Inventing the Future

- Genomes to Life - a compelling partnership at the frontier between computing and biology, enabled through DOE’s programmatic partnership
- bVisionary, bPractical, bLucky – innovative partnerships will create unique DOE role and sustain impact
- Computing has changed biology forever, most just don’t know it yet – Michael Levitt
- Computational Biology is as essential for next Q Century as Molecular Biology was the last
Computational Biology for GTL in the High-Throughput Era

Genome-enabled Science and Biological Machines

- Scientific Challenges
- Algorithmic Challenges
- Data Integration Challenges
- Computational Challenges
Computation within Genomes to Life

• All Goals involve computation
• Must integrate SC-IT with studies on molecular machines; gene regulatory circuits, metabolic pathways, signalling networks; microbial ecology and microbial cell project – not just with the study of biological complexity
• GTL is a program for “the rest of time” with an unlimited scope and need for partnerships
Meso-scale Federal Programs – GTL prototype enablers

• NIGMS – Complexity, Glue Grants, Structural Genomics
• NCHGR – Centers of Excellence
• DOE – Microbial Genomes
• NSF – Microbe Project, Systems Biology, Genome enabled Science; Mathematical and Computational Biology; Database Activities; ITR
Crosscutting Themes Underlying Major Computational Biology Challenges Beyond The Genome

• Bridging Vast Scales of Time, Space and Organization
• Specifying the Relationship of Sequence, Structure and Function
• Understanding the Complexity of Living Systems
Objectives & Overview

• Computational tools developed to extract implicit genome information will allow biologists to understand how genes are regulated; how DNA sequences dictate protein structure and function; how genetic networks function in cellular development, differentiation, health and disease

• Knowledge intensive computing – organized and understood data – theme of computational biology
Biomedical Data: High Complexity and Large Scale

Protein-Protein Interactions
- metabolism pathways
- receptor-ligand
- 4\textsuperscript{o} structure

Proteins
- sequence
- 2\textsuperscript{o} structure
- 3\textsuperscript{o} structure

Polymorphism and Variants
- genetic variants
- individual patients
- epidemiology

Genetics and Maps
- Linkage
- Cytogenetic
- Clone-based

Physiology
- Cellular biology
- Biochemistry
- Neurobiology
- Endocrinology
- etc.

DNA sequences
- alignments
- \text{ATCGAATTCCAGCGTCACATTCTCAATTCCA}

ESTs
- Expression patterns
- Large-scale screens
Biology is becoming an Information Science

- Information S&T provides tools, language for integrating experimental observations.
- BioInformatics’ role is analogous to math in physical sciences
- CBT will capture emergent behavior between organizational levels.
Information-driven Drug Discovery

**Target Identification**
- Gene Expression Microarray
  - Computational Tools

**Proteins Structures Functions Interactions Pathways**
- Computational Tools

**Lead Identification Optimization**
- HTS Target-lead Interaction ADMET
  - Computational Tools

**Clinical Trials**
- Pharmaco-genomics (SNPs)
  - Computational Tools

Databases
Disciplines in the Top50 Projects at the NSF Supercomputer Centers

Particle Physics
Materials
Biomolecular
Industry
Chemistry
Earth Sciences
Engineering
CFD
Planetary Sciences

60,000 to 2 Million NUs Per Project

6/1/97 to 5/31/98 CTC, NCSA, PSC, SDSC - calculation by Smarr
Bottstein-Smarr report - BISTI

- Research Centers of Excellence – 8M$/yr
- Major focus in training new generation
- Bring CS, IT, physical scientists into bio, fund core algorithms, software, simulation
- Use existing resources, esp. NSF’s, for larger scale efforts
Experiments and Simulation for Structural Genomics

DNA Sequence Implies Structure Implies Function

- DNA Sequence Provides Protein Sequence
- DOE Synchrotron Facilities Provide 3-D Protein Structure
- Basis for 21st Century Medicine, Sustainable Development: Enhanced U.S. Competitiveness, Environmental Quality

Advances in Computational Biotechnology are Required to Extract Implicit Genome Information

Accelerated Drug Development
    Individualized Medicine
    Productive, Healthy Citizens

Environmental Remediation
    Biofuels, Biocatalysts
    Improved Agriculture
Challenge of Protein Classification: Pfam Example; S. Eddy

- Pfam contains 900 protein family profiles (as of 1999)
- Uses HMM-base sequence comparison
- Look for genes related to these families in human genome; must be rerun frequently
- Requires $10^{14}$ matrix cells
- 70 days on 1024 node processor
- Family profiles expected to grow to 5000
- High performance code development needed
Mol Bio “Classic Benchmark” for Scientific Computing

