant GO terms
- MAPPFinder Results of Increased and Decreased Criterion
- GenMAPP MAPP Relevant to Results

DNA Microarray Paper Experimental Design
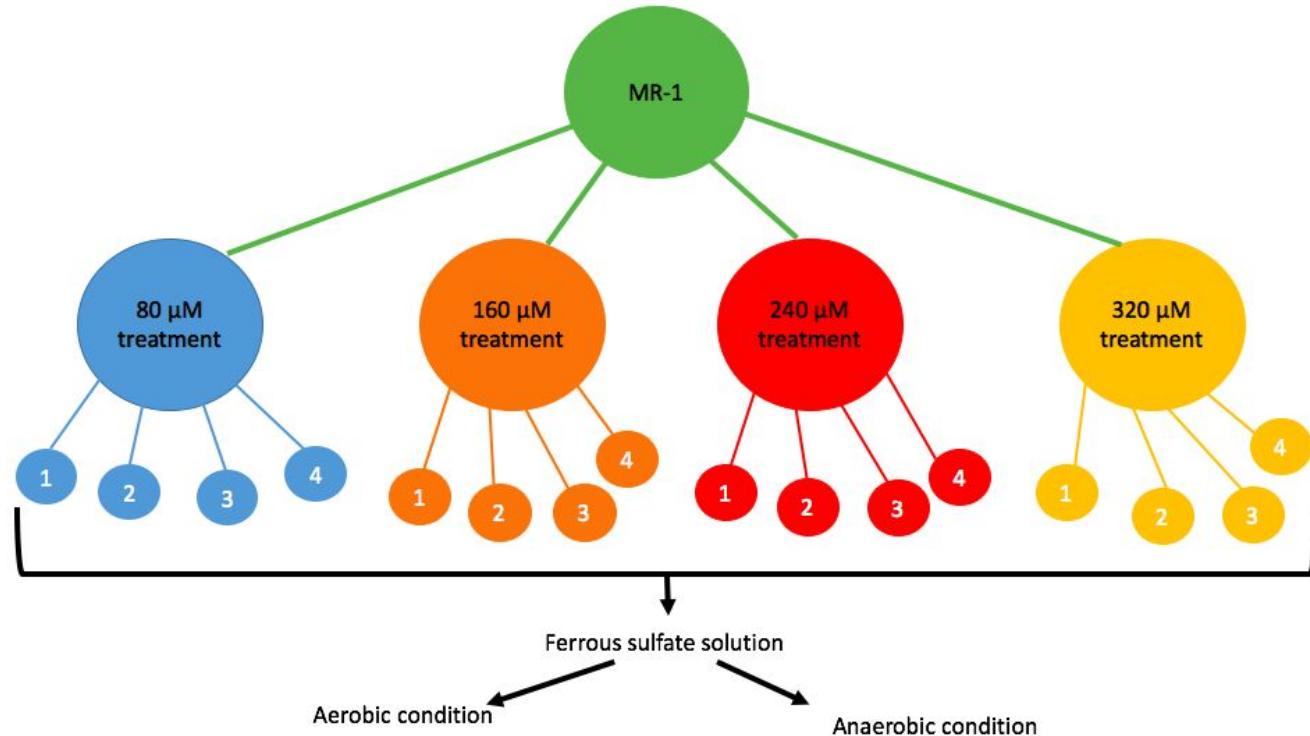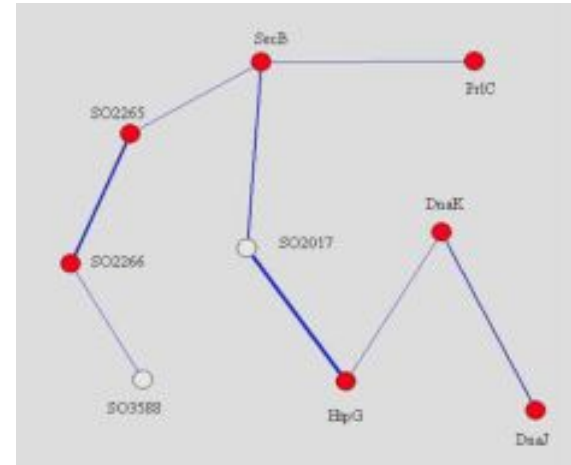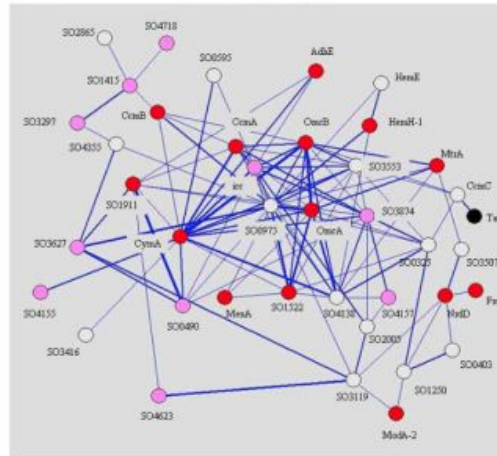
DNA Microarray Paper Iron Depletion in the Anaerobic State

