Metabolic Reconstructions from Global Ocean Sampling (GOS) Marine Metagenome

Mathangi Thiagarajan
J. Craig Venter Institute

Pathways Tools Workshop 2010
- Metagenomics
- The Global Ocean Sampling (GOS) Project
- GOS - Community Makeup
- High Throughput Data Processing
- Metabolic Reconstruction – Mapping to MetaCyc and KEGG
- Metarep (Visualization) – Integrating with MetaCyc and KEGG
- Pathways Tools for GOS & metagenomic projects
- Conclusion
- Acknowledgements
Metagenomics

- Examining genomic content of organisms in community/environment to better understand
  - Diversity of organisms
  - Their roles and interactions in the ecosystem

- Cultivation independent approach to study microbial communities
  - DNA directly isolated from environmental sample and sequenced
Global Ocean Sampling Expedition

Investigate the fundamental microbial contributions from the Ocean waters to energy and nutrient cycling by analyzing its
a) biogeochemical cycling
b) community structure and function
c) microbial diversity
d) adaptation and evolution

GOS Phase I - Published in PLOS Biology 2007
GOS Circumnavigation - Analysis Phase
Sample Filtration

- **Zooplankton**
- **> 20 μm Prefilter**
- **20 - 3.0 μm**
  - Microzooplankton / Larger Phytoplankton
- **3.0 - 0.8 μm**
  - Picoplankton and Large Cyanobacteria
- **0.8 - 0.1 μm**
  - Prokaryotes

Virus / Phage
GOS circumnavigation data
229 stations and 291 samples

- 0.1µm
- 0.8µm
- Viral
- 3.0µm
<table>
<thead>
<tr>
<th></th>
<th>Reads</th>
<th>Proteins</th>
<th>Sequencing Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase I</td>
<td>7.6 Million</td>
<td>9.8 Million</td>
<td>Sanger</td>
</tr>
<tr>
<td>Circumnavigation</td>
<td>48 Million</td>
<td>~53 Million</td>
<td>Sanger + 454</td>
</tr>
</tbody>
</table>
GOS dataset is expanding the protein universe

Extrapolation based on amount of GOS sequence data currently available but not yet released to public domain
Community makeup
## Taxonomic makeup of GOS samples based on 16S data from shotgun sequencing

<table>
<thead>
<tr>
<th>Phylum or Class</th>
<th>Fraction$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha Proteobacteria</td>
<td>0.32</td>
</tr>
<tr>
<td>Unclassified Proteobacteria</td>
<td>0.155</td>
</tr>
<tr>
<td>Gamma Proteobacteria</td>
<td>0.132</td>
</tr>
<tr>
<td>Bacteroidetes</td>
<td>0.13</td>
</tr>
<tr>
<td>Cyanobacteria</td>
<td>0.079</td>
</tr>
<tr>
<td>Firmicutes</td>
<td>0.075</td>
</tr>
<tr>
<td>Actinobacteria</td>
<td>0.046</td>
</tr>
<tr>
<td>Marine Group A</td>
<td>0.022</td>
</tr>
<tr>
<td>Beta Proteobacteria</td>
<td>0.017</td>
</tr>
<tr>
<td>OP11</td>
<td>0.008</td>
</tr>
<tr>
<td>Unclassified Bacteria</td>
<td>0.008</td>
</tr>
<tr>
<td>Delta Proteobacteria</td>
<td>0.005</td>
</tr>
<tr>
<td>Planctomycetes</td>
<td>0.002</td>
</tr>
<tr>
<td>Epsilon Proteobacteria</td>
<td>0.001</td>
</tr>
</tbody>
</table>

$^a$Values shown are averages over all samples.
Phylogenetic Distribution in the Indian Ocean across size-classes

- 0.1 µm
- 0.8 µm
- 3.0 µm

- Synechococcus sp.
- Bacteroidetes
- Verrucomicrobia
- ds DNA viruses
- Planctomycetes

[Diagram showing phylogenetic distribution across different size-classes]
GOS increases size and diversity of known protein families

GOS: prokaryotes, eukaryotes
Known: prokaryotes, eukaryotes
Viruses in the Marine Environment