- Protein folding using molecular dynamics package Amber (Duan and Kollman)
  - 36 residue villin headpiece
  - Simulated 1 microsecond using several months, dedicated 256 processors of T3D (75 Gflops peak)
Recent Envelope-pushing
“Routine” Mol Bio Driver

Folding landscape of small protein (Charles Brooks)

- 56 residues
- 1 month dedicated 192 T3E processors (170Gflops peak) [at PSC]
- 200 residues, an average sized or typical protein, would take 4 months at this computer power
- 40 Tflop machine would deliver in 12 hours
- 10K to 100K proteins by 10K simulations
Swimming in Data:
Exploding ability to capture and manipulate data

• Across Scales, Disciplines, National Boundaries

• From Acquisition, Refinement, Reduction and Deposition

Pointing to need of analysis and linking across these interfaces
GTLTestbed

DATA ACQUISITION & COLLABORATORY
- Python
- Volume
- Viewer
- Continuity
- COSGMS
- CIF

Anatomy
- Anatomy
- Physiology
- Cell Biology
- Proteomics
- Genomics
- Medicinal Chemistry

Organism
- Organ
- Organ
- Organ

SCIENTIFIC VISUALIZATION & PORTALS
- CMEM
- Angles
- MAST
- Meta MEME
- MIA
- QMView
- GAMESS
- FPS
- Flex
- MPS
- MMQL
- XtalView

Cell Biology
- Cell Biology
- Cell Biology
- Cell Biology

Proteomics repository
- Proteomics
- Genomics
- Sequence
- Structure

Genome DB’s

DATABASE QUERY & VISUALIZATION
- CMS-MBR
- Angles
- MotKit
- Rabbit/Dog
- Ventricles
- Finite Element
- Anatomic Models

Pioneering Technologies and Services for the Biomedical Community
Modeling the Cellular Program – Experimental Advances Required

Three mammalian signal transduction pathway that share common molecular elements (i.e. they cross-talk). From the Signaling PAthway Database (SPAD) (http://www.grt.kyushu-u.ac.jp/spad/)

Integrating Computational/Experimental Data at all levels
   Sequence and its structural and functional annotation
   Simulating biochemical/genetic networks to mode cellular decisions
   Modeling of network connectivity (sets of reactions: proteins, small molecules, environment along DNA)
   Functional analysis of that network (kinetics of the interactions)
Challenges of Systems Simulation

- Track expression of 100,000 genes as a function of time, cells, internal conditions and environment
- Genes and proteins interact in complex non-linear pathways
- Dynamic systems capable of chaotic behavior
- A very large analysis problem for the future
Exemplar GTL Research Problems that can only be addressed by BioIT & Comp_Bio

1. Full genome-genome comparisons
2. Rapid assessment of polymorphic genetic variations
3. Complete construction of orthologous and paralogous groups of genes
4. Structure determination of large macromolecular assemblies/complexes
5. Dynamical simulation of realistic oligomeric systems
6. Rapid structural/topological clustering of proteins
7. Prediction of unknown molecular structures; Protein folding
8. Computer simulation of membrane structure and dynamic function
9. Simulation of genetic networks and the sensitivity of these pathways to component stoichiometry and kinetics
10. Integration of observations across scales of vastly different dimensions and organization to yield realistic environmental models for basic biology and societal needs
Summary: GTL Program is a direct progeny of revolution in computing and genome

- Bioscience research used virtually no supercomputer time in 1985 but by 1998 is the number 1 application at NSF-supported supercomputer centers.
- Bioscience is ideally positioned to exploit the hardware and software under discussion for future high performance computing.
Synergistic Revolution in Computing and Biology: Frontiers between Scientific Computing and Life Sciences empowers Genomes2Life

- Maturation of experimental and integrative biology & appreciation of data intensity, system complexity and effect of scaling
- Vigorous training in high-performance computing techniques; experience in physical sciences
  - Utility of massively parallel computing has improved dramatically in this decade
- Development of widely-used advanced software.

These now justify massive applications of the technology to the basic and applied life sciences – a driver for the foreseeable future
An Era of “Meso-scale” Biology?