DNA Microarray Paper Showed Variations in Depletion vs Repletion Responses

- Three major modules
  - Iron acquisition, anaerobic energy metabolism, and protein degradation
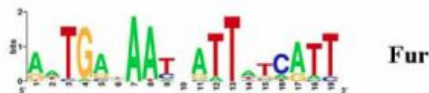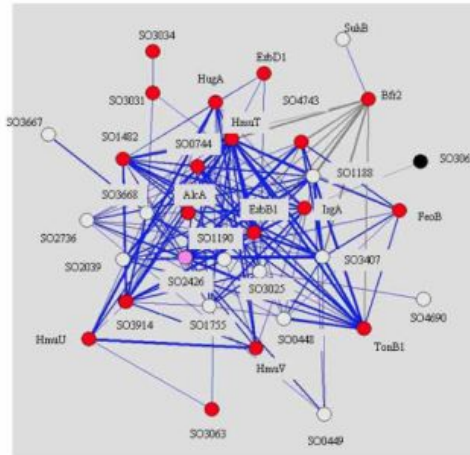
DNA Microarray Paper Analysis of Significant Results

- Anaerobic energy metabolism model
- Lack of oxidative stress
- Rapid iron uptake

- Therefore, our question is: how can we use these explanations to interpret the microarray data using programs such as GenMAPP and MAPPFinder?

Overview

- Background of *Shewanella oneidensis*
- Results of Gene Database creation
  - Gene Database Schema Figure
  - Summary of how many OrderedLocusNames were found
  - Match command and PGAdmin III Query
  - Quantity and identity of gene IDs that weren't exported
  - Code was changed to handle ID's without underscores
- DNA Microarray paper
- **Sanity check showed that F60 and C60 had most relevant results**
- Results of DNA microarray analysis
  - Iron acquisition, anaerobic energy metabolism, protein degradation
- Non-redundant GO terms
- MAPPFinder Results of Increased and Decreased Criterion
- GenMAPP MAPP Relevant to Results

# Sanity Check Showed That F60 and C60 Had The Most Relevant Results

- F60 and C60 had the most genes that were significantly changed.

