Supercomputing and The New (Quantitative) Biology

Grant S. Heffelfinger, PhD
Deputy Director
Materials Science & Technology
Sandia National Laboratories

gsheffe@sandia.gov

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

Outline

Quantitative biology as a computational application

Who is doing what, from commercial purchases to federal agencies

DOE Office of Science’s Genomes to Life Program (GTL)

The Sandia-Oak Ridge GTL Project
Quantitative Biology
A Definition

[A] biology in which mathematics and computing serve as essential tools in framing experimental questions, analyzing experimental data, generating models, and making predictions that can be tested. In quantitative biology, the multifaceted relationships between molecules, cells, organisms, species, and communities are characterized and comprehended by finding structure in massive data sets that span different levels of biological organization. It is a science in which new computational, physical, and chemical tools are sought and applied to gain a deeper and more coherent understanding of the biological world that has strong predictive power.


Full text available at http://www.nap.edu/books/0309085357/html

Quantitative Biology Challenges

Of the approximately 300 genes essential for life, more than 100 have no known function


Biological processes are exquisitely complex - biochemical reactions are energetically subtle and usually enzymatically catalyzed and many biochemical processes are far from equilibrium and driven by reactant/product transport and/or coupled reactions

Implications for Computing:
More Data, New Modeling and Simulation Methods, High End Computing
The Computing “Revolution” in Biology

Before genomics (and the advent of high throughput experimental biology), high-performance computing didn’t impact biology:

- Computers used mainly for experimental support
- Desktop power provided more than enough horsepower
- All other biology-relevant large-scale apps were really chemistry
- There wasn’t enough DATA to provide systems-level understanding or populate complex system models

In the genomics era ...

- bioinformatics (for data from high throughput experiments) and
- new efforts to apply modeling and simulation to understand complex biological systems drive vast new needs for computing.

Quantitative Biology
What’s Different? Experimental Philosophy

Get the answer; get it cheap. Fundamental understanding will follow. Massive numbers of rapid, cheap, often low accuracy measurements.

Versus

A few, time-consuming, carefully targeted, high accuracy measurements.

an anathema to measurement scientists
A Market Example

The Estimated Biochips Market
(Research report by Bioinsights, quoted in Electronic Business, April 2000)

<table>
<thead>
<tr>
<th>Year</th>
<th>DNA Chips</th>
<th>Protein Chips</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>$158M</td>
<td>$4M</td>
</tr>
<tr>
<td>2001</td>
<td>$249M</td>
<td>$8M</td>
</tr>
<tr>
<td>2005</td>
<td>$745M</td>
<td>$68M</td>
</tr>
</tbody>
</table>

Quantitative Biology
What’s Different? Computational Philosophy

- The Established Model: Modeling and simulation help understand and guide experimental results & approaches.

- The "ASCI Model": Experimental methods provide fundamental parameters for and validation for modeling and simulation (and enables predictive capability for coupled phenomena across multiple time and length scales).

- The New Biology Model: Vast amounts of data from massive numbers of rapid, cheap, often low accuracy measurements drive the need for the application of informatics (e.g. bioinformatics), often with little or no regard for developing models which capture fundamental physical and chemical phenomena (like combichem).

  an anathema to computational scientists ... but not necessarily to computer scientists
Quantitative Biology
What’s Different? Computing Requirements

Most physical science/engineering applications start with a single initial system and evolve it
– little or no data exchange with disk
  (checkpointing, final results, interprocessor communication)
– compute intensive

Quantitative biology will coupling to huge databases, both on disk as well as the internet
– large amounts of data exchange
– I/O intensive

Many high-performance computing architectures are inadequate for these challenges

Why?

Technical
Insufficient attention has been paid to communication requirements
– Interprocessor communication (comparing sequences between processors, dividing long sequences among multiple processors)
– I/O (searching large sequence libraries on disk, receiving many requests at a time from the outside world)

Market Drivers
Commodity is King, but ...

