The Biological, Algorithmic and Computational Challenges of Systems Biology

Rick Stevens
Argonne National Laboratory
The University of Chicago
Stevens@anl.gov
Stevens@ci.uchicago.edu
IBM Blue Gene P Supercomputer (Intrepid)
MiRA at Argonne is 48 Racks
10 Petaflops/sec \((10^{16} \text{ ops/s})\)
800 TB RAM
\(~800,000\) cores
35 Petabytes of Storage
6-8 Megawatts of Power
Jan 9th, 2010, BGI (previously known as Beijing Genomics Institute) officially launched the 1,000 plant and animal reference genome project and calls for the collaboration from all over the world. The project aims to generate reference genomes for 1,000 economical and scientific important plant/animal species in two years.

- >200 Solexa High-Seq Machines each capable of producing >600 Gbases every 10 days for an output of >150 Terabases per month, or about >2-3 petabytes of sequence per year
- 3,000 staff of which 1,500 are working in bioinformatics
- 50,000 CPUs in their datacenter, $Ms per year power bill
Some Big Questions in Biology

- Are there still new life forms to be discovered?
- What role does life play in the metabolism/climate of planet earth?
- How do cells really work?
- What are the engineering principles of life?
- What is the information that defines and sustains life?
- What determines how organisms behave in their worlds?
- How much can we tell about the past - and predict about the future - by studying life on earth today?
Describe

Explain

Predict

Control
John Henry Woodger

Professor of Biology
University of London
Axiomatic Method in Biology 1937
Biology and Language 1952
John von Neumann

John von Neumann is well known for his work on self-replicating cellular automata. His ultimate goal, however, was to design self-replicating physical machines, and cellular automata were simply the first step towards the goal. Before his death, he had sketched some of the other steps. One step was to move away from the discrete space of cellular automata to the continuous space of classical physics.
Biological Science is Fundamentally Different from the Physical Sciences

- Dual Causality (Mayr)
  - Biology is more than physics
    - Differentiate inanimate and living processes
  - Physiochemical laws
    - Data light, reversible, time invariant
  - Genetic programs
    - Data rich, irreversible, time variant
  - Ecological and Social
    - Higher-order principles that govern the behavior of collections of autonomous entities

Ernst Mayr (1904-2005)
Reverse Engineering Living Systems: 
Genome + Environment = Phenotype

DNA

Gene Expression

Proteins

Transcription

Translation

Biochemical Circuitry

Phenotypes (Traits)

Environment

“The Central Dogma”

Adapted From Bruno Sobral VBI
Scale of Data (prok only, Euks 10-100x)

- 40K gns → ~40GB
- ~100 gns → 100 GB/prjt
  Total → 1PB

- 1K gnms @ 1000 conds
  Ea. cond @ ~200MB
  → 200GB
  Total 200 TB

- 1K gnms @ 1000 conds
  Ea cond ~> 1GB = 1TB
  Total ~1 PB

Environmental Sample Datasets growing 10x-100x faster
Genes → Proteins → Cell Networks → Cells → Populations → Communities → Ecosystems
The Theory/Modeling Ecosystem

Conceptual Models
- correspondence with theoretical concepts and principles

Functional Models
- behavioural/functional (input-output) correspondence

Generic Models
- abstract and generic correspondence with multiple phenomena

Mechanistic Models
- one-to-one correspondence

Hypothesis

Principles

Experimental and Data

Biological and Environmental Systems
The Theory/Modeling Ecosystem

Conceptual Models
- correspondence with theoretical concepts and principles

Principles
- e.g. Selection, Homeostasis, Stoichiometry, etc.

Generic Models
- abstract and generic correspondence with multiple phenomena

Functional Models
- behavioural/functional (input-output) correspondence

Hypothesis

Mechanistic Models
- one to one correspondence

- e.g. Protein Complexes, Active Sites, Ion Channels, Membrane Dynamics, etc.

Experimentation and Data
- e.g. Biochemical Reaction Network, Regulatory Network, Genotype/Phenotype etc.

- e.g. Species and Speciation, Guild and Biome, RNA World, LUCA, HGT, etc.

- e.g. Predator/Prey, Convergent Evolution, Neutral Models of Island Biogeography, etc.

Biological and Environmental Systems
An Integrated View of Modeling, Simulation, Experiment, and Bioinformatics

- Bioinformatics Analysis Tools
- Integrated Biological Databases
- Experimental Design
- High-Throughput Experiments
- Analysis & Visualization
Genomics is Powering the New Biology, but Computing is in the Drivers Seat

Experiments

- Genes, Proteins, RNAs, and other Biomolecules
- Large-scale Genome Sequencing
- Genome Sequences
- Tissue and Organismal Physiology
- Cellular & Developmental Processes
- Biochemical Pathways & Processes

Computation

- Predicting Three-Dimensional Structures of Proteins and RNAs
- Predicting Protein Sequence
- Predicting Functions
- Simulating and Understanding Gene Expression Networks
- Reconstructing Phylogeny, Homology, and Comparative Approaches
- Predicting Effects of Variation
- Structures of Multi-molecular complexes
- Predicting Catalysis, Molecular Dynamics
- Simulation of Metabolic and Signal Transduction Pathways
- Morphogenesis and Development
- Ecological Processes and Populations

Genomics is Powering the New Biology, but Computing is in the Drivers Seat
What do we need to advance theory and modeling in (Systems) Biology?

• Collect all observations, hypotheses and experimental results (i.e. capture all the relevant data)
• Classify, organize and integrate the data
• Devise means to search, query, analyze, compare, annotate and update the data
• Ability to build models, design experiments, create predictions (and simulate) and compare models derived from the data and produce updated models
• Capability to synthesize new principles, generalizations and abstractions from collections of models (i.e. theory)
• Enable collaboration in all of the above
Biology Databases (335 in 2001; 1,000+ in 2011)

- Major Seq. Repositories (7)
- Comparative Genomics (7)
- Gene Expression (19)
- Gene ID & Structure (31)
- Genetic & Physical Maps (9)
- Genomic (49)
- Intermolecular Interactions (5)
- Metabolic Pathways & Cellular Regulation (12)
- Mutation (34)
- Pathology (8)
- Protein (51)
- Protein Sequence Motifs (18)
- Proteome Resources (8)
- Retrieval Systems & DB Structure (3)
- RNA Sequences (26)
- Structure (32)
- Transgenics (2)
- Varied Biomedical (18)

annotation
mid-15c., from L. annotationem (nom. annotatio), from annotatus, pp. of annotare "to add notes to," from ad- "to" + notare "to note, mark”.

