Managing the Data Bonanza: Generating, Analyzing and Sharing Data for Megasequencing Projects

Narayan Desai

Mathematics and Computer Science Division
Argonne National Laboratory
and
University of Chicago
XLDB 5 – October 19, 2011
Talk Overview

- Metagenomics Primer
- Data Production
- Analysis
- Data Sharing
- Megasequencing Projects
Metagenomics ...

Definition:
“random shotgun DNA sequencing applied directly to environmental samples”

- whole shotgun metagenomics
- Result is a combination of short reads of DNA from all organisms in sampled community
  - Mixed together

Who are they?

What are they doing?
Metagenomics and Discovery

Today:
Mapping our knowledge to help understand microbial ecology
→ using existing knowledge

Via:
Mapping to curated databases

Future:
Discover new biology from computationally mining the unknowns
→ Patterns
→ co-occurrence
→ Exclusion
→ ..

Example
Systematic discovery of patterns
e.g. CRSPR by Jill Banfield
Biology rapidly changing. From this ...

The leatherback turtle, Dermochelys coriacea

These are: “biology.png”, “biology.gif” and “biology.jpg”.

http://www.the-aps.org/education/

http://www.ferrum.edu/majors/biology.jpg

http://www.oneocean.org/ambassadors/track_a_turtle/biology
... to this. (in ~2003)
... benchtop scale now in 2011

From factory to bench-top in 5 years

And

>70% of Illumina machines go to “small” customers\(^{(1)}\)

\(^{(1)}\) From Illumina at 2010 GIA meeting
Metagenomics and computing

Cost per Megabase of DNA Sequence

Moore’s Law

2001-2011
Data generation cost

- Sequencing cost for single lane of Illumina 1x100bp is $1170.62 for approx 10 GBp (gigabasepairs)
- For many large studies sequencing cost are dominated by sample preparation (and sequencing library construction)

Example cost:

100,000 16S amplicon reads
(Source: Rob Knight, Colorado)
- sample extraction: $8
- PCR and pooling: $5
- Sequencing: $4
Sum: $17

Metagenomics 10GBp (or 20GBp)
(Source: Marc Domanus, ANL)
- Sample prep: $100
- Lib prep: $64 (non mate-pair) $389 (mate-pair)
- Sequencing: $1171 @ 10GBp $2174 @ 20GBp
Sum: $1335 Sum: $2663
MG-RAST
metagenomics analysis server

## Metagenome Analysis

### Analysis Views
- Organism Classification
- Functional Classification
- Recruitment Plot
- QIIME Report

### Data Selection
- **Metagenomes**
- Annotation Sources
- Max. e-Value Cutoff
- Min. % Identity Cutoff
- Min. Alignment Length Cutoff

**Workbench**
- use proteins from workbench

### Data Visualization
- bar chart
- tree
- table
- heatmap
- PCA

### Workbench (0 Proteins) | Getting Started | Organism table 1

This data was calculated for metagenomes 4441679.3, 4441680.3, 4441681.3 and 4441682.3. The data was compared to M5NR using a maximum e-value of None and a minimum identity of None%. The data has been normalized to values between 0 and 1. If you would like to view raw values, redraw using the form below.

**group table by**

Download data matching current filter

**display**

15 items per page

Displaying 1 - 15 of 607

<table>
<thead>
<tr>
<th>metagenome</th>
<th>source</th>
<th>domain</th>
<th>phylum</th>
<th>class</th>
<th>abundance</th>
<th>avg eValue</th>
<th>%Ident</th>
<th>proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>4441681.3</td>
<td>M5NR</td>
<td>Bacteria</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>332962</td>
<td>-2.92</td>
<td>83.23</td>
<td>12146</td>
</tr>
<tr>
<td>4441682.3</td>
<td>M5NR</td>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>50026</td>
<td>3.76</td>
<td>77.94</td>
<td>24605</td>
</tr>
<tr>
<td>4441679.3</td>
<td>M5NR</td>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>46841</td>
<td>3.67</td>
<td>78.77</td>
<td>24069</td>
</tr>
<tr>
<td>4441680.3</td>
<td>M5NR</td>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>31308</td>
<td>3.87</td>
<td>78.34</td>
<td>17701</td>
</tr>
<tr>
<td>4441681.3</td>
<td>M5NR</td>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>26796</td>
<td>3.56</td>
<td>78.18</td>
<td>15415</td>
</tr>
<tr>
<td>4441680.3</td>
<td>M5NR</td>
<td>Bacteria</td>
<td>Bacteroidetes</td>
<td>Bacteroides</td>
<td>21190</td>
<td>-3.18</td>
<td>81.07</td>
<td>11495</td>
</tr>
<tr>
<td>4441679.3</td>
<td>M5NR</td>
<td>Bacteria</td>
<td>Bacteroidetes</td>
<td>Bacteroides</td>
<td>20000</td>
<td>-4.74</td>
<td>80.47</td>
<td>10584</td>
</tr>
</tbody>
</table>
**Brief history of MG-RAST**

