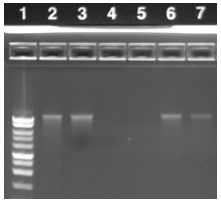
MG-RAST: Scalable Metagenomic Analysis

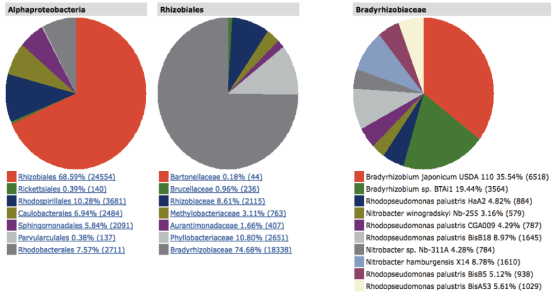
Narayan Desai
desai@mcs.anl.gov
Mathematics and Computer Science Division
Argonne National Laboratory

Metagenomics …

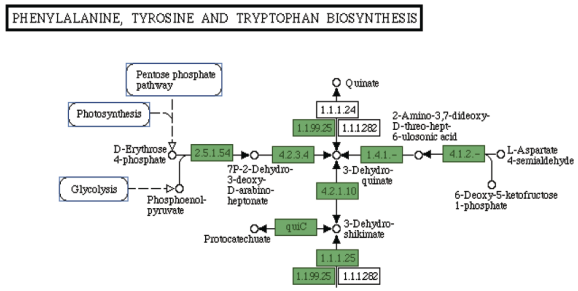
- random shotgun DNA sequencing applied directly to environmental samples
- whole shotgun metagenomics
- Result is a combination of short reads of DNA from all organisms in sampled community
  - Mixed together in unknown proportions



Who are they?



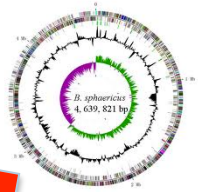
What are they doing?



Distinctions from Genomic Analysis

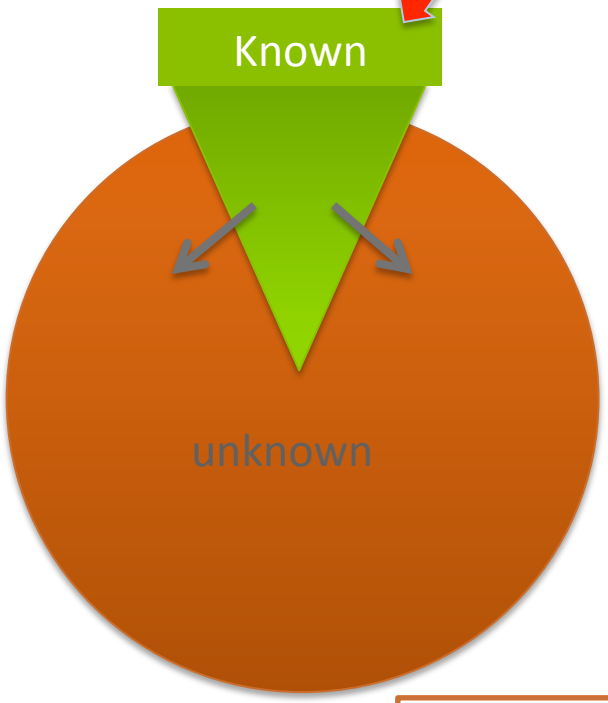
- Metagenomics is far more widely applicable
  - Genomic studies depend on cultivation
    - Only a small % of microbes have known cultivation conditions
  - Community composition is directly measured
    - Useful for microbial ecosystems
    - Builds towards community metabolic models
- Sequence Error
  - Single base pair errors from PCR or sensor failures
  - Often indistinguishable from evolution in communities
- Many bioinformatics heuristics are built for clonal data sets
  - Assume a single genome
  - Depend on inherent structure of a single genome for error correction
- Pan Genomes
  - Closely related organisms in the same community
  - Assembly can result in chimeric contigs

Metagenomics and Discovery



Today:
Mapping our knowledge to help understand microbial ecology

→ **using existing knowledge**



Future:
Discover new biology from computationally mining the unknowns

→ Patterns
→ co-occurrence
→ Exclusion
→ ..