In genomics annotation involves parsing the DNA, determining features, functions and associations.

Ultimately we want to determine the relationship of every gene to the whole organism, population, community and environment.
Nature Reviews Genetics Reed et al. 2006
From Genotype to Phenotype
RAST workflow -- Rapid propagation

New paradigm for annotation:
reverse search: search for protein for a given prediction function
SEED/Model SEED Database

Model SEED

- Metabolic Model
  - Reaction
    - Compound
      - Publication
        - Culture Conditions
  - Biomass Reaction
    - Growth Phenotype

SEED

- Role Set
  - GPR Association
    - Identifier
      - Protein sequence
        - DNA sequences
          - Genome
            - Variant
              - Subsystem
                - Subsystem Class

Is Modeled By
Mapped To
Depends On
Named By
Named By
Named By
Named By
Named By
Is Role Of
Encodes protein For
Is Located In
Is Made Up Of
Is In Taxa
Is comprised of
Experimentally Observed For
Related to
Includes
Consistent with
Experiment level
Expression
Expression Experiment
Goto http://pubseed.theseed.org
SEED/Model SEED Database

SEED/RAST
• 3,000+ publically available genome annotations
• 30,000+ private genomes annotated
• 1000 subsystems
• 14,000 functional roles

Model SEED
• 3000+ publically available metabolic models (available this week through “reviewer” account)
• 10,000+ private models constructed
• 13,000 reaction
• 16,000 compounds
• 589 media conditions
Model SEED: Converting Annotated Genomes into Genome-scale Metabolic Models

RAST annotation server

Annotated genome in SEED

Preliminary reconstruction

Auto-completion

Analysis-ready models

Model accuracy

66%

71%

74%

82%

87%

22 optimized models

130 new metabolic models

• 965 reactions
• 688 genes
• 876 metabolites

19 Predicted gene

essentiality

19 Predicted growth media

19 Predicted phenotypes

19 Correction for 202 annotations inconsistent with essentiality data

Essential gene A

Essential Nonessential gene B

Gene essentiality consistency analysis

Model opt: GapFill

Model opt: GapGen

Optimized models

Predicting 69 missing transporters/model

Predicting 56 missing metabolic functions/model

Predicted cell-host interactions

Correcting reversibility constraints

Correcting missing and extra metabolic functions

Model: Converting Annotated Genomes into Genome-scale Metabolic Models
Reconstructing Models in High Throughput: Model SEED

• A biochemistry database was constructed combining content from the KEGG and 13 published genome-scale models into a non-redundant set of compounds and reactions

  **Acetinobacter**: iAbaylyiv4 (874 rxn)

  **B. subtilis**: iAG612 (598 rxn)

  **B. subtilis**: iYO844 (1020 rxn)

  **E. coli**: iAF1260 (2078 rxn)

  **E. coli**: iJR904 (932 rxn)

  **H. pylori**: iIT341 (476 rxn)

  **L. lactis**: iAO358 (619 rxn)

  **M. barkeri**: iAF692 (620 rxn)

  **M. genitalium**: iPS189 (263 rxn)

  **M. tuberculosis**: iNJ661 (975 rxn)

  **P. putida**: iJN746 (949 rxn)

  **S. aureus**: iSB619 (649 rxn)

  **S. cerevisiae**: iND750 (1149 rxn)

  **B. subtilis**: iYO844 (1020 rxn)

  **E. coli**: iJR904 (932 rxn)

  **H. pylori**: iIT341 (476 rxn)

  **L. lactis**: iAO358 (619 rxn)

• Reactions were then mapped to the functional roles in the SEED based on EC number, substrate names, and enzyme names:

<table>
<thead>
<tr>
<th>REACTION</th>
<th>COMPLEX</th>
<th>FUNCTIONAL ROLE</th>
<th>GENE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAD⁺ + NADPH ⇌ NADH + NADP⁺</td>
<td>Gene complex</td>
<td>NAD(P) transhydrogenase subunit beta (EC 1.6.1.2)</td>
<td>peg.100</td>
</tr>
</tbody>
</table>

  NAD(P) transhydrogenase alpha subunit (EC 1.6.1.2) | peg.101 |
To test growth of the model, we build a biomass objective function template:

**Nutrients**

- ATP + H₂O → ADP + Pi
- dATP, dGTP, dCTP, dTTP
- ATP, GTP, CTP, UTP
- Amino acids
- Cofactors and ions
- Various acylglycerols
- Peptioglycan
- Teichoic acid
- Core lipid A

**Energy**

- Universal

**DNA**

- Universal

**RNA**

- Universal

**Protein**

- Universal

**Lipids**

- Universal

**Cell wall**

- Any genome with cell wall
- Gram positive
- Gram negative

**Misc**

- Depends on genome

Each biomass component may be rejected from the biomass reaction of a model based on the following criteria:

- Subsystem representation
- Functional role presence
- Taxonomy
- Cell wall types

Reconstructing Models in High Throughput: Model SEED
Identifying Trends in Gene Expression

Fraction of datasets with role expressed

- svr_exp_genomes
- svr_genome_experiments
- svr_fraction_gene_expression
- svr_function_of

Genome list
Experiment list
Experiments where gene is “on”
Fraction where gene is “on”
Fraction where function is on
Identifying Trends in Gene Expression

Color Key for Subsystem Classes

- Amino Acids and Derivatives
- Carbohydrates
- Cell Division and Cell Cycle
- Cell Wall and Capsule
- Cofactors, Vitamins, Prosthetic Groups, Pigments
- DNA Metabolism
- Fatty Acids, Lipids, and Isoprenoids
- Membrane Transport
- Metabolism of Aromatic Compounds
- Motility and Chemotaxis
- Nitrogen Metabolism
- Nucleosides and Nucleotides
- Phosphorus Metabolism
- Potassium metabolism
- Protein Metabolism
- Regulation and Cell signaling
- Respiration
- RNA Metabolism
- Stress Response
- Sulfur Metabolism
- Virulence
- Miscellaneous

Genome list

-svr_exp_genomes

Experiment list

-svr_genome_experiments

Experiments where gene is “on”

-svr_fraction_gene_expression

Fraction where gene is “on”