Comparison	How many genes have a pvalue<0.05 (with percentage)	How many genes have a pvalue<0.01 (with percentage)	How many genes have a pvalue<0.001 (with percentage)	How many genes have a pvalue <0.0001 (with percentage)	How many genes are p<0.05 for the Bonferroni-corrected p value (with percentage)	How many genes are p<0.05 for the Benjamini and Hochberg corrected p value (with percentage)	Genes with an average log fold change greater than zero	Genes with average log fold change less than zero	Genes with average log fold change of >0.25 and p<0.05 (with percentage)	Genes with average log fold change <-0.25 and p<0.05 (with percentage)
C5 and C0	344 genes, 6.36%	94 genes, 1.74%	18 genes, 0.33%	5 genes, 0.09%	2 genes, 0.04%	2 genes, 0.037%	180 genes, 3.33%	164 genes, 3.03%	161 genes, 2.98%	149 genes, 2.76%
C20 and C0	868 genes, 16.05%	342 genes, 6.32%	79 genes, 1.46%	14 genes, 0.26%	1 gene, 0.01%	34 genes, 0.63%	452 genes, 8.36%	416 genes	437 genes, 7.69%	405 genes, 7.49%
C60 and C0	1017 genes, 18.81%	471 genes, 8.71%	163 genes, 3.01%	53 genes, 0.98%	13 genes, 0.24%	229 genes, 4.23%	487 genes, 9.01%	530 genes, 9.80%	475 genes, 8.78%	513 genes, 9.49%
F5 and C60	969 genes, 17.95%	315 genes, 5.82%	40 genes, 0.74%	7 genes, 0.13%	1 gene, 0.01%	4 genes, 0.07%	479 genes, 8.86%	490 genes, 9.06%	441 genes, 8.15%	431 genes, 7.97%
F20 and C60	1838 genes, 33.99%	892 genes, 16.49%	239 genes, 4.42%	54 genes, 1.00%	10 genes, 0.18%	707 genes, 13.07%	826 genes, 15.27%	1012 genes, 18.71%	788 genes, 14.57%	963 genes, 17.81%
F60 and C60	2070 genes, 38.28%	1140 genes, 21.08%	387 genes, 7.16%	120 genes, 2.22%	33 genes, 0.61%	1193 genes, 22.06%	870 genes, 16.09%	1200 genes, 22.19%	828 genes, 15.31%	1146 genes, 21.19%

Overview

- Background of *Shewanella oneidensis*
- Results of Gene Database creation
    - Gene Database Schema Figure
    - Summary of how many OrderedLocusNames were found
    - Match command and PGAdmin III Query
    - Quantity and identity of gene IDs that weren't exported
    - Code was changed to handle ID's without underscores
- DNA Microarray paper
- Sanity check showed that F60 and C60 had most relevant results
- **Results of DNA microarray analysis**
    - **Iron acquisition, anaerobic energy metabolism, protein degradation**
- Non-redundant GO terms
- MAPPFinder Results of Increased and Decreased Criterion
- GenMAPP MAPP Relevant to Results

Results of DNA Microanalysis - Iron acquisition
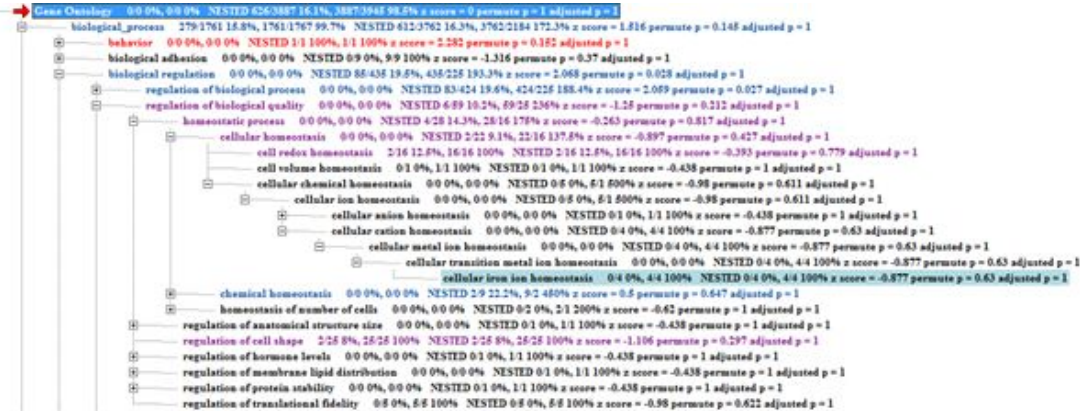
SO1111
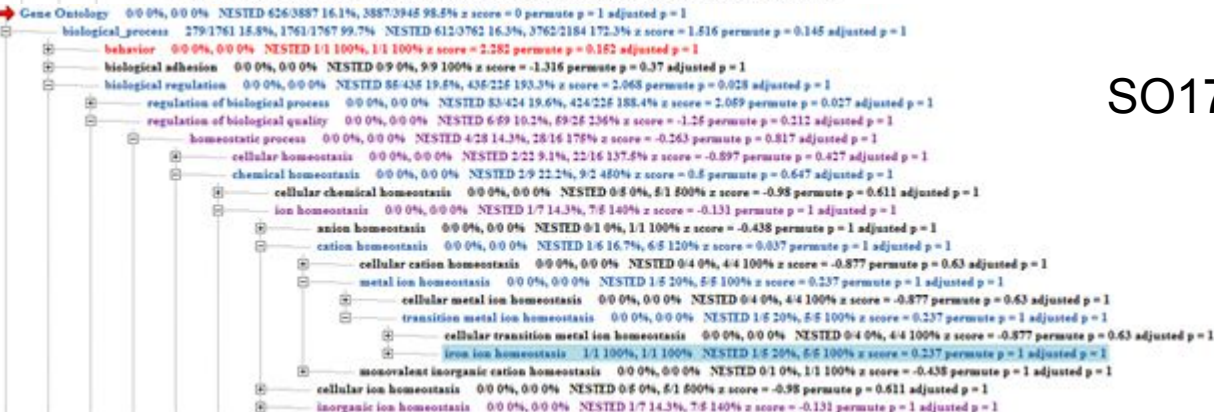
SO1784

