ASCAC-BERAC Joint Subcommittee on Computational Biology and GTL

Rick Stevens
John Wooley
Co-chairs
The Subcommittee Charge

- Convene a joint panel with BERAC to examine the issue of computational models for GTL, including:
  - How progress could be accelerated through targeted investments in applied mathematics, and
  - How computer science can be incorporated to meet the needs of computational biology.
- The joint panel should consider whether the current ASCR long-term goal is too ambitious, given the status and level of buy-in from the community.
- It needs to consider what is happening in the computational-science and life-sciences communities. It should discuss possible intermediate goals that might be more relevant to the two programs.
- And it should identify the key computational obstacles to developing computer models of the major biological understandings necessary to characterize and engineer microbes for DOE missions, such as biofuels and bioremediation.
Joint Subcommittee Members

- Rick Stevens, Argonne-Uchicago (co-chair)
- John Wooley, UCSD (co-chair)
- Barbara Wold, Caltech
- David Galas, Battelle and ISB
- Thomas Zacharia, Oak Ridge-UT
- Michael Banda, Berkeley Lab
- Virginia Torczon, William and Mary
- David Kingsbury, Moore Foundation
- Chris Somerville, Carnegie Institution
Overall Plan of Attack

- Organizational and Framing Meeting (8/6/07)
- Teleconference (mid 9/07)
- Two Day Community Input Meeting at Planned for early October at the Moore Foundation in the Bay Area (~10/04/07)
- Writing Meeting/Teleconference (mid 10/07)
- Report Concurrence Teleconference (11/07)
- Reporting out at the November ASCAC and BERAC meetings.
Observations from the First Meeting

• Both BER and ASCR need a common set of explicit science goals to drive advances in bioinformatics, computational and theoretical biology and the associated high-throughput experimental techniques in systems biology

• The existing PART Performance Target is not an ideal goal as currently stated since it is somewhat vague and not focused on a scientific outcome

• Also progress towards the existing PART Performance Target is also not easily measured

• The recent shifting of the BER agenda to nearer term bioenergy research may make it more difficult to focus on basic joint goals
Genes → Proteins → Cell Networks → Cells → Populations → Communities → Ecosystems
Microbial Cell Modeling an Example of one Approach

- There are many groups building models of cells at different levels of abstraction
- Significant progress has been made in the last 17 years on extending flux balance methods to whole genomes
- The state-of-the-art reconstruction now boasts over 1200 reactions incorporated into the model and is now covering nearly 70% of metabolic genes and increasing fraction of other cellular functions
- Soon it will be possible to semi-automatically produce 100’s of reconstructions from existing microbial genomes
E. coli K-12 Metabolic Overview

Source: EcoCyc
17 Years of Progress in FBA Model Development

Molecular Systems Biology 3 Article number: 121
doi:10.1038/msb4100155
Published online: 26 June 2007
Prokaryotic Intracellular Environment – Gel Like Media

- 100 nm$^3$
- 450 proteins
- 30 ribosomes
- 340 tRNA molecules
- Several long mRNAs
- 30,000 small organic molecules
- 50,000 Ions
- Rest filled with water 70%

From: David Goodsell, The Machinery of Life
Hierarchical Modeling in Biological Systems

- Genetic Sequences
- Molecular Machines
- Molecular Complexes and modules
- Networks + Pathways [metabolic, signaling, regulation]
- Structural components [ultrastructures]
- Cell Structure and Morphology
- Extracellular Environment
- Populations and Consortia etc.
In Quest of the Minimum Genome

- What are the Smallest number of genes needed to create a viable organism?
  - Free living on a rich, but defined culture medium
- Experimentally determined essential genes
  - Bacillus subtilis ~300 CDS
  - Escherichia coli ~400 CDS
- Reduced organisms in nature
  - Mycoplasma ~500
  - Nanoarchaea ~400
- Bioinformatics predicts a conserved core
  - ~200-400 CDS
Table 1