-svr_function_of

Fraction where function is “on”
Identifying genes candidates for gaps

Genome

svr_gapfilled_rxns_and_roles

Gapfilled reactions

svr_find_clusters_for_rxn

Gene clusters near gap

svr_find_hypos_for_cluster

Gene candidates

svr_function_of

New annotation

A. *Tumefaciens*

R. *leguminosarum*

S. *Meliloti*

M. sp. *BNC1*

B. *Suis 1330*

Gene Clusters Associated with Reactions Surrounding Pathway Gap

Gapfilled step in pathway

rxn03130

Candidate hypothetical gene: peg.1726
Assembling Alignments and Trees

- **svr_align_seqs**: Aligned sequences
- **svr_trim_alignment**: Trimmed alignment
- **svr_psiblast_search**: Iterative BLAST search
- **svr_motif**: Conserved domains
- **svr_alni_to_html**: Concatenated motif
- **svr_tree**: Gene tree
- **svr_tree_to_html**: Tree painted with annotation

**Function**
- Protein of unknown function DUF1009 clustered with KDO2-Lipidgenes (38)
- Protein of unknown function DUF1009 (10)
- Hypothetical protein (8)
- Putative protein (5)
- Conserved hypothetical protein (1)
- Nucleic acid conserved (1)
- Oxidoreductase, Gfd/dh/Moca family (1)
- Others (1)

**Query gene**

**Current annotations of the hypothetical gene family**
Buchnera aphidcola sp.

100 copies of the genome per cell, lacks cell defense genes
Snapshot of one of ~30,000 complete genomes

/Buchnera_aphidicola_str_APS_Acrythosiphon_pismum
First few thousand base pairs of Buchnera

>Buchnera aphidicola str. APS (Acyrthosiphon pisum)

attttttgtattaaagacagtaatattataacaaaagagctaatattcattctcttctactcttctacagttgatcaagcaccgcaacatttttgattgaataagacgtaaattaatacaaaaaggctaatattctttatatattataactcataaatagattataattttatggtgtcagtaaaattttattttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttt
### Jobs Overview

The overview below lists all genomes currently processed and the progress on the annotation. To get a more detailed report on an annotation job, please click on the progress bar graphic in the overview.

In case of questions or problems using this service, please contact: rast@mcs.anl.gov.

#### Progress bar color key:
- not started
- queued for computation
- in progress
- requires user input
- failed with an error
- successfully completed

#### Jobs you have access to:

<table>
<thead>
<tr>
<th>Job</th>
<th>Owner</th>
<th>ID</th>
<th>Name</th>
<th>Num contigs</th>
<th>Size (bp)</th>
<th>Creation Date</th>
<th>Annotation Progress</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>15755</td>
<td>Henry, Christopher</td>
<td>3702.5</td>
<td>Arabidopsis thaliana</td>
<td>11</td>
<td>19695780</td>
<td>2010-10-04 11:48:55</td>
<td>[view details]</td>
<td>complete</td>
</tr>
<tr>
<td>15754</td>
<td>Henry, Christopher</td>
<td>3702.4</td>
<td>Arabidopsis thaliana</td>
<td>59</td>
<td>30263488</td>
<td>2010-10-04 11:44:31</td>
<td>[view details]</td>
<td>in progress</td>
</tr>
<tr>
<td>7383</td>
<td>Hicks, Christina</td>
<td>211586.8</td>
<td>Shewanella oneidensis MR-1</td>
<td>2</td>
<td>5131416</td>
<td>2009-09-19 17:56:26</td>
<td>[view details]</td>
<td>complete</td>
</tr>
<tr>
<td>7382</td>
<td>Bradley, Shirley</td>
<td>319224.15</td>
<td>Shewanella putrefaciens CN-32</td>
<td>1</td>
<td>4659220</td>
<td>2009-09-19 17:56:17</td>
<td>[view details]</td>
<td>complete</td>
</tr>
<tr>
<td>7381</td>
<td>boyd, kimberly</td>
<td>3992500.5</td>
<td>Shewanella woodyi ATCC 51908</td>
<td>1</td>
<td>5935403</td>
<td>2009-09-19 17:55:04</td>
<td>[view details]</td>
<td>complete</td>
</tr>
<tr>
<td>7376</td>
<td>Kieft, Aria</td>
<td>398579.6</td>
<td>Shewanella poaezae ATCC 51908</td>
<td>1</td>
<td>5174581</td>
<td>2009-09-18 11:49:54</td>
<td>[view details]</td>
<td>complete</td>
</tr>
</tbody>
</table>
ATP gamma subunit

250 organisms
From KEGG

ATP SYNTHESIS

Proton channel

E-type ATPase (Bacteria)