Results of DNA Microanalysis - Iron acquisition

- Increased in acquisition (5.61)
- SO1111: bacterioferritin subunit

- Cellular iron ion homeostasis
- Iron ion transport
- Oxidation-reduction process
- Cellular component → cell
- Metal ion binding
- Ferric iron binding

- Decreased in iron acquisition (0.14)
- SO1784: ferrous iron transport protein

- Iron ion homeostasis
- Transport
- Ferrous iron transport
- Ferrous iron transmembrane transport
- Ferrous iron transmembrane transporter activity

Results of DNA Microanalysis - Anaerobic energy metabolism
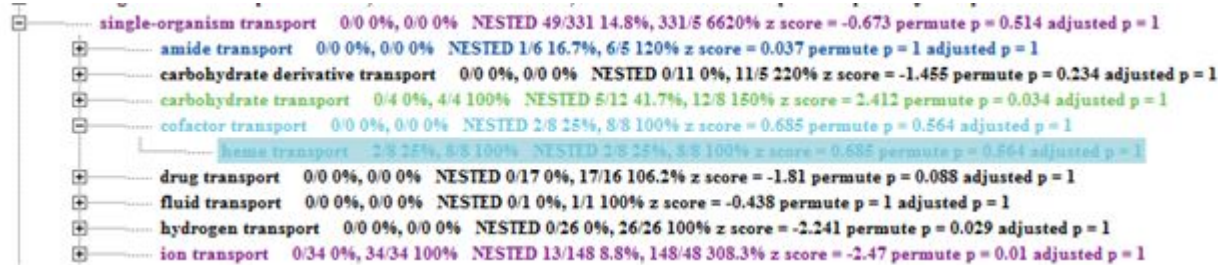
SO0261

SO0262

Results of DNA Microanalysis - Anaerobic energy metabolism

- Increased during anaerobic energy metabolism (2.57)
- SO0261: heme exporter protein

- Heme transport
- Heme binding
- Heme transporter activity

- Not a part of the pathway for repletion in paper (was found to be in depletion)
- SO0262: heme exporter protein

- Cytochrome complex assembly
- Establishment of localization → transport
- Membrane
- Integral component of membrane
- Plasma membrane
- Heme transport activity