Via:
Mapping to curated databases

Example
Systematic discovery of patterns
e.g. CRSPR by Jill Banfield

MG-RAST

metagenomics analysis server

METAGENOME ANALYSIS

Analysis Views

- » Organism Classification
- Functional Classification
- Recruitment Plot
- QIIME Report

Data Selection

Metagenomes	4441679.3, 4441680.3, 4441681.3, 4441682.3
Annotation Sources	M5NR
Max. e-Value Cutoff	None
Min. % Identity Cutoff	None
Min. Alignment Length Cutoff	None
Workbench	use proteins from workbench

Data Visualization

barchart | tree | table | heatmap | PCA | generate

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 class

download data matching current filter

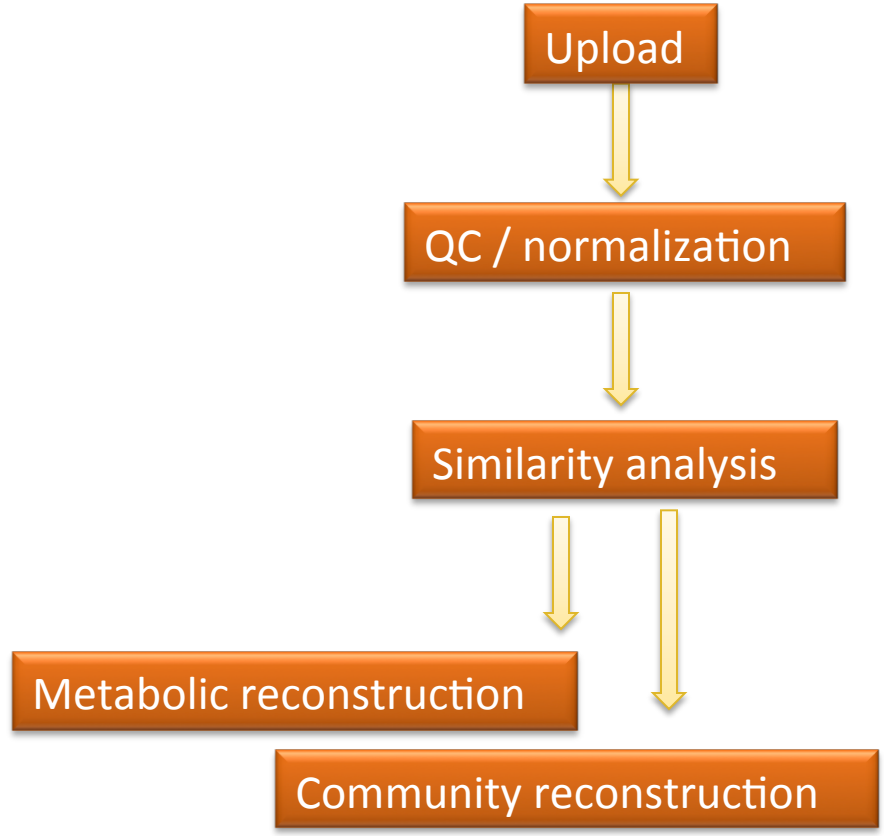
display 15 items per page

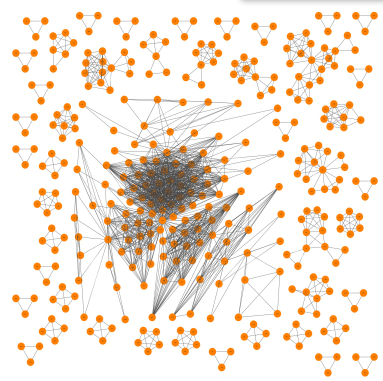
displaying 1 - 15 of 607 next» last»