| beta | alpha | gamma | delta | epsilon | c | a | b |

3H^+

ATP Synthesis vs ATP-driven H^+ Pumping


mitochondria

chloroplast

bacteria

A top view of alpha3-beta3-gamma
By Heungjun Uhm & George Oster, U.C. Berkeley
<table>
<thead>
<tr>
<th>Genome ID</th>
<th>Organism</th>
<th>Variant Code</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>alpha</th>
<th>beta</th>
<th>gamma</th>
<th>delta</th>
<th>epsilon</th>
<th>l</th>
</tr>
</thead>
<tbody>
<tr>
<td>1769.1</td>
<td>Mycobacterium leprae [B]</td>
<td>1</td>
<td>1141</td>
<td>1143</td>
<td>1142</td>
<td>1145</td>
<td>1147</td>
<td>1146</td>
<td>1144</td>
<td>1148</td>
<td></td>
</tr>
<tr>
<td>233413.1</td>
<td>Mycobacterium bovis subsb. bovis AF2122/97 [B]</td>
<td>1</td>
<td>1333</td>
<td>1337</td>
<td>1336</td>
<td>1332</td>
<td>1334</td>
<td>1330</td>
<td>1328</td>
<td>1322</td>
<td></td>
</tr>
<tr>
<td>107090.2</td>
<td>Buchnera aphidicolae str. APS (Acyrthosiphon pisum) [B]</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>8</td>
<td>2</td>
<td>5</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>224915.1</td>
<td>Buchnera aphidicolae str. Bp (Brazoria picea) [B]</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>2</td>
<td>5</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>119804.1</td>
<td>Buchnera aphidicolae str. Sig (Schizaphis graminum) [B]</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>8</td>
<td>2</td>
<td>5</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>83131.1</td>
<td>Mycobacterium tuberculosis CDC1551 [B]</td>
<td>1</td>
<td>1383</td>
<td>1383</td>
<td>1384</td>
<td>1387</td>
<td>1386</td>
<td>1384</td>
<td>1386</td>
<td>1390</td>
<td></td>
</tr>
<tr>
<td>83132.1</td>
<td>Mycobacterium tuberculosis H37Rv [B]</td>
<td>1</td>
<td>1306</td>
<td>1308</td>
<td>1307</td>
<td>1310</td>
<td>1312</td>
<td>1311</td>
<td>1309</td>
<td>1313</td>
<td></td>
</tr>
<tr>
<td>159288.1</td>
<td>Staphylococcus aureus EMRSA-16 (Str. 252) [B]</td>
<td>1</td>
<td>873</td>
<td>871</td>
<td>872</td>
<td>869</td>
<td>867</td>
<td>868</td>
<td>870</td>
<td>874</td>
<td></td>
</tr>
<tr>
<td>93061.1</td>
<td>Staphylococcus aureus NCTC 8325 [B]</td>
<td>1</td>
<td>1454</td>
<td>1456</td>
<td>1455</td>
<td>1458</td>
<td>143,1460</td>
<td>1459</td>
<td>1457</td>
<td>1453</td>
<td></td>
</tr>
<tr>
<td>196620.1</td>
<td>Staphylococcus aureus subsp. aureus MW2 [B]</td>
<td>1</td>
<td>2023</td>
<td>2021</td>
<td>2022</td>
<td>2029</td>
<td>2027</td>
<td>2028</td>
<td>2030</td>
<td>2026</td>
<td>2034</td>
</tr>
<tr>
<td>158878.1</td>
<td>Staphylococcus aureus subsp. aureus Mn50 [B]</td>
<td>1</td>
<td>2109</td>
<td>2107</td>
<td>2108</td>
<td>2105</td>
<td>2103</td>
<td>2104</td>
<td>2106</td>
<td>2102</td>
<td>2110</td>
</tr>
<tr>
<td>176280.1</td>
<td>Staphylococcus epidermidis ATCC 12228 [B]</td>
<td>1</td>
<td>1206</td>
<td>1204</td>
<td>1203</td>
<td>1202</td>
<td>1200</td>
<td>1201</td>
<td>1203</td>
<td>1207</td>
<td></td>
</tr>
<tr>
<td>206435.1</td>
<td>Streptococcus agalactiae 2603W/R [B]</td>
<td>1</td>
<td>836</td>
<td>837</td>
<td>835</td>
<td>839</td>
<td>841</td>
<td>840</td>
<td>838</td>
<td>842</td>
<td></td>
</tr>
<tr>
<td>211110.1</td>
<td>Streptococcus agalactae NEM316 [B]</td>
<td>1</td>
<td>875</td>
<td>876</td>
<td>874</td>
<td>878</td>
<td>880</td>
<td>879</td>
<td>877</td>
<td>881</td>
<td></td>
</tr>
<tr>
<td>1336.1</td>
<td>Streptococcus equi [B]</td>
<td>1</td>
<td>1105</td>
<td>1106</td>
<td>1104</td>
<td>1108</td>
<td>1110</td>
<td>1109</td>
<td>1107</td>
<td>1111</td>
<td></td>
</tr>
<tr>
<td>210007.1</td>
<td>Streptococcus mitis UA159 [B]</td>
<td>1</td>
<td>1390</td>
<td>1389</td>
<td>1391</td>
<td>1387</td>
<td>1385</td>
<td>1386</td>
<td>1388</td>
<td>1384</td>
<td></td>
</tr>
<tr>
<td>714.2</td>
<td>Actinobacillus actinomycetemcomitans HK1651 [B]</td>
<td>1</td>
<td>1715</td>
<td>1717</td>
<td>1716</td>
<td>1719</td>
<td>1721</td>
<td>1720</td>
<td>1718</td>
<td>1722</td>
<td></td>
</tr>
<tr>
<td>228391.1</td>
<td>Actinobacillus pleuropneumoniae 1 str. 4074 [B]</td>
<td>1</td>
<td>818</td>
<td>820</td>
<td>819</td>
<td>822</td>
<td>824</td>
<td>823</td>
<td>821</td>
<td>825</td>
<td></td>
</tr>
<tr>
<td>181661.1</td>
<td>Agrobacterium tumefaciens str. C58 (Cerese) [B]</td>
<td>1</td>
<td>699</td>
<td>702</td>
<td>700</td>
<td>2520</td>
<td>2518</td>
<td>2519</td>
<td>2521</td>
<td>2517</td>
<td>698</td>
</tr>
<tr>
<td>180351.1</td>
<td>Agrobacterium tumefaciens str. C58 (U. Washington) [B]</td>
<td>1</td>
<td>699</td>
<td>702</td>
<td>700</td>
<td>2524</td>
<td>2522</td>
<td>2523</td>
<td>2525</td>
<td>2519</td>
<td></td>
</tr>
<tr>
<td>224324.1</td>
<td>Aquifex aeolicus VF5 [B]</td>
<td>1</td>
<td>172</td>
<td>1136</td>
<td>1137</td>
<td>170</td>
<td>512</td>
<td>1442</td>
<td>1444</td>
<td>1128</td>
<td>509</td>
</tr>
<tr>
<td>354.1</td>
<td>Azotobacter vinelandii [B]</td>
<td>1</td>
<td>1776</td>
<td>1778</td>
<td>1777,2560</td>
<td>1780</td>
<td>1782</td>
<td>2565</td>
<td>1781</td>
<td>2556</td>
<td>1779</td>
</tr>
<tr>
<td>191210.1</td>
<td>Bacillus anthracis str. A2012 [B]</td>
<td>1</td>
<td>716</td>
<td>714</td>
<td>715</td>
<td>712</td>
<td>710</td>
<td>711</td>
<td>712</td>
<td>717</td>
<td></td>
</tr>
<tr>
<td>198094.1</td>
<td>Bacillus anthracis str. Ames [B]</td>
<td>1</td>
<td>513</td>
<td>5137</td>
<td>5139</td>
<td>5135</td>
<td>5133</td>
<td>5134</td>
<td>5136</td>
<td>5132</td>
<td>1133,5149</td>
</tr>
<tr>
<td>86665.1</td>
<td>Bacillus halodurans [B]</td>
<td>1</td>
<td>3760</td>
<td>3759</td>
<td>3756</td>
<td>3754</td>
<td>3753</td>
<td>3755</td>
<td>3753</td>
<td>3753</td>
<td>3753</td>
</tr>
<tr>
<td>224308.1</td>
<td>Bacillus subtilis subsp. subtilis str. 168 [B]</td>
<td>1</td>
<td>4094</td>
<td>4092</td>
<td>4093</td>
<td>3690</td>
<td>3688</td>
<td>3689</td>
<td>3691</td>
<td>3687</td>
<td>3693</td>
</tr>
<tr>
<td>817.2</td>
<td>Bacteroides fragilis NCTC9343 [B]</td>
<td>1</td>
<td>1638</td>
<td>1640</td>
<td>1639</td>
<td>1642</td>
<td>1635</td>
<td>1643</td>
<td>1641</td>
<td>1636</td>
<td></td>
</tr>
<tr>
<td>226186.1</td>
<td>Bacteroides thetaiotaomicron VPI-5482 [B]</td>
<td>1</td>
<td>714</td>
<td>716</td>
<td>715</td>
<td>718</td>
<td>711</td>
<td>719</td>
<td>717</td>
<td>712</td>
<td></td>
</tr>
<tr>
<td>205913.1</td>
<td>Bifidobacterium longum DJ010A [B]</td>
<td>1</td>
<td>1550</td>
<td>1552</td>
<td>1551</td>
<td>1554</td>
<td>1556</td>
<td>1555</td>
<td>1553</td>
<td>1557</td>
<td></td>
</tr>
<tr>
<td>206672.1</td>
<td>Bifidobacterium longum NCC2705 [B]</td>
<td>1</td>
<td>534</td>
<td>532</td>
<td>533</td>
<td>530</td>
<td>328</td>
<td>329</td>
<td>331</td>
<td>327</td>
<td></td>
</tr>
</tbody>
</table>
Multiple Sequence Alignments
An Integrated View of Modeling, Simulation, Experiment, and Bioinformatics
An Integrated View of Modeling, Simulation, Experiment, and Bioinformatics
Knowledgebase enabling **predictive** systems biology.