Results of DNA Microanalysis - Protein degradation
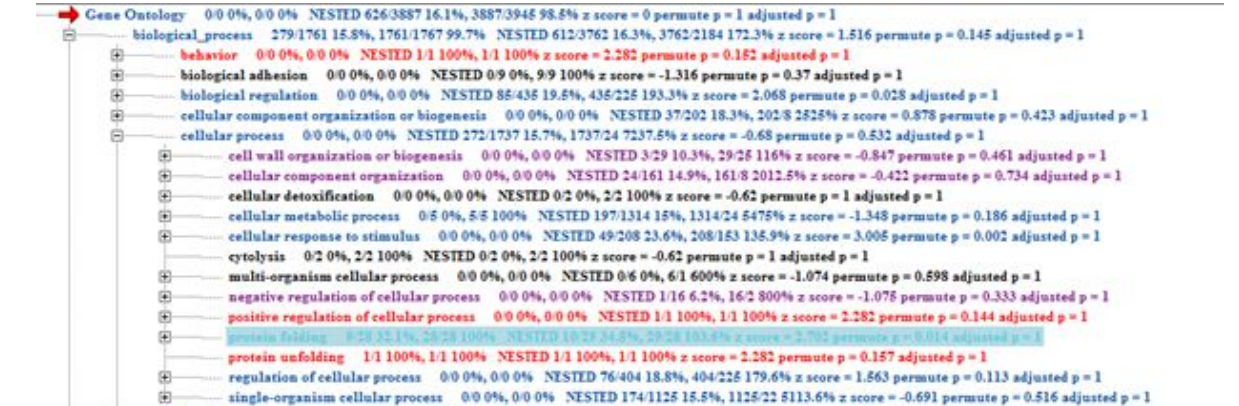
SO2016
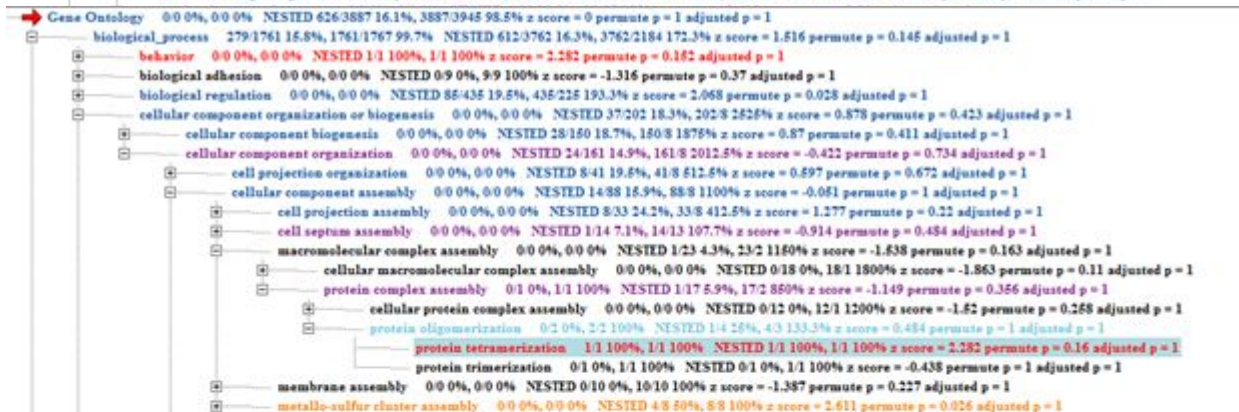
SO0052

Results of DNA Microanalysis - Protein degradation

- Decreased (0.06)
- SO2016: heme shock protein

- Protein folding
- Response to stress
- Nucleotide binding
- ATP binding
- Unfolded protein binding

- Not involved (was present in depletion)
- SO0052: protein export chaperone

- Protein tetramerization
- Protein folding
- Transport
- Protein transport
- Unfolded protein binding

Overview

- Background of *Shewanella oneidensis*
- Results of Gene Database creation
    - Gene Database Schema Figure
    - Summary of how many OrderedLocusNames were found
    - Match command and PGAdmin III Query
    - Quantity and identity of gene IDs that weren't exported
    - Code was changed to handle ID's without underscores
- DNA Microarray paper
- Sanity check showed that F60 and C60 had most relevant results
- Results of DNA microarray analysis
    - Iron acquisition, anaerobic energy metabolism, protein degradation
- **Non-redundant GO terms**
- MAPPFinder Results of Increased and Decreased Criterion
- GenMAPP MAPP Relevant to Results