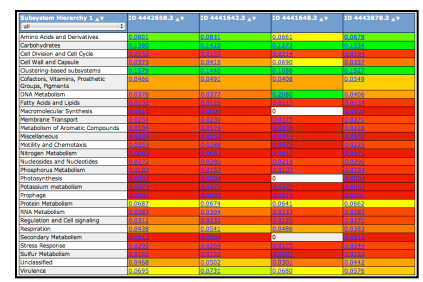
metagenome	source	domain	phylum	class	abundance	avg eValue	%ident	# proteins	to workbench
all	M5NR	all							
4441681.3	M5NR	Bacteria	Proteobacteria	Gammaproteobacteria	332902	-2.92	83.23	12146	☐
4441682.3	M5NR	Bacteria	Firmicutes	Clostridia	50026	-3.76	77.94	24605	☐
4441679.3	M5NR	Bacteria	Firmicutes	Clostridia	46641	-3.67	78.77	24069	☐
4441680.3	M5NR	Bacteria	Firmicutes	Clostridia	31808	-3.87	78.34	17701	☐
4441681.3	M5NR	Bacteria	Firmicutes	Clostridia	26796	-3.55	78.18	15415	☐
4441680.3	M5NR	Bacteria	Bacteroidetes	Bacteroidia	21190	-5.18	81.07	11495	☐
4441679.3	M5NR	Bacteria	Bacteroidetes	Bacteroidia	20000	-4.74	80.47	10594	☐

Brief MG-RAST History

- 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 sets
  - 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
- May 2012 (v3.1.2)
  - 48K data sets
  - 13.3 TBp analyzed
  - 500 users / day
  - 7500+ users total
  - ~750X performance improvement compared with the project start