- **Collaboratories**: intermediate scale environments for remote access to specialized instruments, & for interdisciplinary, multidisciplinary +/- multi-institutional interactions - both local and distant

- **Portfolio Balance**: instrument development (high versus low risk, exploratory versus implementation) and prototype applications
  - Teams are needed; “ferment” for selecting direction, then driving instrument development (by situation dependent biology pull and technology push); facilitate commercialization and getting tools to the community for a broad impact
  - portfolio policy considerations include cost sharing in instrument acquisition and mechanisms to enable PIs to sustain applications / user facilities

- **Data Deluge**: rapid data acquisition for high throughput biology, organization, long term maintenance, query and integration tools

- **Quantitative Approaches**: beyond binary biological questions, and (renewal of) adequate training to address the complexity of biology
GTL Telescience/Collaboratory

Changing How Science is Done
Providing the Tools to Swim in the Data
Scientific Teams @ Research
(*’s)

Stable, large scale funding of virtual collaborations built around a balanced and integrated systems approach for the systems package of hardware, software and networking to achieve understanding of specific biology research problems

- **Software Environment – DOE Nat Labs experience**
  - Algorithms
  - Verification and Validation
  - Support for Application Innovation
  - Problem solving environments
  - Collaborative environments

- **Institutional Partnerships – Universities, NFPs, Industry**
  - Trained People for next generation
  - Universities – core strengths can be harnessed
  - Involvement of applications researchers
  - Industrial involvement-both vendors and users/researchers
  - Good technology transfer mechanisms for spinoffs
Major Organizational Challenges for GTL *’s

• Maintaining essential triangle of theory, computation and experimentation for basic and applied life sciences requires concomitant advances in instrumentation & experimental methodology
  * Each component - genes, metabolism, signalling pathways - requires teamwork
• Maturation of High Throughput Biology provides the opportunity
• Success requires focus on deliverables for biology, justifying major expenditures by OSCR and BER
Devil’s Details Discussed

- By other speakers
- As needed
As Our Anatomy “Moves On-Line”
Bioengineering and Bioinformatics Merge

- New Sensors—Israeli Video Pill
  - Battery, Light, & Video Camera
  - Images Stored on Hip Device

- Next Step—Putting You On-Line!
  - Wireless Internet Transmission
  - Key Metabolic and Physical Variables
  - Model -- Dozens of 25 Processors and 60 Sensors / Actuators Inside of our Cars

- Post-Genomic Individualized Medicine
  - Combine
    - Genetic Code
    - Body Data Flow
  - Use Powerful AI Data Mining Techniques
The Future of Computing *Fuel* Data Analysis

- Computing Platforms will exploit emerging trends
  - The rise of Linux
  - Commodity Clusters
  - Ubiquitous access
  - The maturation of Grid Computing

*Source: Dave Turek, IBM*
Data Storage Platforms Changing

- Growth of archival storage to petabyte sizes
- Increasing on-line storage area networks to terabyte size

<table>
<thead>
<tr>
<th>Voxel size</th>
<th>B&amp;W (1 B/p)</th>
<th>High res (2 B/p)</th>
<th>Color (3 B/p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cm</td>
<td>1.5 KB</td>
<td>3 KB</td>
<td>4.5 KB</td>
</tr>
<tr>
<td>mm</td>
<td>1.5 MB</td>
<td>3 MB</td>
<td>4.5 MB</td>
</tr>
<tr>
<td>10 μm</td>
<td>1.5 TB</td>
<td>3 TB</td>
<td>4.5 TB</td>
</tr>
<tr>
<td>μm</td>
<td>1.5 PB</td>
<td>3 PB</td>
<td>4.5 PB</td>
</tr>
</tbody>
</table>

Volume sizes by resolution - brain = 1500 cm$^3$
New Results Possible on TeraGrid

• Modeling Cell Structures
  – Pre-Blue Horizon
    • Possible to model electrostatic forces of a structure with up to 50,000 atoms -- a single protein or small assembly
  – Pre-TeraGrid:
    • Possible to model one million atoms -- enough to simulate drawing a drug molecule through a microtubule or tugging RNA into a ribosome
  – TeraGrid:
    • Models of 10 million atoms will make it possible to model function, structure movement, and interaction at the cellular level for drug design and to understand disease

New Results Possible on TeraGrid

- **Biomedical Informatics Research Network [BIRN] (NIH)**
  - Evolving reference set of brains provides essential data for developing therapies for neurological disorders (Multiple Sclerosis, Alzheimer’s disease).

- **Pre-TeraGrid:**
  - One lab
  - Small patient base
  - 4 TB collection

- **Post-TeraGrid**
  - Tens of collaborating labs
  - Larger population sample
  - 400 TB data collection: more brains, higher resolution
  - Multiple scale data integration and analysis
Recent Hardware History

Can economically configure systems with very large sizes with tailorable speeds and latencies because

- Technology has progressed
- Building blocks can be commodity PC’s, workstations or SMP’s even for the most capable machines.
- Interconnects range from inexpensive, commodity network connections through high-speed and proprietary
Recent Software History

- Overall Computational Environment and Infrastructure have become stable and advances “more transparent” - a prerequisite for sustained scientific investigation
- Emergence of widely accepted, public standards for parallel code such as MPI and PVM are supported on virtually every parallel platform,

As a result, much more applications code including BIO has been parallellized in the last few years