GTL Software Infrastructure: A Computer Science Perspective

Rick Stevens
Argonne National Laboratory
University of Chicago
Tool Building Will Be Critical for GTL Progress

“If I had eight hours to chop down a tree, I’d spend six sharpening my axe.”

-- Abraham Lincoln
The Outline

- Database Tools and DB Frameworks
- Computational Toolkits Focused on Biological Systems
- Computational Mathematical Biology Topics
- Hierarchical Biological Systems Modeling Environment
- International Systems Biology Grid
- Workflow for Understanding Microbial Life
- Modeling Example
- Simulation Targets
- Technologies Needed for Biological CAD: Tomita’s E-cell List
- Top 10 Bioinformatics Goals: Richard Smith’s List
- Conclusions
Database and DB Frameworks

- New Types of Data Support extend existing RDBMS [Oracle, db2]
  - Sequences and Strings
  - Trees and Clusters
  - Networks and Pathways
  - Deep Images
  - 3D Models and Shapes
  - Molecules and Coordinate Structures
  - Hierarchical Models and Systems Descriptions
  - Time Series and Sets
  - Probabilities and Confidence Factors
  - Visualizations
- Store, Retrieve, index, search, compare, transform, input/output, import/export, etc.
Database and DB Frameworks [continued]

- Scalable performance [parallel, open source OS support]
- Robust Data Exchange
- Data Provenance and curation Tools
- Query Optimization including computing in the query
  - Fast operator support for biological structures
    - Regular expression, sequence similarity, tree comparisons, confidence factors, etc.
- Robust linkages to:
  - Simulation [logical and numerical]
  - Automated Deduction and Knowledge Structures [production systems]
  - Interfaces to experimental systems [chips, detectors, microscopes, etc.]
  - Workflow and planning
  - Visualization and statistical analysis
Some Mathematical Biology Topics Relevant to GTL

- Population models, symbiosis and stability
- Discrete growth models
- Reaction kinetics
- Biological oscillators and switches
- Coupled oscillators
- Reaction-diffusion, chemotaxis and non-locality
- Oscillator generated wave phenomena and patterns
- Spatial pattern formation with population interactions
- Mechanical models for generating pattern and form in development
- Evolution and morphogenesis
Mathematical Toolkits Focused on Biological Systems

• “A Mathematica for molecular, cellular and systems biology”
  • Core data models and structures [see db]
  • Optimized functions [see core libraries]
  • Scripting environment [e.g. Python, PERL, ruby, etc.]
  • Database accessors and built-in schemas
  • Simulation interfaces
  • Parallel and accelerated kernels
  • Visualization interfaces [info-vis and sci-vis]
  • Collaborative workflow and group use interfaces
Some of the Core Library Functions

- Algorithms for Restriction Maps and Map Assembly
  - Planning cloning and clone libraries, building physical Genome Maps
- Sequence assembly, Multiple sequence assembly
  - Data models and sequence analysis algorithms, multiple sequence alignment
  - Probability and statistics for sequence alignment and patterns
- Gene prediction, Mutation analysis
- Trees and sequence comparisons, construction
  - Phylogenetic tree construction and analysis
  - Comparative Genomics
- Proteomics analysis
  - Protein structure prediction and kinetics prediction
  - Array analysis
Hierarchical Biological System Modeling Environment

- 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.
An International Systems Biology Grid

- A Data, Experiment and Simulation Grid Linking:
  - People [biologists, computer scientists, mathematicians, etc.]
  - Experimental systems [arrays, detectors, MS, MRI, EM, etc.]
  - Databases [data centers, curators, analysis servers]
  - Simulation Resources [supercomputers, visualization, desktops]
  - Discovery Resources [search servers perhaps optimized]
  - Education and Teaching Resources [classrooms, labs, etc.]
- Different than and more fine grain than current Grid Projects
- More laboratory integration [need small laboratory]
A Systems Biology Approach to Understanding Cellular Life Integrating:

- Theory
- Numerical methods
- Software development
- Data models and Visualization
- DB support
- Education
A GTL Focus on Prokaryotic life [bacteria]

• ~4,000 known species of prokaryotic life
  • It is estimated that we have identified only 1%-5% of extant species ⇒ 80,000-400,000 species
  • Can we attack the culturability problem via simulation?
• More diversity than Eukaryotic life forms
  • More diverse metabolisms, more diverse environments
• A Human contains $10^{12}$ cells and $10^{13}$ bacterial cells
• Estimated that protoctista 250,000 species are extant [10,000 known today, 40,000 from paleontology]
Typical bacteria [E. coli]