Non-redundant GO Terms

Gene Ontology Results

purine nucleotide binding   NESTED 87/347 25.1%, 347/309 112.3% z score = 4.761 permute p = 0 adjusted p = 0.029

fatty acid catabolic process   NESTED 7/10 70%, 10/7 142.9% z score = 4.642 permute p = 0 adjusted p = 0.038

nonribosomal peptide biosynthetic process   NESTED 4/4 100%, 4/2 200% z score = 4.567 permute p = 0 adjusted p = 0.103

monocarboxylic acid catabolic process   NESTED 7/12 58.3%, 12/1 1200% z score = 3.985 permute p = 0 adjusted p = 0.226

small molecule binding   NESTED 109/495 22%, 495/361 137.1% z score = 3.832 permute p = 0 adjusted p = 0.664

phosphorelay signal transduction system   NESTED 32/116 27.6%, 116/117 99.1% z score = 3.415 permute p = 0 adjusted p = 0.767

intracellular signal transduction   NESTED 32/118 27.1%, 118/119 99.2% z score = 3.305 permute p = 0 adjusted p = 0.834

receptor signaling protein activity   NESTED 18/57 31.6%, 57/50 114% z score = 3.201 permute p = 0 adjusted p = 1

cell communication   NESTED 40/165 24.2%, 165/153 107.8% z score = 2.906 permute p = 0 adjusted p = 1

intracellular organelle   NESTED 5/115 4.3%, 115/58 198.3% z score = -3.482 permute p = 0 adjusted p = 0.738

intracellular non-membrane-bounded organelle   NESTED 3/110 2.7%, 110/58 189.7% z score = -3.872 permute p = 0 adjusted p = 0.662

purine ribonucleotide binding   NESTED 86/346 24.9%, 346/309 112% z score = 4.639 permute p = 0.001 adjusted p = 0.039

short-chain fatty acid metabolic process   NESTED 4/4 100%, 4/3 133.3% z score = 4.567 permute p = 0.001 adjusted p = 0.103

propionate metabolic process   NESTED 4/4 100%, 4/3 133.3% z score = 4.567 permute p = 0.001 adjusted p = 0.103

adenyl nucleotide binding   NESTED 78/310 25.2%, 310/309 100.3% z score = 4.521 permute p = 0.001 adjusted p = 0.115

purine ribonucleoside binding   NESTED 84/342 24.6%, 342/309 110.7% z score = 4.454 permute p = 0.001 adjusted p = 0.143

purine ribonucleoside triphosphate binding   NESTED 84/342 24.6%, 342/309 110.7% z score = 4.454 permute p = 0.001 adjusted p = 0.143

purine nucleoside binding   NESTED 84/342 24.6%, 342/309 110.7% z score = 4.454 permute p = 0.001 adjusted p = 0.143

nucleoside binding   NESTED 84/343 24.5%, 343/309 111% z score = 4.424 permute p = 0.001 adjusted p = 0.143

ribonucleoside binding   NESTED 84/343 24.5%, 343/310 110.6% z score = 4.424 permute p = 0.001 adjusted p = 0.143

Clicking on a specific term will locate that term in the hierarchy.

To sort, filter, or print this list, you can open the .txt files in a spreadsheet program such as Microsoft Excel. The files "yourfilename-GO.txt" and "yourfilename-Local.txt" are tab-delimited files containing all of the data you see here.

Overview

- Background of *Shewanella oneidensis*
- Results of Gene Database creation
    - Gene Database Schema Figure
    - Summary of how many OrderedLocusNames were found
    - Match command and PGAdmin III Query
    - Quantity and identity of gene IDs that weren't exported
    - Code was changed to handle ID's without underscores
- DNA Microarray paper
- Sanity check showed that F60 and C60 had most relevant results
- Results of DNA microarray analysis
    - Iron acquisition, anaerobic energy metabolism, protein degradation
- Non-redundant GO terms
- **MAPPFinder Results of Increased and Decreased Criterion**
- GenMAPP MAPP Relevant to Results