- Powerful modeling framework.
- **Community-driven**, extensible and scalable **open-source** software and application system.
- Infrastructure for integration and reconciliation of algorithms and data sources.
- Framework for standardization, search, and association of data.
- Resource to enable **experimental design** and **interpretation** of results.
Engineering a Microbe for Biofuel Production

Annotated Genome

Feed Stock

Stresses
Hydrolysat, pH, Salt, End product, intermediates

DNA replication
transcription
protein folding
Regulation

Biomass
Isoprene

KBase Tool Integration

Annotation algorithms
Metabolic model generation
Regulatory network inference
Fitting kinetic model parameters
Proposing strain optimizations

Metabolic reconstruction
Model optimization algorithms
Other functional modeling
Predicting pathway fluxes

Genome Sequence
KEGG
Brenda
BioCyc
Published models

Gene KO Phenotypes

Comparative Genomics

Biolog

Metabolomics
Proteomics
Growth curves
Flux tracing experiments
KBase Data Integration

Transcriptomics
Modifying Lignin Biosynthesis

SNPs3D
PolyPhen

SNP influenced changes in protein structure and function
Pathway predictions
Model optimization
Model validation
Plant systems modification

• Genome annotation algorithms
• Comparative genomics
• Network inference
• Pathway reconstruction
• Omics & SNP overlay

Phylogenomics
Modeling phase I

Phenotype
Mutant population
Resequencing data
Transcriptomics
Proteomics
Metabolomics

NCBI
phytozome
iPlant Collaborative
BESC
Joint BioEnergy Institute
jbei
Culturing Recalcitrant Microbes from Communities

Population Statistics
- Isolate Genomes and Models
- Annotation and Metabolic Reconstruction
- Predict Syntrophic Interactions
- Isolate vs. Community Phenotype

Comparative Metagenomics
- Genome Assembly from Metagenomics
- Regulation and Functional Modeling
- Predict Culturing Conditions

Species Abundance
- Functional Gene Abundance
- Phylo binning and scaffolding
- Transcriptomics
- Proteomics

Temp
- pH
- Salinity
- Amino Acids
- Cofactors
- Syntrophies
Integrating Microbial Metabolism into Soil Ecology Models

Metagenomic data collection

Collecting samples

Sequencing

Sequence fragments

Associating fragments to taxonomical groups

Assembly of most prevalent microbes into complete genomes

Combining physiochemical descriptions of soil content and structure with microbial models in agent-based simulations

Integrating these models into a flux balance community model:

Forming flux balance models of individual microbe metabolism

Biomass

Air and water

CO₂ and organic matter

Soil nutrients

Air and water

CO₂ and organic matter

Soil nutrients

Air and water

CO₂ and organic matter

Soil nutrients

Air and water

CO₂ and organic matter

Soil nutrients

Air and water

CO₂ and organic matter

Soil nutrients

Air and water

CO₂ and organic matter

Soil nutrients
What the KBase Needs To Provide?

- Scalable compute and data capabilities beyond that available locally (100x – 1000x)
- Distributed infrastructure available 24x7 worldwide
- Integration with local databases and systems for seamless computing and data management
- Enables leverage of remote systems administration and support via service providers (EC2 compatible)
- Enables access to state of the art facilities at fraction of the cost (SPs just add more servers)
- Centralized support of tools and data
- Bottom line ⇒ enables biologists to focus on biology
Our vision is to put users in the drivers seat.
Leverage Existing Investments

• We leverage the considerable investments in existing integrated databases and analysis environments

• Key challenge: How we build on these systems yet provide to the community an integrated view for future development
Infrastructure Applications

Scientific Applications

Community Plug-ins

KB Unified Programming Interface API

Application Platform

Core Services APIs

SEED
Model Seed
MicrobesOnline
REGFamily
Phytozome
RAST
MGRAST

Public APIs
Protected APIs
Private APIs

Community development
Kbase development
Core Services
Storing a diverse representation of biological data ranging from highly structured data in relational databases to large bulk data to frequently generated and changing user data entails multiple approaches.