*a system of commodity systems

is not the same as

an engineered system using commodity components
A Figure of Merit
In Silico Drug Development

Embarrassingly Parallel Massively Parallel

1 10 100 1000 TeraOps

Bioinformatics Molecular Biophysics Complex Systems

Sequence Genome Assemble Gene Finding "Identification" Annotate Gene to Protein Map Protein Protein Interaction Pathways Normal & Aberrant Function in pathway Structure Drug Targets Cellular Response

Another Possible Scenario

Bioinformatics

Sequence Genome Assemble Gene Finding "Identification" Annotate Gene to Protein "Map" Protein Protein Interaction Pathways Normal & Aberrant Function in pathway Structure Drug Targets

shortcut

1 10 100 1000 1000 100
The Cytoprint Business Model
Focus on Cellular Regulatory Processes Controlling Metabolism

"Shotgun Metabolomics"

4000 X Cell Lines
2,000 X Stains and Dyes
10,000 Small Molecules
8 Trillion Robotic Control Experiments
(1 Machine = 1M Experiments/Year)

Digital Image Comparison & Data Mining

Drug Targets

So ... Who’s Doing What?
## Industry

The Stakes are High for Big Pharma & Biotech

### Background

#### Annual Sales
$300B

#### Historic Growth
10-14%/year

#### R&D Expenditures
$62B ($26.4B in US)

- $4.6B external R&D
- 11% of sales in 1980
- 20% of sales in 2000

- 70% of patented drugs come off patent in the several years.
- 80% average drop in sales revenue when patent expires.
- $600M average drug development cost.
- Diminishing pool of easy targets.

### Some Recent Responses

- **Glaucus Proteomics**: Teraflop system - 1024 processors (two SGI Origin 3800 512 CPU systems)
- **Vertex Pharmaceuticals**: 100+ Processor cluster, company featured in Economist article
- **IBM, Compaq**: $100 million investments in the life sciences market
- **NuTec**: 7.5 Tflops IBM cluster (US, & Europe-planned)
- **GeneProt**: Large-Scale Proteomic Discovery And Production Facility: 1,420 Alpha processors.
- **Celera**: builds 1st tera-cluster for biotechnology - speeds up genomics by 10x
- **Blackstone & Others**: Linux/Intel Clusters (Pfizer, Biogen, AstraZeneca, and so on)

## Accelerating Biology with Advanced Algorithms and Massively Parallel Computing

*A Cooperative Research and Development Agreement (CRADA) between Sandia National Laboratories and Celera Genomics*

**A 100 Tflops, Scalable to a Petaflop, Massively Parallel Supercomputer**

*Cost effective from informatics to simulation*

- HW Architecture
- Scalable OS
- Parallel I/O
- Reliability
- Parallel Algorithms for Advanced Bioinformatics Tools
- Integrated Computing Environment
Federal Agencies
  e.g. DHS, NIH, Darpa, NSF, FDA, EPA, and ...

NIH
  • The Biomedical Information Science and Technology Initiative (BISTI)
  • National Programs of Excellence in Biomedical Computing
  • Training grants for physical scientists and engineers
  • NIBIB
DARPA DSO
  • Biospice
NSF
  • Biology with Distributed Terascale Facility
  • Educational Programs (e.g. Interdisciplinary postdocs)
EPA
  • Computational Toxicology Initiative
  • Partnerships with DOE Laboratories

... and so on

DOE Office of Science
Genomes to Life Program

“To achieve the most far-reaching of all biological goals: a fundamental, comprehensive, and systematic understanding of life.”

Sponsored by Two Elements in the DOE Office of Science
Biological and Environmental Research & Advanced Scientific Computing Research

• Beyond characterizing individual life components (e.g. genes and sequences)
• Towards a more comprehensive, integrated view of biology at a whole-systems level
• Ultimately, an integrated and predictive understanding of biological systems and insights into how both microbial and human cells respond to environmental changes.
• The applications of this level of knowledge will be extraordinary and will help DOE fulfill its broad missions in energy, environmental remediation, and the protection of human health.