- Abundant: $\sim 10^7 \text{ ml}^{-1}$ of surface seawater
- Diverse: VBR $\equiv 10$; $\sim 10$-fold greater diversity than microbial hosts
- Influence microbial diversity through infection and host cell lysis
- Mediators of horizontal gene transfer
- Influence biogeochemical cycling, particularly carbon
High-throughput Metagenomic Data Analysis

- Metagenomic Assembly
  - Sanger data
  - 454 data
  - Illumina data (HMP)

- Protein Clustering
- Annotation Pipeline
  - Structural Annotation (coding + non coding)
  - Functional Annotation

- Taxonomic Classification

- Fragment Recruitment

- Sample Comparison
  - Taxonomic level
  - DNA library level
  - Protein level
  - Functional and metabolic profiles

- Functional linkages via Operons

- Metabolic Reconstruction

- Linking to Metadata
Metagenomic Data Processing - Annotation pipeline

**Structural Annotation**
- Nucleotide sequence in multi fasta file
  - BLAST vs. ncRNA db
- tRNA nucleotide sequence in multi fasta file
  - Soft-masked nucleotide sequence in multi fasta file
  - Clear range option
- Structural annotation via naive ORF and MetaGene
- Peptides (ORFs) in multi fasta file

**Functional Annotation**
- PRIAM
- BLAST vs. PANDA
- HMMer vs. Pfam & TIGRFAM
- TMHMM
- Lipoprotein motif string search
- Functional annotation via rules Heirarchy
- Annotated peptides in multi fasta file

Published in SIGS
Annotation Rules Hierarchy

Evidences

- HMM Hits
  - PFAM
  - TIGRFAM
  - above trusted cut off

- BlastP Hits
  - 1e-5; 35% ident; 80% cov

- Rps PRIAM Hits
  - 1e-10 cutoff
  - EC Numbers

- CHAR database

- TmHMM
- Lipoprotein Motif

Annotation Rules

1. TIGRFAM/PFAM (Equivalog)
2. Characterized (CHAR) BlastP Hit
3. TIGRFAM/PFAM (Non-Equivalog)
4. CDP (conserved domain protein) blastp hit
5. TmHMM hit: “membrane protein”
6. Lipoprotein motif: “lipoprotein”
7. “hypothetical protein”

Common Names, Gene Symbols, EC Numbers, GO Terms, TIGR Role ids
Viral Metagenomic (functional)Pipeline
Annotation Rules Hierarchy (Viral)

- PFAM/TIGRFAM_HMM, equivalog above trusted cutoff
- ACLAME_PEP, %id>= 50, coverage >= 80, e-value <= 10^{-10}
- ALLGROUP_PEP, %id>= 50, coverage >= 80, e-value <= 10^{-10}
- ACCLAME_HMM matches, > 90% coverage, e-value < 10^{-5}
- PFAM/TIGRFAM_HMM, non-equivalog above trusted cutoff
- CDD_RPS, %id>= 35%, coverage >= 90% of CDD-domain, e-value <= 1e^{-10}
- FRAG_HMM, e-value < 1e^{-5}
- ACLAME_PEP, %id >= 30%, coverage >= 70%, e-value <= 1e^{-5}
- ALLGROUP_PEP, %id >= 30%, coverage >= 70%, e-value <= 1e^{-5}
- No evidence -> hypothetical protein
Metagenomic Assembly

Advantages

- Provides genomic context
- Reduces redundancy and complexity
- Improves annotation
- Mechanism to isolate environment specific gene regions

Challenges

- Coverage dependent
- Variation can limit the length of assemblies
- Can mask diversity

• Celera Hybrid Assembler has been updated to work with 454 Titanium reads
• Will further optimize assembly process to capture environmental diversity
Metagenomic Data Processing - Continued

- **Protein Clustering**: JCVI’s Protein clustering (S. Yooseph)
- **Taxonomic Classification**: APIS (J. Badger)
- **Fragment Recruitment**: Advanced Reference Viewer (D. Rusch)
- **Metagenomic Assembly**: Celera Assembler (G. Sutton & J. Miller)
- **Sample Comparison**