- **Petabytes of Raw Data**
- **Flexible Storage For Workspaces**
- **Structured Storage For Curated Data**

**The KBase Infrastructure and Services**
The KBase Infrastructure and Services

Bulk Data Services  Persistent Store  Central Data Store

Data Loads, Caching, Replication, Availability, Performance, Integration

Kbase Core Services

Microbes Online  Meta Microbes Online  RegFam  SEED  RAST  Model SEED  MG-RAST  Phytozome  External Cores  Future KBase Cores
Building a KBase Unified API

- Uniform view of KBase core services
- Access to data management, security and computational services in the KBase infrastructure
- Rich set of standard graphics and UX components for building web applications with single look and feel

![Diagram showing KBase Unified Application Programming Interface (API) with sections for Integration Services, Low-Level System Services, and User Interfaces and Graphics]
The KBase Infrastructure and Services

Bulk Data Services
Persistent Store
Central Data Store
The KBase Unified API is based on a service-oriented architecture. SoA.

In a SoA the system is functionally decomposed into many services each of which is implemented as one or more servers. The API provides a single view to all the capabilities of KBase.

Our long-term goal includes community developed and contributed services. Our initial set of services will be backed by the following example servers:
Workflows from each of the three science domains (microbes, communities, and plants) used to demonstrate the capabilities provided by the first production release.
Figure 1. Four basic types of evolutionary events

Figure 4. Visualized reconciled trees: HKY85 data
EXAMPLE WORKFLOW

SEARCHING FOR CANDIDATES
OF HORIZONTAL GENE TRANSFER

\( G \) is a gene/protein index (think coherent Protein Family or COG)
\( O \) is set of organisms (we have 3000 now - will have 10,000 in three years)
\( X \) is a set of conserved Genes/Proteins common to many \( O \) (order hundreds)

\( dnaSeq[G] \leftarrow \) Extract Coding DNA Sequences for \( X \) from some \( O \)
\( proteinSeq[G] \leftarrow \) Translate DNA to Protein for Each \( dnaSeq[G] \)
\( Bbhs[G] \leftarrow \) Compute Bidirectional Best Hits for Each \( proteinSeq[G] \)
\( TC[G] \leftarrow \) Compute Transitive Closure of \( Bbhs[G] \)
\( MSA[G] \leftarrow \) Compute Multiple Sequence Alignment of Each \( TC[G] \)
\( GeneTree[G] \leftarrow \) Compute Maximum Likelihood Tree for Each \( MSA[G] \)
\( OrgTree \leftarrow \) Compute Organism tree* for \( O \)
\( ReconTrees[G] \leftarrow \) Reconcile \( OrgTree \) with each \( GeneTree[G] \)
\( Candidates[G] \leftarrow \) Score and Sort \( ReconTrees[G] \)
\( TreeGraphs[G] \leftarrow \) Compute Readable Reconciled Trees for \( Candidates[G] \)
• ESnet backbone (ESnet4) is a national nx 10 Gbps optical circuit infrastructure

• ESnet shares its optical network with Internet2

• ESnet's IP network functions as a Tier 1 internet service provider

KBase leverages ESNet for 10+ Gb/s data transfer between all nodes
The DOE KBase Cloud

**Built on DOE ASCR Originated Magellan Cloud Infrastructure**

- Open Stack Cloud @ Argonne
- Open Stack Cloud @ Oak Ridge
- Cluster system @ Berkeley
- Cluster system @ Brookhaven
- 3 Petabytes of Storage

**Argonne KBase Magellan Hardware**

> 700 nodes (> 12,000 cores) for KBase
The KBase Cloud Architecture

- Commodity Compute Cluster Hardware
- OpenStack IaaS Cloud Software Stack (EC2/S3 APIs)
- Data Intensive Science
- Method Development
- KBase Application Development
- Large Scale Computation
- KBase Images
- Ubuntu Image
- MapReduce Image
- HPC Cluster Image

Large Scale Computation

HPC Cluster Image

MapReduce Image

Ubuntu Image

KBase Images
The Release Schedule for Kbase

Feb 2012 – Development release (internal target)
- debug release engineering, prototype deployments, initial data models and data loads, non-unified API, performance testing, architecture refinement

May 2012 – Alpha release (internal target, invited testing)
- draft tutorials, v0.0 database loads, draft API (performance and ubiquity unified prototypes), draft UI library, domain workflow drafts, cloud and cluster services

Aug 2012 – Beta Release (early adopter beta testing)
- workflow function complete, API refinement, v1.0 database loads, prototype plug-in interfaces, prototype galaxy support, performance debugging

Nov 2012 – Production Release Candidate (public beta testing)
- draft website, draft documentation, full functionality API, draft UI v1.0, database loads v2.0, significant number of external beta test users

Feb 2013 – KBase Production Release
- public website, unified API, initial production UI, database loads v3.0 (microbes, community, plant databases), production demonstration workflows, and replication
Large-Number of Bioinformatics Tools and Environments $O(100)$ core codes

- Raw sequencing clean up
- Assembly (prok, euk, meta), Gene Calling (prok)
- Comparative analysis (basic tools, e.g. BLAST, BLAT, etc.)
- Protein family maintenance and decision procedures
- MG Fragment id and function/taxa inference (e.g. MG-RAST)
- Annotation, reconstruction (e.g. RAST, k-mers, etc)
- Multiple sequence alignment (e.g. T-COFFEE, Clustalw)
- Statistical analysis (k-mer spectra, co-occurrences, etc.)
- Sequence data mining (e.g. motifs, meme, ncRNA)
- Evolutionary analysis (16S, phylogenetic trees)
- Large-scale data integrations (SEED, IMG, MO, etc.)
- Systems level integrations (e.g. Kbase modeling environments)
Three ideas that will change bioinformatics

• Alignment Free Methods
  • Inferring relationships between sequences by computing statistical measures of small words
  • Better scaling and can use hardware acceleration

• Computing on Compressed Data (Directly)
  • Most sequence data can be compressed (esp raw reads)
  • Can we develop methods that compute directly on these compressed representations

• Representing Sequences as Graphs with Probabilities
  • Most sequence data is inferred from assemblies of short reads
  • During assembly many alternatives are considered and heuristics used to discard all but one representation
The whole-proteome FFP (Feature Frequency Profile) tree at feature length $l = 13$. 