MAPP Finder Results of Increased Criterion

GOID	GO Name	Number Changed	Percent Changed	Z Score	PermuteP
5840	ribosome	21	36.2069	4.8	0
30529	ribonucleoprotein complex	21	35.59322	4.706	0
3735	structural constituent of ribosome	19	35.18518	4.413	0
5198	structural molecule activity	20	30.76923	3.823	0
9156	ribonucleoside monophosphate biosynthetic process	13	33.33333	3.411	0.001
190260 0	hydrogen ion transmembrane transport	9	37.5	3.255	0.002
30151	molybdenum ion binding	7	46.66667	3.586	0.003
9124	nucleoside monophosphate biosynthetic process	14	32.55814	3.438	0.003
16836	hydro-lyase activity	11	32.35294	3.019	0.003
44205	'de novo' UMP biosynthetic process	4	66.66666	3.666	0.004
9311	oligosaccharide metabolic process	6	46.15385	3.285	0.004
15078	hydrogen ion transmembrane transporter activity	9	36	3.108	0.004
6818	hydrogen transport	9	34.61538	2.968	0.004
15992	proton transport	9	34.61538	2.968	0.004
19843	rRNA binding	12	30.76923	2.951	0.004
15412	molybdate transmembrane-transporting ATPase activity	4	57.14286	3.24	0.008
190127 1	lipooligosaccharide biosynthetic process	5	50	3.228	0.009
9247	glycolipid biosynthetic process	5	50	3.228	0.009
46467	membrane lipid biosynthetic process	5	50	3.228	0.009

MAPPFinder Results for Decreased Criterion

GOID	GO Name	Number Changed	Percent Changed	Z Score	PermuteP
9062	fatty acid catabolic process	7	70	4.642	0
19184	nonribosomal peptide biosynthetic process	4	100	4.567	0
72329	monocarboxylic acid catabolic process	7	58.33333	3.985	0
46459	short-chain fatty acid metabolic process	4	100	4.567	0.001
19541	propionate metabolic process	4	100	4.567	0.001
44242	cellular lipid catabolic process	7	63.63636	4.294	0.001
16042	lipid catabolic process	7	63.63636	4.294	0.001
30163	protein catabolic process	6	66.66666	4.131	0.001
15891	siderophore transport	6	66.66666	4.131	0.001
15688	iron chelate transport	6	66.66666	4.131	0.001
4176	ATP-dependent peptidase activity	4	66.66666	3.372	0.008
5315	inorganic phosphate transmembrane transporter activity	4	57.14286	2.956	0.008
190167	phosphate transmembrane transporter activity	4	57.14286	2.956	0.008
51082	unfolded protein binding	6	50	3.199	0.01
34440	lipid oxidation	4	57.14286	2.956	0.011
6635	fatty acid beta-oxidation	4	57.14286	2.956	0.011
19395	fatty acid oxidation	4	57.14286	2.956	0.011
9401	phosphoenolpyruvate-dependent sugar phosphotransferase system	4	57.14286	2.956	0.014
3684	damaged DNA binding	4	57.14286	2.956	0.02
46915	transition metal ion transmembrane transporter activity	4	50	2.611	0.022
16226	iron-sulfur cluster assembly	4	50	2.611	0.026
3774	motor activity	4	50	2.611	0.026
31163	metallo-sulfur cluster assembly	4	50	2.611	0.026

Overview

- Background of *Shewanella oneidensis*
- Results of Gene Database creation
    - Gene Database Schema Figure
    - Summary of how many OrderedLocusNames were found
    - Match command and PGAdmin III Query
    - Quantity and identity of gene IDs that weren't exported
    - Code was changed to handle ID's without underscores
- DNA Microarray paper
- Sanity check showed that F60 and C60 had most relevant results
- Results of DNA microarray analysis
    - Iron acquisition, anaerobic energy metabolism, protein degradation
- Non-redundant GO terms
- MAPPFinder Results of Increased and Decreased Criterion
- **GenMAPP MAPP Relevant to Results**