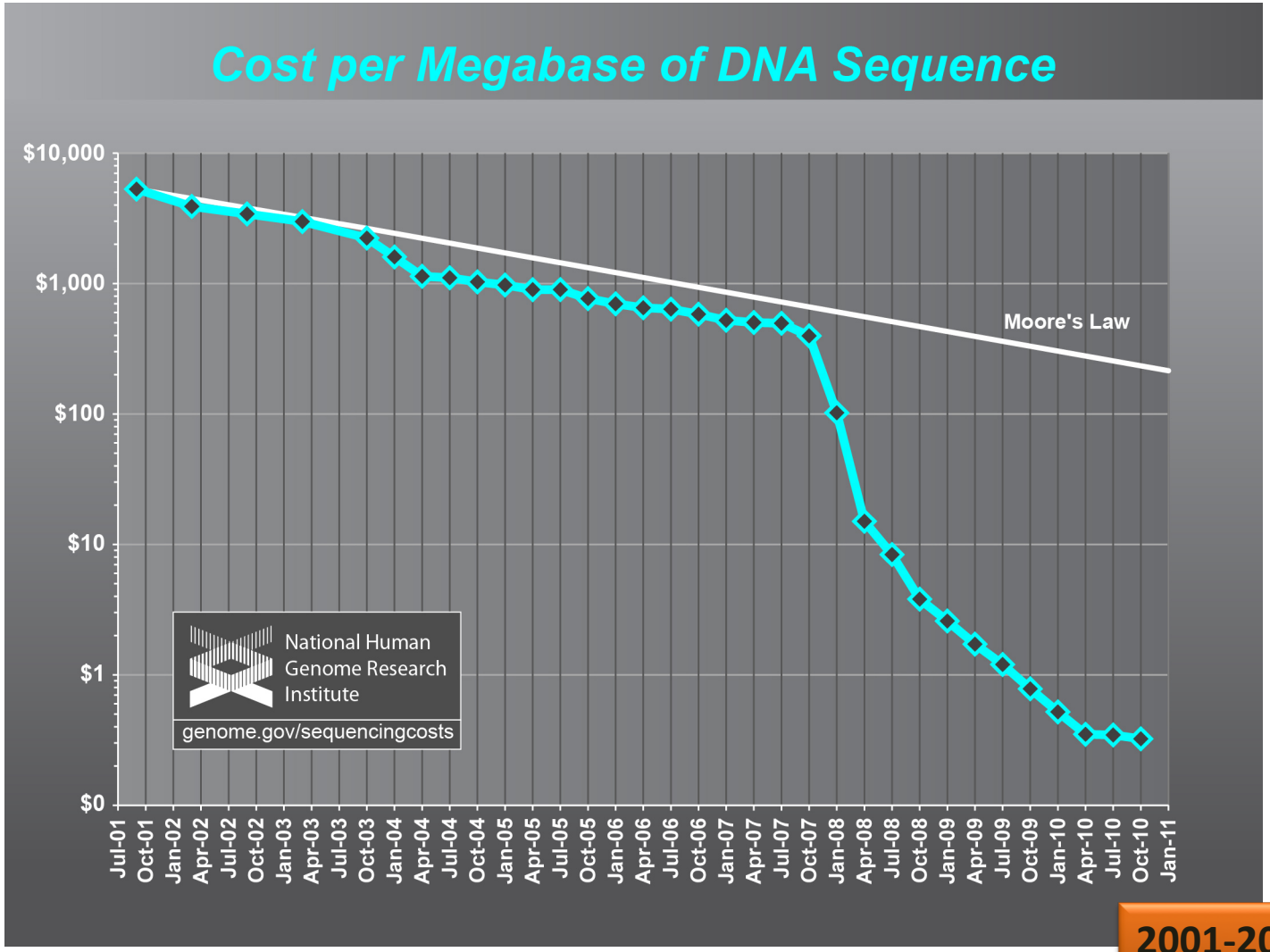


simplified

MG-RAST Analysis Overview

- Quality Control
  - Construction of technical replicate proxy reference + multiple sequence alignment
  - Nucleotide per read position (for adapter contamination/systematic bias)
  - Used to fail data sets as well as discard/trim bad reads
- Feature Prediction
  - Find protein coding regions and infer reading frames
- Clustering
  - Cluster predicted proteins using 90% sequence similarity
- Sequence Similarity Search
  - Search against non redundant database of proteins with functional and identity annotations
  - SuperBLAT
    - In house modified OpenMP version of BLAT code

Metagenomics and computing



2001-2011

Complicating Factors in Metagenomic Analysis

- More sequencing depth is needed for complex communities
  - Price changes are likely to result in larger data sets, not cheaper data sets of the same size
- Computational culture in bioinformatics is underdeveloped
  - Real progress requires a combination of biological and computer science expertise
- Progress is limited by data analysis capabilities
- Computational costs dominate
- Changes in data production costs are producing extreme budgetary shocks
- 20K of chemistry will produce 600 giga-base pairs of sequence data
  - 3.6 million for blast-based gold standard analysis (ec2 list prices)
  - 5-10K for MG-RAST analysis approach
- Value in the ecosystem is shifting from raw data to data products
- These costs require informed decisions by the metagenomic community
  - Due to cultural/expertise divide, these discussions are rare

Metagenomic Analysis Open Challenges

- Need a robust data ecosystem
  - maximize scientific progress with limited computational resources
  - Enable/incentivize data product re-use
  - Support virtual surveys
- Economics of data point at service oriented architecture
  - Application centric model isn't ideal when data re-analysis isn't affordable
  - Service architectures aren't well suited to HPC systems
  - …but we still need efficient computations
- Current analysis approaches cannot scale
  - Need new analysis algorithms
  - Data reduction techniques
  - Fast screening/classification
- Removal of the annotation bias in metagenomics

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
  - This architecture is likely wrong 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

Shock - Building a robust data ecosystem

- Shock is a data management service
  - Provides an archival interface to data analysis products
  - Sequence data workbench
  - Building toward a federated archive
- Built to layer on top of a distributed object storage system (Ceph, Triton, etc)
- Support for rich metadata
  - Environmental information
  - Computational (and wetlab) provenance
  - Direct support for metadata queries
- Enabling efficient use of storage and network resources
  - Moving computations to data, when it makes sense
  - Minimizing data transmission when it does not (subselection, indexing)
- Built to store and move large data sets
  - Goal is NGS sequence data sets (1-10TB+) at the high end
  - Support distributed computation on these data sets
- Primary interfaces are RESTful HTTP interfaces
  - Designed to span systems

Results sharing

- Raw data sharing is established
  - GenBank, SRA, EMBL, …
- Suitable for low volume data science
- <u>Reproducibility no longer exists for current data</u>
- 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