Genome-scale gene essentiality studies in bacteria.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Mutagenesis</th>
<th>Mutant outgrowth</th>
<th>Readout</th>
<th>ORFs total</th>
<th>N</th>
<th>E</th>
<th>%E</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>M. genitalium</em></td>
<td>Random</td>
<td>Clones</td>
<td>Sequencing</td>
<td>2600</td>
<td>n/a</td>
<td>168</td>
<td>n/a</td>
<td>[2]</td>
</tr>
<tr>
<td><em>S. aureus WCUH29</em></td>
<td>Random</td>
<td>Clones</td>
<td>Sequencing</td>
<td>2892</td>
<td>n/a</td>
<td>658</td>
<td>23%</td>
<td>[3]</td>
</tr>
<tr>
<td><em>S. aureus RN4220</em></td>
<td>Random</td>
<td>Population</td>
<td>Footprint-PCR</td>
<td>1657</td>
<td>602</td>
<td>670</td>
<td>40%</td>
<td>[5]</td>
</tr>
<tr>
<td><em>H. influenzae Rd</em></td>
<td>Random</td>
<td>Population</td>
<td>Colony formation</td>
<td>2043</td>
<td>560</td>
<td>133</td>
<td>n/a</td>
<td>[4]</td>
</tr>
<tr>
<td><em>S. pneumoniae Rx-1</em></td>
<td>Targeted</td>
<td>Clones</td>
<td>Colony formation</td>
<td>2043</td>
<td>560</td>
<td>133</td>
<td>n/a</td>
<td>[13]</td>
</tr>
<tr>
<td><em>S. pneumoniae D39</em></td>
<td>Targeted</td>
<td>Clones</td>
<td>Colony formation</td>
<td>2043</td>
<td>560</td>
<td>133</td>
<td>n/a</td>
<td>[13]</td>
</tr>
<tr>
<td><em>M. tuberculosis H37Rv</em></td>
<td>Random</td>
<td>Population</td>
<td>Microarray</td>
<td>3989</td>
<td>2567</td>
<td>614</td>
<td>15%</td>
<td>[6*]</td>
</tr>
<tr>
<td><em>B. subtilis 168</em></td>
<td>Targeted</td>
<td>Clones</td>
<td>Colony formation</td>
<td>4105</td>
<td>3830</td>
<td>271</td>
<td>7%</td>
<td>[7**]</td>
</tr>
<tr>
<td><em>E. coli K-12 MG1655</em></td>
<td>Random</td>
<td>Population</td>
<td>Footprint-PCR</td>
<td>4308</td>
<td>3126</td>
<td>620</td>
<td>14%</td>
<td>[8]</td>
</tr>
<tr>
<td><em>E. coli K-12 MG1655</em></td>
<td>Targeted</td>
<td>Clones</td>
<td>Colony formation</td>
<td>4308</td>
<td>2001</td>
<td>n/a</td>
<td>n/a</td>
<td>[12]</td>
</tr>
<tr>
<td><em>E. coli K-12 BW25113</em></td>
<td>Targeted</td>
<td>Clones</td>
<td>Colony formation</td>
<td>4390</td>
<td>3985</td>
<td>303</td>
<td>7%</td>
<td>[15**]</td>
</tr>
<tr>
<td><em>P. aeruginosa PAO1</em></td>
<td>Random</td>
<td>Clones</td>
<td>Sequecing</td>
<td>5570</td>
<td>4783</td>
<td>678</td>
<td>12%</td>
<td>[9]</td>
</tr>
<tr>
<td><em>P. aeruginosa PA14</em></td>
<td>Random</td>
<td>Clones</td>
<td>Sequencing</td>
<td>5688</td>
<td>4469</td>
<td>335</td>
<td>6%</td>
<td>[16*]</td>
</tr>
<tr>
<td><em>S. typhimurium</em></td>
<td>Random</td>
<td>Clones</td>
<td>Sequencing</td>
<td>4425</td>
<td>n/a</td>
<td>257</td>
<td>~11%</td>
<td>[10]</td>
</tr>
</tbody>
</table>

This table provides a short summary. We refer the reader to the Supplementary material (Table S1) for details. Genome-wide screens for genes essential for virulence are beyond the scope of this review and are not listed here. The complete gene essentiality datasets obtained in these studies have been incorporated in the SEED (http://theseed.uchicago.edu/Fig/index.cgi) and National Microbial Pathogen Data Resource (NMPDR; http://www.nmpdr.org/) genomic databases. ORFs total, an estimate of a total number of protein-encoding genes in a genome; N, genes detected as nonessential; E, deemed essential for survival (for datasets generated via clonal strategy) or essential for fitness (for datasets generated via populational screens); %E, fraction of essential genes in a genome.

a Mutant collection is available for public distribution.
b Direct essentiality screening methods (e.g., via antisense RNA) do not provide information about nonessential genes.
c Only partial dataset is available.
d This list also includes predicted gene essentiality and data compilation from published single-gene essentiality studies.
e Project in progress.
f Deduced by comparison of the two gene essentiality datasets obtained independently in the *P. aeruginosa* strains PA14 [16*] and PAO1 [9].
Modeling of Simple Microbial Communities

• Many environmental metagenomics projects have identified modest sized microbial communities in novel environments
  • Dozens of Studies have been done and hundreds are being planned
• Reconstruction of these communities (genomic, proteomic, metabolic) has begun
  • When the diversity is small, it is possible to essentially sequence the dominate organisms without culturing
• In addition there are well studied model communities that can serve as laboratory models for development (e.g. Winogradsky columns, biofilms, etc.)
## Acid Mine Drainage Sediment Community