GenMAPP MAPP Relevant to Results - Ribosome

Ribosomal Proteins

Gene Database
So-Std_20151201special.gdb

Expression Dataset
Name: CompiledRawData_ForGenMAPP_20151210_HMH
Color Set: F60-C60
Gene Value: F60-C60: AverageLogRatio_F60-C60
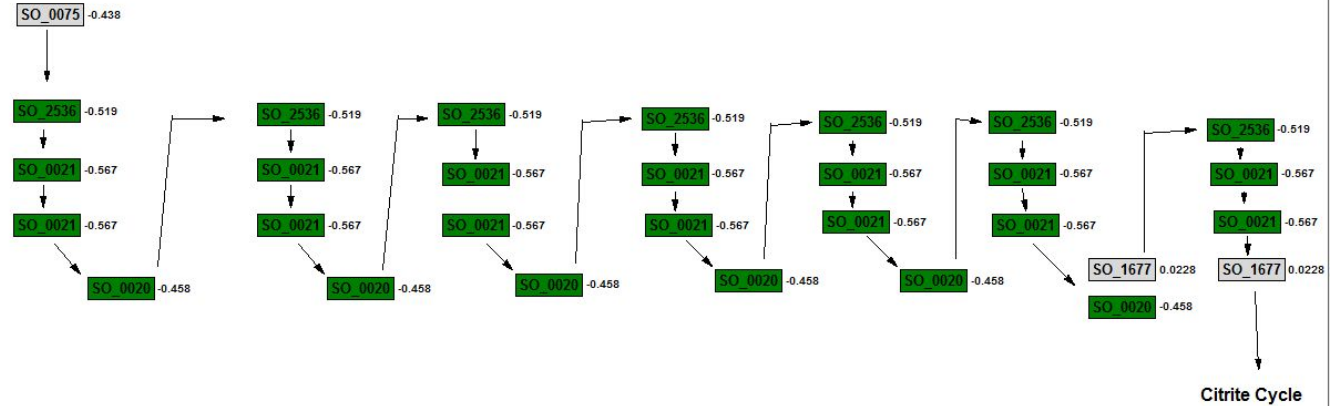Heavy Metal HaterZ 4 Lyfe

Legend: F60-C60
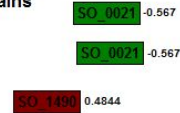- Increase Log Change
- Decrease Log Change
- No criteria met
- Not found

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
SO_0230 0.5146	SO_0240 0.2852	SO_0252 0.0389	SO_0227 0.3909	SO_1629 0.4474	SO_4247 1.1676
SO_0231 0.3324	SO_0241 -0.123	SO_0253 0.744	SO_0226 0.1959	SO_1207 0.1988	SO_4246 0.7042
SO_0232 0.6153	SO_0242 0.2248	SO_0254 1.0529	SO_0223 1.0311	SO_2301 0.8982	SO_3652 0.7151
SO_0233 1.0171	SO_0243 0.2593	SO_0255 0.5653	SO_0222 0.9601	SO_2302 0.0516	SO_3651 0.0441
SO_0234 0.2941	SO_0244 0.6862	SO_0257 1.1646	SO_0221 0.5616	SO_0007 1.1624	SO_1357 0.7616
SO_0235 1.0538	SO_0245 0.1871	SO_3940 0.5927	SO_0220 0.426	SO_4120 1.0637	SO_1360 0.4877
SO_0236 0.4983	SO_0246 0.7667	SO_3939 0.4383		SO_2780 -0.242	SO_2402 0.3236
SO_0237 0.1608	SO_0247 1.49			SO_3927 0.8738	SO_3537 0.1991
SO_0238 1.1968	SO_0248 0.7147			SO_3928 0.4861	SO_1288 0.4317
SO_0239 -0.571	SO_0249 0.51			SO_3930 0.3979	S0_2112 0.1732
	SO_0250 0.8393				

GenMAPP MAPP Relevant to Results - Fatty Acid Degradation

Acknowledgements

- Dr. Dahlquist

- Dr. Dionisio

- Departments of Biology and Computer Science at LMU

- Biological Databases Students

  - Thank you for listening!

References

Heidelberg, J. F., Paulsen, I. T., Nelson, K. E., Gaidos, E. J., Nelson, W. C., Read, T. D., ... & Fraser, C. M. (2002). Genome sequence of the dissimilatory metal ion–reducing bacterium Shewanella oneidensis. *Nature biotechnology, 20*(11), 1118-1123. doi:10.1038/nbt749

Yang, Yunfeng, et al. "Snapshot of iron response in Shewanella oneidensis by gene network reconstruction." *BMC genomics* 10.1 (2009): 131.