Making sense of everything in the context of **METADATA**
General Questions

- Who are they?
  - Species, Taxonomic distribution...
- How many?
  - Distribution across sites and filters
- What are they doing?
  - Functional profiles
  - Metabolic profiles
MR Specific Questions

- Metabolic profiles across sites and filters
- Pathways coverage and abundance
- What known characterized pathways and how many?
- What novel pathways are there?
- Metabolic network
Metabolic Reconstruction

- From the Annotation Pipeline (orf based)
  Proteins → EC assignment → Pathways prediction
  (EC to MetaCyc/Kegg mapping)

  **Sources for EC:**
  - TIGRFAM
  - PFAM
  - High confidence blast hit to Uniref100/Panda
  - RPSblast to EC profiles from PRIAM

- From BlastX to a Functional database (read based)
  Reads → Blastx Metacyc/Kegg → Pathways prediction
Browse/analyze/compare pathways across datasets in the context of annotation and Metadata

METAREP
JCVI Metagenomics Reports

METAREP is a web interface designed to help scientists to view, query and compare annotation data derived from proteins called on metagenomics reads

Developer: Johannes Goll
Published in Bioinformatics

website www.jcvi.org/metarep
source code http://github.com/jcvi/METAREP
blog http://blogs.jcvi.org/tag/metarep
contact metarep-support@jcvi.org

www.jcvi.org/metarep
Browse Metacyc Pathways
gos-phase-l-sanger (GOS Phase I)

Pathway Classification:

- Degradation/Utilization/Assimilation
- Other
- Metabolism of Small Molecules
- Metabolism of Amino Acids
- Metabolism of Nucleotides
- Metabolism of Lipids
- Metabolism of Carbohydrates
- Other

MetaCyc Class: Degradation/Utilization/Assimilation

Synonyms: assimilation, utilisation

Summary:
This class contains pathways by which various organisms degrade substrates to serve as sources of nutrients and energy, utilize exogenous sources of essential metabolites, or assimilate certain sources of essential treatments.

Parent Class:
Pathways

Child Classes:
- Amino Acid Degradation (16)
- Amino Acid Degradation (18)
- Amino Acid Degradation (20)
- Amino Acid Degradation (21)
Aromatic Compounds Biosynthesis

MetaCyc Pathway: β-alanine betaine biosynthesis

Detoxification (level 1) [375,280 hits]

Generation of precursor metabolites and energy (level 1) [1,458,128 hits]

Signal transduction pathways (level 2) [5,076 hits]

Secondary metabolites biosynthesis (level 1) [1,369,338 hits]

Siderophore Biosynthesis (level 2) [5,641 hits]

Repression/Utilization/Assimilation (level 1) [2,150,781 hits]

β-alanine betaine biosynthesis (level 2) [11,461 hits]
Pathways Tools for GOS

- Metagenomic specific predictions - Incorporate taxonomic resolution when predicting pathways
- Confidence Scores for the pathways
- Incorporate more annotation evidence types in predictions other than EC
- Ability to overlay and visualize expression data
- Full integration of pathways tools into Metarep
- Performance enhancements to handle metagenomic data volume
Conclusion

- Who are they?
  Species, Taxonomic distribution...
- How many?
  Distribution across sites and filters
- What are they doing?
  Functional profiles
  Metabolic profiles
Acknowledgements

Metagenomic PI’s & Coordinators
Shibu Yooseph
Barbara Methe

Metagenomic Bioinformatics & Software Engineers
Johannes Goll
Jeff Hoover
Alex Richter
Aaron Tenney
Daniel Brami
Monika Bihan
Kelvin Li

GOS Funded by
DOE Genomics: GTL Program
Gordon and Betty Moore Foundation
J. Craig Venter Science Foundation

Metagenomic PI’s
Doug Rusch
Andy Allen
Shannon Williamson
Andrey Tovtchigretchko
Jonathan Badger

Postdocs
Seung-Jin Sul
Youngik Yang

Leadership
Robert Friedman, Karen Nelson
& J. Craig Venter
Questions

Thank You