### Table 2: BLAST analysis of 16S rDNA sequences of acidophiles in the sediment

<table>
<thead>
<tr>
<th>Representative sequence</th>
<th>Length (bp)</th>
<th>Frequency</th>
<th>Microbial group affiliation</th>
<th>Closest relative (Genebank accession number)</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>H5</td>
<td>1461</td>
<td>4.4%</td>
<td><em>Acidobacteria</em></td>
<td>Uncultured bacterium (AB179509)</td>
<td>1378/1462 (94%)</td>
</tr>
<tr>
<td>H6</td>
<td>1494</td>
<td>4.4%</td>
<td><em>γ-Proteobacteria</em></td>
<td>Uncultured bacterium clone 1013-28-CG34 (AY532575)</td>
<td>1423/1492 (95%)</td>
</tr>
<tr>
<td>H11</td>
<td>1517</td>
<td>51.1%</td>
<td><em>δ-Proteobacteria</em></td>
<td>Uncultured bacterium BA71 (AF225447)</td>
<td>1418/1451 (97%)</td>
</tr>
<tr>
<td>H12</td>
<td>1521</td>
<td>6.7%</td>
<td><em>Nitrospira</em></td>
<td>Uncultured bacterium clone ASL9 (AF544226)</td>
<td>1476/1515 (97%)</td>
</tr>
<tr>
<td>H24</td>
<td>1502</td>
<td>4.4%</td>
<td><em>β-Proteobacteria</em></td>
<td>Uncultured bacterium clone DSBACT9 (AY762628)</td>
<td>1030/1109 (92%)</td>
</tr>
<tr>
<td>H40</td>
<td>1432</td>
<td>2.2%</td>
<td>Candidate Division TM7</td>
<td>Uncultured soil bacterium clone Cl29 (AF507687)</td>
<td>1297/1411 (91%)</td>
</tr>
<tr>
<td>H50</td>
<td>1520</td>
<td>6.7%</td>
<td><em>Nitrospira</em></td>
<td>Uncultured bacterium (DQ23212)</td>
<td>1473/1517 (97%)</td>
</tr>
<tr>
<td>H65</td>
<td>1500</td>
<td>4.4%</td>
<td><em>γ-Proteobacteria</em></td>
<td><em>Acidithiobacillus ferrooxidans</em> strain QXS-1 (DQ168465)</td>
<td>1491/1498 (99%)</td>
</tr>
<tr>
<td>H70</td>
<td>1491</td>
<td>2.2%</td>
<td>Low G + C Gram-positives</td>
<td>Uncultured Low G + C Gram-positive bacterium</td>
<td>782/799 (97%)</td>
</tr>
<tr>
<td>H74</td>
<td>1484</td>
<td>13.3%</td>
<td><em>δ-Proteobacteria</em></td>
<td></td>
<td>967/992 (97%)</td>
</tr>
</tbody>
</table>
Yellowstone Microenviroments

Geobiology (2005), 3, 211–227
Extending the Models to Include the Environment is Key to Progress

Diagram:
- Extracellular Organic Material
- Bacterium
- Cr, Pb, Hg, U
- (Bio)mineralization Products
- Mineral Surface
Emergent Biogeography of Microbial Communities in a Model Ocean

Michael J. Follows,1* Stephanie Dutkiewicz,1 Scott Grant,1,2 Sallie W. Chisholm3

Fig. 1. Annual mean biomass and biogeography from single integration. (A) Total phytoplankton biomass (μM P, 0 to 50 m average). (B) Emergent biogeography: Modeled photo-autotrophs were categorized into functional groups; color-coding is according to group locally dominating annual mean biomass. Green, analogs of Prochlorococcus; orange, other small photo-autotrophs; red, diatoms; and yellow, other large phytoplankton. (C) Total biomass of Prochlorococcus analogs (μM P, 0 to 50 m average). Black line indicates the track of AMT13.

Fig. 2. Observed and modeled properties along the AMT13 cruise track. Left column shows observations (27), right column shows results from a single model integration. (A and B) Nitrate (μmol kg⁻¹), (C and D) total Prochlorococcus abundance (log (cells m⁻³)). (E, G, I, and K) Distributions of the four most abundant Prochlorococcus eukaryotes (log (cells m⁻³)). (F, H, and J) The three emergent model eukaryotes ranked vertically by abundance. Model Prochlorococcus biomass was converted to cell density assuming a quota of 1 fg P cell⁻¹ (27). Black lines indicate isotherms.
The Subcommittee is Focusing on a Short List of Exemplar Goals for Consideration

- Possibilities so far considered include:
  - Microbial Communities Associated with Carbon Sequestration
  - Microbial Communities Associated with Bioremediation
  - Communities Associated with Cellulose Degradation
  - A Synthetic Model Cell that Can be used for analysis of what needs to be measured for systems identification.
Predictive Modeling and Simulation

- **Goal:** Develop Integrated predictive models relating cell processes, phenotypes and response to environment

- **Predictive Models**
  - Metabolism
  - Transport
  - Regulation
  - Signaling
  - Replication
  - Development
  - Motility

- **Databases**
  - Sequences and Expression
  - Phenotypes and Imaging

- **Bioinformatics**
  - Annotation and Informatics
  - Network and Model Reconstruction

- **Modeling Targets**
  - Model Organisms
  - Diverse Organisms
  - Limited Communities
  - Natural Communities
An Integrated View of Modeling, Simulation, Experiment, and Bioinformatics
Challenges for Cell and Ecosystem Simulation

- Modeling cells rivals the complexity of climate and earth systems models
  - Multiple space and time scales
  - Millions of interacting parts
  - Populations of cells to understand emergent behavior
  - Integrated modeling necessary to advance theory in systems biology
- Cell modeling and systems biology could be a driver for Petascale computing and beyond