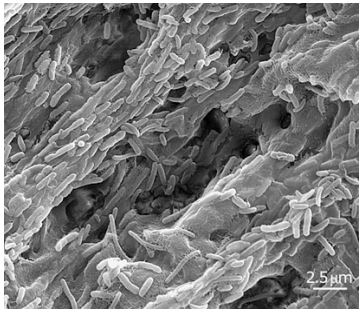
Moving towards Service Architectures

- Without a viable scalable computational strategy, we need to aggressively re-use existing data products
  - Switch from a de novo approach to one of recombination
  - In addition to access to the raw data, we need a network of services that provide analytical services in this data ecosystem
- This problem is a poor fit to traditional HPC systems
  - The batch queue model doesn't accommodate long running services
  - Security policies don't fit
  - Software infrastructure is not really correct either
- So the Department of Energy Systems Biology Knowledgebase is taking a different approach..

Enable DOE Mission Science

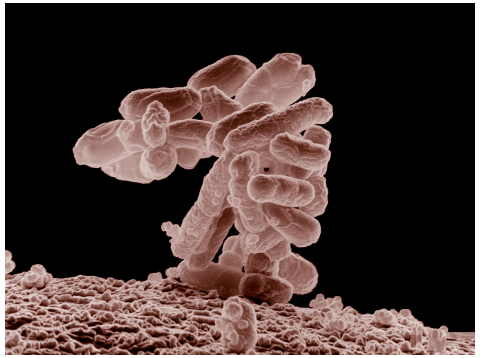
KBASE predictive biology

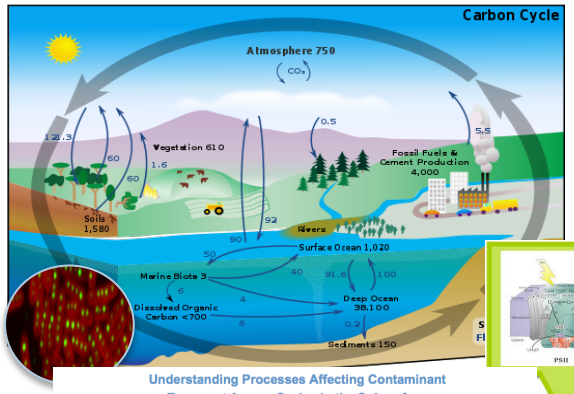
Communities

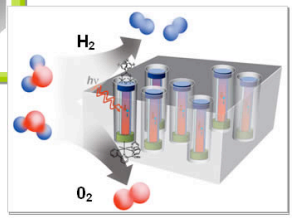


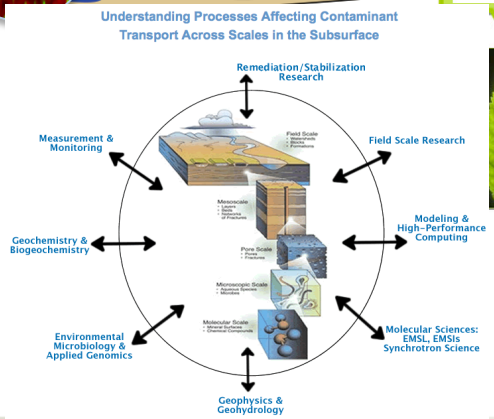
Plants

Microbes



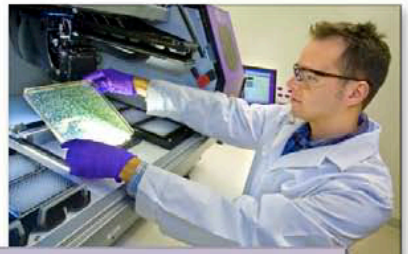






KBASE
predictive biology

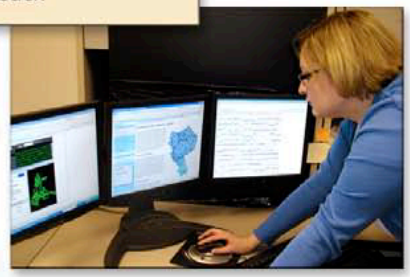
Data generators



Open-Access Data and Information Exchange

- Access through several flexible interfaces
- Retrieval of experimental data and products of modeling and simulation
- Working environment for testing hypotheses by *in silico* experimentation
- Provision and sharing of user feedback