www.doegenomestolife.org
Why DOE?
Payoffs for the Nation

Within a Decade | Long Term
---|---
Develop knowledge base for cost-effective cleanup strategies | 2020
Understand earth’s natural carbon cycle and design strategies for enhanced carbon capture | 2040
Increase biological sources of fuels and electricity | 2050

Why DOE?
Continuing a Tradition of Achievements

In 1986 the DOE Office of Science launched the Human Genome Project to understand, at the DNA level, the effects of energy production on human health. DOE has the historic perspective, track record, and infrastructure to conduct the large-scale, complex, mission-driven science needed to achieve these goals. DOE, whose successes in advancing biological and computational sciences are well known:

- Sponsored most key technical advances and resources that made possible the sequencing of the human genome in the public- and private-sector efforts.
- Is acknowledged as the national leader in microbial genomics. Work accomplished in its Microbial Genome Program has determined almost 50 microbial genomes and led to the verification of a third branch of life on earth.
- Launched the field of nuclear medicine using radioisotopes to target specific cells in the body, and laid the foundation for such modern imaging technologies as PET and CT scans.
- Established the first National Computer Center and high-speed interconnects for supercomputers to enable researchers to analyze, model, simulate, and predict complex phenomena important to DOE missions. DOE advanced scientific computing has become crucial for research problems that are insoluble by traditional theoretical and experimental approaches, hazardous to study in the laboratory, or time-consuming and expensive to solve by traditional means.
Why DOE?
Cutting Edge Facilities

The Department conducts scientific investigations through DOE user facilities located at its national laboratories and other institutions. The strengths of these laboratories include:

- Production-scale DNA sequencing
- Major laboratory research facilities—X rays, neutrons, and other probes of material structure, properties, and phenomena—necessary for making critical measurements in biological systems
- Field research facilities for bioremediation, carbon management, and bioproduct (e.g., clean energy) research and development
- Unparalleled resources and expertise in high-performance computing and networking
- Skills and capabilities for technology development, including microfabrication and nanotechnologies
Five Initial Awards

Sandia, Oak Ridge, Lawrence Berkeley, U Mass., Harvard

Carbon Sequestration in Synechococcus: From Molecular Machines to Hierarchical Modeling
Sandia, Oak Ridge, Lawrence Berkeley, Los Alamos, National Center for Genome Resources, California/San Diego, Tennessee, Michigan, The Molecular Science Institute, California/Santa Barbara, Illinois

Genomes to Life Center for Molecular and Cellular Systems: A Research Program for Identification and Characterization of Protein Complexes
Oak Ridge, Pacific Northwest, Argonne, Sandia, North Carolina/Chapel Hill, Utah

Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria
Lawrence Berkeley, Sandia, Oak Ridge, California/Berkeley, Missouri, Columbia, Washington, Diversa Corp.

Analysis of the Genetic Potential and Gene Expression of Microbial Communities Involved in the in situ Bioremediation of Uranium and Harvesting Electrical Energy from Organic Matter
Umass., The Institute for Genomic Research, Argonne, Tennessee/Memphis

Microbial Ecology, Proteogenomics and Computational Optima
Harvard, MIT, Brigham and Women's Hospital, Massachusetts General Hospital

Grant S. Heffelfinger, Sandia Labs, PI
Project Philosophy
“A Goal 4 Proposal”

The major goal of this effort is to develop computational methods and capabilities to advance understanding of complex biological systems and predict their behavior.

The initial target for the development and testing of new methods and tools is *Synechococcus* Sp.

The major biological objective of this work is to elucidate the relationship of the *Synechococcus* genome to *Synechococcus’* relevance to global carbon fixation.