Jun S et al. PNAS 2010;107:133-138
Reference Based Compression

- Align reads to reference sequences
- Encode mapping
- Encode variation
- For unmapped reads, assemble and align

.02 bits/base compression ratio achieved in some cases (1000x improvement)

(Genome Research 2011, Fritz, et. al.)
Figure 1
The ALLPATHS assembly of *S. aureus*. Each edge represents a contiguous and unambiguous sequence of bases and, for this assembly, each component is its own scaffold. Longer edges are in red, short edges in gray. The sizes of the gray edges and regions are in bases. Several key features are called out in blue boxes. Five short sequences totaling 9 kb are not shown. Images of the graphs for all five ALLPATHS assemblies of this paper are available at [16].
LESSONS LEARNED
Data Intensive Computing Usage Models Differ from Existing Simulation Usage Models

**Data Intensive**
- Continuous access require based on data generation/submission rates
- CPU time, I/O and data volume all important
- Data products typically used in future computations via an integration or pipeline
- Data products made available for external users and curated over time

**Simulation**
- Batch oriented access based on allocations for specific projects
- CPU time centric
- Output not necessarily used in future runs but often significant time used for visualization
- Output generally (but not always) used “privately” and rarely curated
Management Policies Need to be Different

- Long term (many years) access commitment at a continuous or increasing level of service
- Support for persistent services
- Storage allocation that grows over time
- Rich software environment with high-performance database support
- Mechanism to publish the data to a community
- Archival support for data, links and citations
What is Important?

• While the data volumes are increasing in biology (sequencing will soon be pushing many PBs per year) it is complexity, errors, inconsistencies and incomplete knowledge that are at least as challenging as scale

• Scale of raw data is increasing faster than we are closing the loop on understanding

• This implies that the strategies for data analysis and integration are changing rapidly

• It also means that new sources of high-throughput data many force depreciation of many lower throughput data sources (keeping older less accurate data is questionable)
What is Important for Biology?

• Pushing forward on core capability development (e.g. reconstruction of regulatory networks, assembly of genomes from metagenomes, automated model generation)

• Integration of data and tools to enable higher level investigations to build on lower levels (e.g. reasoning over all genomes, protein families, trees, etc.)

• Packaging capabilities as services such that communities with little technical skill but significant scientific skill can exploit them
  • Examples (NCBI BLAST, RAST, ModelSEED and MG-RAST)
The Trends

• Genomics data will continue to decline in cost
  • Detectors, imaging, protons, ph, etc.
  • Microfluidics and micropatterning
  • Chemistry control
  • Molecular placement and manipulation
• Imaging technologies (improved resolution, performance, interpretation, etc.)
• Robotics (lower cost, improved programmability, liquids, solids, etc.)
• Microfluidics (lower volumes, improved control, coupling to electronics, improved integration, scale up)
• Single Cell manipulation, Soft tissue manipulation
• Genomification and Moore’s lawification
Trends in Questions

• Whole Genome vs One Gene
• Knock out all genes rather than one
• Pan-genome vs One Genome
• All Proteins in a Family vs one Protein
• All taxa in a clade vs one organism
• All unknown functions vs target function
• All known/unknown \{X\} vs an element of \{X\}
• Etc.
Adam has autonomously generated functional genomics hypotheses about the yeast *Saccharomyces cerevisiae* and experimentally tested these hypotheses by using laboratory automation.

*Science 2009 King, et. al.*
Examples of B-Factory Runs

- Test knockouts of all 5000 genes in a bacterium for essentiality and enhanced growth on 200 media types
- Screen 100,000 genetic variants of a protein for enhanced biofuel activity
- Isolate single cells from 1,000 soil samples containing given protein or DNA markers
- Isolate DNA from 1,000 soil samples for DNA sequencing
- Clone, express and test 1,000 unknown proteins against 10,000 functional assays
- Collect phenotype data for 500 bacterial taxa for 2000 growth conditions
Acknowledgements

• Many thanks to DOE, NSF, NIH, DOD, ANL, UC, Moore Foundation, Sloan Foundation, Apple, Microsoft, Cray, Intel and IBM for supporting our research groups over the years
Acknowledgements

• Many many people are working to build the things I’ve talked about today.. I can’t mention them all but you should know about a few

• Ross Overbeek, Chris Henry, Fangfang Xia, Folker Meyer, Tom Brettin, Bob Olson, Terry Disz, Andreas Wilke, Sam Seaver, Veronika Vonstein, Scott Devoid, Gordon Pusch, Bruce Parello, Jared Wilkens, Adam Arkin, Daniel Quest, Chris Bun, Jennifer Salazar, Elizabeth Glass, Shane Canon, Narayan Desai, Matt Dejongh, Aaron Best, Tobias Paczian, Peter Larson and many more
Thank You for Listening
Modeling in Microbial Ecology

• Simon Levin in his McArthur Award Lecture (Levin, 1992) on ‘The problem of pattern and scale in ecology’ described the essence of modeling thus:

• ‘to facilitate the acquisition of this understanding (scaling), by abstracting and incorporating just enough detail to produce observed patterns. A good model does not attempt to reproduce every detail of the biological system; the system itself suffices for that. Rather, the objective of a model should be to ask how much detail can be ignored without producing results that contradict specific sets of observations...’
Models of Microbial Communities

- Community Structure (static and time series)
  - Species Abundance (distributions, model free, etc.)
  - Assemblages (island biogeography theory, neutral models)
  - Diversity Measures ($\alpha$, $\beta$, $\gamma$)
- Functional Profiles
  - Gene functions in a community
- Community (mixed) Metabolic Network
- Community (compartmentalized) Metabolic Network
- Dynamic Agent Based Models of Communities
- Predator-Prey Ecosystems Models
Estimates of Global Diversity

\[ \frac{N_T}{N_{\text{max}}} \sim 10 \text{ for soil} \]

\[ \frac{N_T}{N_{\text{max}}} \sim 4 \text{ for aquatic} \]

\( N_T/N_{\text{max}} \) method estimates global diversity between \( 10^6 \) and \( 10^9 \).

Figure 2. A ‘quick and dirty’ way to estimate diversity by assuming a distribution. (a) The total number of taxa in a community with a lognormal species abundance curve is simply the area under that curve (called the species curve). The individuals curve is the number of species at each abundance (the species curve) multiplied by their abundance (the x-axis). There is therefore a mathematical relationship between the area under a species area curve, the number of individuals \( N_T \) (the area under the individuals curve), and the maximum and the minimum abundance (\( N_{\text{max}} \) and \( N_{\text{min}} \)). (b) The relationship, over 30 orders of magnitude in population size, for various ratios of \( N_T/N_{\text{max}} \) by assuming that \( N_{\text{min}} \) is equal to one (Curtis et al. 2002). As a rule of thumb, soil has a ratio of 10 and seas and lakes have a ratio of 4. There are about \( 10^{30} \) bacteria in the world (Whitman et al. 1998).