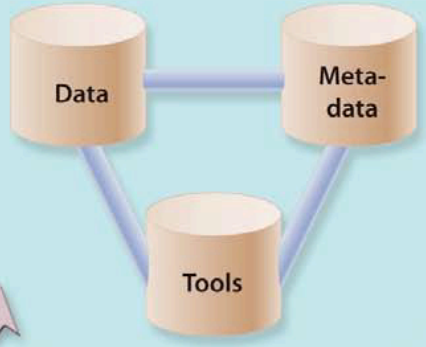
Data users



Seamless Submission and Incorporation of Diverse Data

- Standards for data and metadata representation
- Quality control and assurance capabilities
- Automated systems for depositing and updating bulk data
- Tracking and evaluation of data use

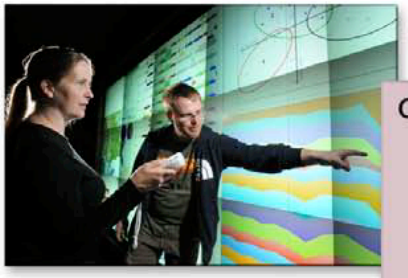
DOE Systems Biology Knowledgebase



Community-Wide Stewardship

- Advisory committee
- New social networking tools and capabilities
- Development
- Value-added analysis
- Training, tutorials, support staff

Software and tool developers



Open Development of Open-Source Software and Tools

- Data analysis and visualization tools
- Resources for *in silico* experimentation
- Modeling and simulation tools
- Customizable tools with layers of functionality
- Application programming interfaces (APIs) for open development of software
- Tracking and evaluation of tool use

Kbase Computational Infrastructure

- Distributed across the 4 sites (ANL, BNL, LBL, ORNL)
  - Primary resource is the Magellan system at ANL
- Magellan is a moderate scale private openstack cloud
  - ~8000 commodity cores
  - ~1PB of storage
  - 30 TB of memory
  - QDR Infiniband
  - Nvidia GPUs
- Cloud management features are a good match for Kbase
  - Multi-tenancy, flexible software stack
- But performance is still a challenge
  - Moderate virtualization overhead
- But there are some bright spots
  - Have demonstrated wide area data transfers (memory to memory) at 95 gigabits
    - With only 10 VMs and 40 TCP streams
  - We suspect we will even be able to build out storage properly to source/sink to disk at similar speeds
  - Paves the way for on demand provisioned DTNs!

A Quick Word About Metagenomics Algorithms

- Metagenomics is rapidly moving out of an easy computational regimen
  - Loosely coupled tasks are trending towards more HPC-like applications
  - With a big data flavor
  - More statistics and machine learning, as well as graph algorithms
- Sequence similarity search will stop being the dominant algorithm soon
  - We just can't afford it for everything
- Assembly of some sort will be key
  - Many large scale assemblers (nucleotide) use De Bruijn graph data structures
  - Combination of individual reads/observations into larger consensus aggregates
- We've begun to investigate protein-level assembly
  - But the eventual goal will be to co-assemble all data in MG-RAST
  - Which would require 20-30 TB of memory at a minimum
- Distributed graph algorithms will definitely be key

Fast Classification Approaches

- Goal is twofold
  - Enable use of more targeted (cheaper) methods
  - Prioritize use of computational resources
- One approach is to build core community metabolic models
  - Small database of critical proteins
  - These markers could be used to select between finer characterization methods tuned to particular biomes
- Another approach could use databases of uncharacterized proteins to screen previously seen fragments
  - Prioritize novelty

Removal of Annotation Bias in Metagenomics

- All current metagenomics approaches depend heavily on protein annotations
  - Either occurrence of proteins in closed genomes
  - Or characterizations of protein function
- Both approaches depend on manual (low throughput) methods
- Due to the plummeting cost of metagenomic data sets, alternative approaches are becoming plausible
- These kinds of tasks will be “big data” problems
  - Large scale data mining
  - Feature prediction
- A few approaches
  - Databases of uncharacterized proteins
  - Correlations between known and unknown proteins across biomes
  - Global protein survey

Questions?