New Tools for New Challenges
Our Goal 4 Proposal Will Impact GTL Goals 1-3

- New high performance methods and software for characterizing protein complexes
- Efficient algorithms for determining regulatory pathways
- New approaches to computational systems biology
- Improved methods for obtaining and evaluating *Synechococcus* data
- Work environments & computational infrastructure for GTL
Carbon Fixation in *Synechococcus*
A Computational Decomposition of the Problem

- Identify candidate proteins involved in carbon fixation through gene expression data analysis, regulatory binding site prediction, and operon/regulon structure prediction
- Identify protein interactions through analysis of affinity data and public protein-protein interaction data
- Protein structure prediction through Rosetta-type algorithms and refinements
- Elucidate gene regulatory pathways via systematic inference methods
- Link to cellular and macroscopic response
- Experimental verification
- Model refinement through an iterative process of computation and experiments

Carbon Fixation & Molecular Machines
Carboxysome, ABC Transporters, and Histidine Kinase-Response Regulators

Carboxysomes have been experimentally characterized
- at least ten polypeptides present
- two inside the core (structures known)
- > 6 or 7 are in the shell (structures not known)

Our computational and experimental efforts will focus on molecular machines key to the carbon fixation process in *Synechococcus.*
Three Synergistic Computational Biology Efforts Form the Core of This Effort

- **Molecular Machines**: Regulatory binding sites, operons, regulons
- **Regulatory Pathways**: Communication between pathways, cellular response
- **Systems Biology**: Binding constants, protein interactions, Physical basis for transporters, carbon fixation unit (carboxysomes)

Computational Investigations & Capability Development

Two Additional Efforts Support the Computational Biology Core

- **Molecular Machines**: Gene regulation networks
- **Regulatory Pathways**: Protein networks
- **Systems Biology**: Microarray data analysis

Experimental Investigation & Data Generation

- Matlab-like systems biology tool
- Inferred networks

Work Environments & Computational Infrastructure

- Gene expression data
- Protein network
- Phage display data

Computational Investigations & Capability Development

- Regulatory network characterization
- Ligand binding data
- Gene regulation networks

Data mgt/clustering for gene expression, protein interactions, & regulatory networks

Biology web portal/Genomic Integrated Supercomputing Toolkit
**Participants**

<table>
<thead>
<tr>
<th>Sandia National Laboratories</th>
<th>National Center for Genome Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bioinformatics &amp; Data Visualization</td>
<td>Complex Systems Modeling</td>
</tr>
<tr>
<td>Experimental Biology</td>
<td></td>
</tr>
<tr>
<td>Spectroscopy &amp; Multivariate Analysis</td>
<td></td>
</tr>
<tr>
<td>Computational Molecular Biology</td>
<td></td>
</tr>
<tr>
<td>Complex Systems Modeling</td>
<td></td>
</tr>
<tr>
<td>Statistics &amp; Experiment Design</td>
<td></td>
</tr>
<tr>
<td>High Performance Computing</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Oak Ridge National Laboratory</strong></td>
<td></td>
</tr>
<tr>
<td>Bioinformatics</td>
<td></td>
</tr>
<tr>
<td>Computational Molecular Biology</td>
<td></td>
</tr>
<tr>
<td>Statistics</td>
<td></td>
</tr>
<tr>
<td>High Performance Computing</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Lawrence Berkeley National Laboratory</strong></td>
<td></td>
</tr>
<tr>
<td>Data Management</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Los Alamos National Laboratory</strong></td>
<td></td>
</tr>
<tr>
<td>Computational Molecular Biology</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**For More Information & Acknowledgements**

www.genomes-to-life.org

Abridged project proposal recently published in OMICS: The Journal of Integrative Biology

Grant S. Heffelfinger1*, Anthony Martino2, Andrey Gorin3, Ying Xu1, Mark D. Rintoul III1, Al Geist3, Hashim M. Al-Hashimi3, Laurie J. Frank1, Andrey Gorin3, William E. Hart1, Erik Jakobsson1, Todd Lane2, Brian Palenik6, Steven J. Plimpton1, Diana C. Roe7, Nagiza F. Samatova3, Charlie E. M. Strauss3

*Author to whom correspondence should be addressed (gsheffe@sandia.gov)

1Sandia National Laboratories, Albuquerque, NM
2Sandia National Laboratories, Livermore, CA
3Oak Ridge National Laboratory, Oak Ridge, TN
4Lawrence Berkeley National Laboratory, Berkeley, CA
5Los Alamos National Laboratory, Los Alamos, NM
6University of California, San Diego
7University of Illinois, Urbana/Champaign
8University of Michigan