- **December 2007 (v1)**
  - 100+ groups and ~250 data submitters
  - 100+ data sets, ~10+ GBp total size

- **October 2009 (v2)**
  - Pre-publication sharing available
  - ~1500 data submitters, ~300 public data sets
  - 6000+ data sets
  - 200+ GBp total data set size
  - About ~30 GBp/month throughput

- **March 2011 (release v3)**
  - 2500+ data submitters
  - ~2000 public data sets
  - 25,000 data sets total
  - Throughput:
    - 47GBp in 24h
    - 3000 submissions in 24h

- **September 2011 (v3.1.1)**
  - 32,000+ data sets
  - 4.4 TBp analyzed
  - 500 users / day
  - 7000+ users total
2011: MG-RAST adoption

- ~32,000 data sets, hundreds of groups
- 7000+ users
- ~4300 public, with metadata, 45GBp
- >> 4.4 Terabases completed (gigabase == \(10^{12}\) basepairs)
Analysis Open Challenges

- Data volume reduction is the key goal
  - Superlinear algorithms
  - Data growth faster than Moore’s law
- Read assembly might help
  - Reduces number of strings
  - But will produce some chimeric contigs
- Need fast screening techniques
  - Which data sets are actually interesting
- Before long, analysis of all data sets may be unsustainable
  - Computational costs dwarfing all others
- Cost breakdown means metagenomic results are more valuable than raw data
  - Sea change in bioinformatics data ecosystem
- Virtual surveys are the long term goal
Data Archives and Sharing

- Value is in analyzed metagenomic data sets, not raw data
  - Though raw data is still useful in many cases
- Sharing of analysis results requires improvements in metadata
  - All sample collection details (biome, ph, etc)
  - Provenance
  - Formats emerging through the Genomic Standards Consortium
- Centralized archives exist (NCBI, EMBL, etc)
  - But can’t possibly scale to handle data volume from decentralized, democratized sequencing
  - Not clear this architecture is even correct for the new workload
- More likely, a moderate number of community brokers will fill the gap
  - Driven by domain or funding
  - Consensus metadata, analysis, and provenance
  - Federation with other archives
  - Bilateral peering/data sharing arrangements
Results sharing

- Raw data sharing is established
  - GenBank, SRA, EMBL, ...
- Suitable for low volume data science
- Reproducibility no longer exists for current data
- Large volume data science requires result sharing
  - Require community agreement and standards
  - GSC’s M5 initiative provides transport encoding
    - Metagenome transport format (MTF)

Fixes the re-computing issue
OSDF: Data access API for community bioinformatics resources

- Open Science Data Framework
- APIs for data discovery, storage and retrieval
  - Metadata-based queries (environmental data, assays, etc)
  - Analysis graph queries (provenance traversal)
- Reference implementation of the archive layer (Shock)
  - Tools ecosystem built on top of OSDF APIs
  - Initial release available
- Support from major metagenomics service providers
  - MG-RAST and IMG/M
  - QIIME (16S amplicon analysis software)
  - CLOVR (Bio VM environment)
- Major projects (EMP & HMP) will support OSDF API for data access
www.earthmicrobiome.org
EMP Products:

- **Earth Microbiome Gene Atlas (EM-GA)** – a repository and database for all sequencing and metadata information.

- **Earth Microbiome Assembled Genomes (EM-AG)** – ~500,000 microbial genomes and extra chromosomal elements derived from metagenomic data.

- **Earth Microbiome Metabolic Reconstruction (EMMR)** – describe changes in metabolite profiles between all samples, providing another metric against which to refine biome descriptions.

- **Earth Microbiome Visualization Portal (EM-VIP)** – A web portal like Google Earth – exploring microbial space.
Challenges

- 2.4 Quadrillion Base Pairs (2.4 Peta bp) = 8000 HiSEQ2000 runs.
- Global Environmental Sample Database (GESD): identification and selection of 200,000 environmental samples, soil, air, marine and freshwater, host-associated, etc.
- The standardization of sampling, sample prep and sample processing, cataloging and sample metadata – Genomic Standards Consortium.
The Earth Microbiome Project

Questions?