- 1000nm x 300nm x 300nm volume
- ~4000 genes and gene products
  - ¼ genes $\Rightarrow$ protein synthesis
  - ¼ genes $\Rightarrow$ glycolysis
  - ¼ genes $\Rightarrow$ citric acid cycle
  - rest genes involved in regulation, synthesis and degrading tasks
  - Relatively few genes related to sensing and motility
- 1,000’s of small molecular species
- ~3 million total molecules to track
Intracellular environment

- 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

R. Stevens  Argonne National Laboratory + University of Chicago
Cell Membranes and Cell Wall

- Cell wall
  - Polysaccharides
  - Porin pores
- Peptidoglycan
  - crosslinked
- Periplasmic space
  - Small proteins
- Complex inner membrane
  - < 50% lipids

From: David Goodsell, The Machinery of Life

R. Stevens
Argonne National Laboratory + University of Chicago
Flagellum and Flagellar Motor

- Transmembrane proton powered rotating motor
- About 10 Flagella per cell
- 5-10 µm long
  - Built from the inside out
- Propels cell ~10-20 µm/sec
  - Medium is extremely viscous

From: David Goodsell, The Machinery of Life

R. Stevens
Argonne National Laboratory + University of Chicago
DNA Replication via DNA Polymerase

- DNA replication about 800 new nucleotides per second
- In circular DNA both directions at once
- 50 minute to duplicate entire circle of 4,700,000 nucleotides
- With cell replication ~30 minutes
- Simultaneous duplication possible

From: David Goodsell, The Machinery of Life
Phototrophic Prokaryotic Cell

From: Lynn Margulis, The Five Kingdoms of Life

R. Stevens Argonne National Laboratory + University of Chicago
Flagellum [prok] vs Undulipodium [euk]
Complexity

From: Lynn Margulis, The Five Kingdoms of Life
R. Stevens
Argonne National Laboratory + University of Chicago
Some Modeling and Simulation Targets for GTL

- Modeling activity of single genes
- Probabilistic models of prokaryotic genes and regulation
- Logical models of regulatory control in eukaryotic systems
- Gene regulation networks and genetic network inference in computational models and applications large-scale gene expression data
- Atomistic level simulation of biomolecules
- Diffusion phenomena in cytoplasm and extracellular environment
- Kinetic models of excitable membranes and synaptic interactions
- Stochastic simulation of cell signaling pathways
- Complex dynamics of cell cycle regulation
- Model simplification
Technologies for Biological CAD: Tomita’s list

- **Enzyme engineering**: to refine enzymes and to analyze kinetic parameters *in vitro*.
- **Metabolic engineering**: to analyze flux rates *in vivo*.
- **Analytical chemistry**: to determine and analyze the quantity of metabolites efficiently.
- **Genetic engineering**: to cut-and-paste genes on demand, for modifying metabolic pathways.
- **Simulation science**: to efficiently and accurately simulate a large-number of reactions.
- **Knowledge engineering**: to construct, edit and maintain large-metabolic knowledge bases.
- **Mathematical engineering**: to estimate and tune unknown parameters.
Bioinformatics Goals: Richard Smith’s top 10 Challenges

• Precise, predictive model of transcription initiation and termination: ability to predict where and when transcription will occur in a genome

• Precise, predictive model of RNA splicing/alternative splicing: ability to predict the splicing pattern of any primary transcript

• Precise, quantitative models of signal transduction pathways: ability to predict cellular response to external stimuli

• Determining effective protein:DNA, protein:RNA and protein:protein recognition codes

• Accurate ab initio structure prediction

• Rational design of small molecule inhibitors of proteins
Bioinformatics Goals: Richard Smith top 10 Challenges

- Mechanistic understanding of protein evolution: understanding exactly how new protein functions evolve
- Mechanistic understanding of speciation: molecular details of how speciation occurs
- Continued development of effective gene ontologies – systematic ways to describe the functions of any gene or protein
- Education: development of appropriate bioinformatics curricula for secondary, undergraduate, and graduate education
Conclusions

- Priorities for R+D software infrastructure investment
  - Extensions to current database capabilities
    - Kernel functions for trees and networks, uncertainty representation
  - New types of toolkits that integrate bioinformatics algorithms
    - “Mathematica” or “matlab” for computational molecular and cell biology
    - Visual representation and analysis of biological systems
  - New algorithms for core problems in systems biology
    - Perhaps linked with new developments in hardware
  - Scalable hierarchical modeling environments
    - Multiple levels of biological detail and levels of abstraction
  - International grid for systems biology