Mechanisms of Microbial Species Diversity

Figure 5 | Molecular evolutionary mechanisms that shape bacterial species diversity: one genome, pan-genome and metagenome. Intra-species (a), inter-species (b) and population dynamic (c) mechanisms manipulate the genomic diversity of bacterial species. For this reason, one genome sequence is inadequate for describing the complexity of species, genera and their inter-relationships. Multiple genome sequences are needed to describe the pan-genome, which represents, with the best approximation, the genetic information of a bacterial species. Metagenomics embraces the community as the unit of study and, in a specific environmental niche, defines the metagenome of the whole microbial population (d).
Mechanisms of Microbial Species Diversity

Figure 5: Molecular evolutionary mechanisms that shape bacterial species diversity: one genome, pan-genome and metagenome. Intra-species (a), inter-species (b) and population dynamic (c) mechanisms manipulate the genomic diversity of bacterial species. For this reason, one genome sequence is inadequate for describing the complexity of species, genera and their inter-relationships. Multiple genome sequences are needed to describe the pan-genome, which represents, with the best approximation, the genetic information of a bacterial species. Metagenomics embraces the community as the unit of study and, in a specific environmental niche, defines the metagenome of the whole microbial population (d).
Island Biogeography View of Microbial Community Structure

Pedros-Alio 2006
**Rank Abundance Plot**

Redraw the below plot using the following taxonomic level: **family**

The plot below shows the family abundances ordered from the most abundant to least abundant. Only the top 50 most abundant are shown. The y-axis plots the abundances of annotations in each family on a log scale.

The rank abundance curve is a tool for visually representing taxonomic richness and evenness.

Download chart data

The image is currently dynamic. To be able to right-click/save the image, please click the static button **static**.
Rarefraction Curve

The plot below shows the rarefaction curve of annotated species richness. This curve is a plot of the total number of distinct species annotations as a function of the number of sequences sampled. On the left, a steep slope indicates that a large fraction of the species diversity remains to be discovered. If the curve becomes flatter to the right, a reasonable number of individuals is sampled: more intensive sampling is likely to yield only few additional species.

Sampling curves generally rise very quickly at first and then level off towards an asymptote as fewer new species are found per unit of individuals collected. These rarefaction curves are calculated from the table of species abundance. The curves represent the average number of different species annotations for subsamples of the complete dataset.

Download chart data

The image is currently dynamic. To be able to right-click/save the image, please click the static button static
Comparing Soil Metagenomes

Grass vs Bare Fallow sites – Abundance of shared Taxa

- Grass: 8,000 OTU, 2500:1
- Bare Fallow: 12,000 OTU, 1600:1

> 5,000 shared OTUs
In the world of microbial ecology, we need theory very badly.

Almost any consequential microbial community will have $10^{10}$ to $10^{17}$ bacteria that could compose more than $10^7$ differing taxonomic groups and countless functional groups.

It seems remarkable that we should even contemplate trying to understand such vast systems without recourse to some form of theory.

Curtis and Sloan Science 2005
Neutral Model for Community Assemblage

Battin et al. Nature Reviews Microbiology 5, 76-81 (January 2007)
Island Biogeography Model

\[ \alpha - \text{diversity} \]

\[ \beta - \text{diversity} \]
Dormancy in Different Environments

Lennon and Jones, Nature Reviews Microbiology 2011
Island Biogeography Model with Dormancy
Chaos in a Microbial Food Web

a, b – stable equilibrium

c-g – chaotic > 0 Lyapunov exp

h, I – limit cycles

Becks et al Nature 2005
Synthetic Predatory Prey Model

Engineering Microbial Consortia
Brenner, You and Arnold Trends in Biotechnology 2008
How stable are microbial communities?

Resistance
Resilience
Redundancy

Allison and Martiny PNAS 2008
Functional Profiles (From MG-RAST)

What gene functions are present in a sample
Relative Metabolic Flux – Community Level Prediction

• Predicting the metabolome from metagenomics data!

• RMF returns a list of metabolites and whether those metabolites are more or less likely to be consumed or synthesized in one environment relative to another.

• When linked to Model-SEED – provides information relevant for ecologists
Modeling Mutualistic Microbial Community

Stolyar et al, Molecular Systems Biology 2007
Integrating Microbial Metabolism into Soil Ecology Models

Metagenomic data collection

- Collecting samples
- Sequencing
- Sequence fragments
- Associating fragments to taxonomical groups
- Assembly of most prevalent microbes into complete genomes

Combining physiochemical descriptions of soil content and structure with microbial models in agent-based simulations

- Integrating these models into a flux balance community model:
  - Air and water
  - CO₂ and organic matter
  - Soil nutrients
  - Forming flux balance models of individual microbe metabolism

Biomass

ACGGCGTTAGATATATCGATCGATCGATGCTATATACGCGTAGTACTGATCGTACGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTA
The KBase Services Mappings

Cloud Services) Support
Argonne, Oak Ridge
Cluster Services Support
NERSC, Brookhaven
**Concept:** Kbase User Experience

**Search**

- Sensory genes: gs_aeb123456
- Flagellar genes: gs_aeb123457

**Narrative**

  (* Bot the guy says I missed one. I have looked and by eye I agree *)
  Add(Sensory genes, gi0123456) - Sensory genes

- [in5] 10:00:28pm 11/16/2011
  I need to figure out in which conditions these genes are expressed. First I am going to aggregate my two sets of genes (I separate them for differential analysis later), then query for all gene expression data concerning them. Hmmm... what's that function again?
  Merge(Sensory genes, Flagellar genes) - GetExpress(geneids:stdin) - Add(Expression_data, stdin)

  ClusterMe(Expression_data) - PickCluster()

  ClusterMe(Expression_data) - PickCluster()

- Add(High Expression, stdin)

**Data management**

- upload

**Team management**

- Metal reduction project
- Chemotaxis project

**Narrative management**

- Chemotaxis Study

**Blog**

C: ClusterMe Clustering

...
Concept: Interactive community knowledge

Narrative Graphs Provide Measures of Activity and Influence

Function and Data contents can be tracked to assess “use”

Narrative Code Versioning and Scenario Branching

Narrative Query Update

Narrative Data Change

Project Linkage and Citation