Computational Biology and High Performance Computing

Manfred Zorn, Teresa Head-Gordon, Adam Arkin, Brian Shoichet, and Horst D. Simon

National Energy Research Scientific Computing Division
Ernest Orlando Lawrence Berkeley National Laboratory
University of California
Berkeley, California 94720

October 1999

This work was supported by the Director, Office of Science, Office of Advanced Scientific Computing Research, Division of Mathematical, Information, and Computational Sciences, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.
Computational Biology and High Performance Computing

Presenters:
Manfred Zorn – Co-Head, Center of Bioinformatics and Computational Genomics, NERSC
Teresa Head-Gordon – Scientist, Physical Biosciences Division, LBNL
Adam Arkin – Scientist, Physical Biosciences Division, LBNL
Brian Shoichet, Northwestern University

Organizer: Horst D. Simon – NERSC Director
November 15, 1999

Abstract

The pace of extraordinary advances in molecular biology has accelerated in the past decade due in large part to discoveries coming from genome projects on human and model organisms. The advances in the genome project so far, happening well ahead of schedule and under budget, have exceeded any dreams by its protagonists, let alone formal expectations.

Biologists expect the next phase of the genome project to be even more startling in terms of dramatic breakthroughs in our understanding of human biology, the biology of health and of disease. Only today can biologists begin to envision the necessary experimental, computational and theoretical steps necessary to exploit genome sequence information for its medical impact, its contribution to biotechnology and economic competitiveness, and its ultimate contribution to environmental quality.
Abstract (cont.)

- High performance computing has become one of the critical enabling technologies, which will help to translate this vision of future advances in biology into reality. Biologists are increasingly becoming aware of the potential of high performance computing. The goal of this tutorial is to introduce the exciting new developments in computational biology and genomics to the high performance computing community.

Tutorial Outline

- 1:30 - 2:00 p.m. Overview of Computational Biology
  -- Teresa Head-Gordon
- 2:00 - 3:00 p.m. Bioinformatics -- Manfred Zorn
- 3:00 - 3:30 p.m. Break
- 3:30 - 4:00 p.m. Protein Structure Prediction and Folding -- Teresa Head-Gordon
- 4:00 - 4:30 p.m. Docking/Molecular Recognition
  -- Brian Shoichet
- 4:30 - 5:00 p.m. Cellular Networks -- Adam Arkin
Computational Challenges in Structural and Functional Genomics

Teresa Head-Gordon
Physical Biosciences and Life Sciences Divisions
Lawrence Berkeley National Laboratory

November 15, 1999

(1) Why computational biology?

(2) Community effort to define problems with genuine computational complexity
    Genome analysis, gene modeling, sequence-based annotation
    Low resolution fold prediction: Single Molecule
    High resolution structure prediction and protein folding: Single Molecule
    Molecular recognition or Docking: Multi-molecule complexes
    Cellular Decision modeling

(3) Putting it all together:
    Deinococcus radiodurans
    Center for Integrative Physiome Analysis (CIpHA)
**Revolutionary Experimental Efforts in Biology**

- Sequence
  - Genome projects
  - Microbial organisms
    - C elegans
    - Human
- Structure
  - Structural Genomics Initiative
  - High throughput effort underway
    - NIH, new beamlines
    - LBNL: ALS
- Function
  - Functional Annotation Initiatives
  - Gene deletion projects
  - Yeast two-hybrid screening
  - Gene expression micro-arrays
  - In vivo GFP protein (kinetics)

**Computational Biology White Paper**

- http://cbcg.lbl.gov/ssi-csb
- A technical document to define areas of biology exhibiting computational problems of scale
  - Organization:
    - Introduction to biological complexity and needs for advanced computing (1)
    - Scientific areas (2-6)
    - Computing hardware, software, CSET issues (7)
    - Appendices
  - For each scientific chapter:
    - Illustrate with state of the art application (current generation hpc platform)
    - Define algorithmic kernals
    - Deficiencies of methodologies
    - Define what can be accomplished with 100 teraflop computing
  - ➢ Community document
  - ➢ More organized CB community in government labs, universities
  - ➢ Support for CB by the broader biological community
The Genome Channel Browser to access and visualize current data flow, analysis and modeling. (Manfred Zorn, NERSC)

- Genome sequencing and annotation → Bioinformatics
- 100,000 human genes; genes from other organism
- Structure/functional annotation at the sequence level
- Computation to determine regions of a genome that might yield new folds
- Experimental Structural Genomics Initiative
- Functional annotation at the structure level by experiment

---

Characterize the Link Between Protein Sequence and Fold Topology

Sequence Assignments to Protein Fold Topology (David Eisenberg, UCLA)

- Experimental Structural Genomics Initiative
- Define basis set of folds: ~10³ structures to be determined
- Predict Fold Topology from Computation (~10⁸ folds)
- Functional annotation at the structural level by computation
Low Resolution Fold Topologies to High Resolution Structure

One microsecond simulation of a fragment of the protein, Villin.
Duan & Kollman, Science 1998

Influenza virus poised above a model of a lipid membrane will involve a 100,000 atom MD simulation over long timescales to understand this step in the mechanism of viral infection. (Tobias, UCI)

Low Resolution Structures from Predicted Fold Topology
Fold class gives some idea of biological function, but....

Higher Resolution Structures with Biochemical Relevance
Drug design, bioremediation, diseases of new pathogen

Simulating Molecular Recognition/Docking

Changes in the structure of DNA that can be induced by proteins. Through such mechanisms proteins regulate genes, repair DNA, and carry out other cellular functions.

Improvements in Methodology and Algorithms of Higher Resolution Structure
Breaking down size, time, lengthscale bottlenecks (IT², algorithms, teraflop computing)
Protein, DNA recognition, binding affinity, mechanism with which drugs bind to proteins
Simulating two-hybrid yeast experiments
Protein-protein and Protein-nucleic acid docking
Modeling the Cellular Program

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

Integrating Computational/Experimental Data at all levels
Sequence, structural functional annotation (Virtually all biological initiatives)
Simulating biochemical/genetic networks to model cellular decisions
Modeling of network connectivity (sets of reactions: proteins, small molecules, DNA)
Functional analysis of that network (kinetics of the interactions)

Implicit Collaborations Across the DOE Mission Sciences

Computer Hardware & Portability
Applications described running on various platforms
T3D, T3E, IBM SP's, ASCI Red, Blue

Information Technologies and Database Management
Integrating biological databases; CORBA and java
Data Warehousing
ultra-high-speed networks

Ensuring Scalability on Parallel Architectures
implicit algorithmic scaling
paradigm/software library support tools for effective parallelization
strategies: 100 teraflop

Meta Problem Solving Environments
geographically distributed software paradigm: “plug and play” paradigm

Visualization
Querying data which is “information dense”
Feedback from Biotech Industry Meeting

LBNL 2/25/99

Jim Cavatcoli, Ph.D.
Bioinformatics Manager, PDLMG
Parke-Davis, Warner-Lambert

Patrick O'Hara
VP, BioMolecular Informatics
Zymogenetics, Inc
Seattle WA

Herve Recipon
Asst. dir. bioinformatics
diaDexus (Incyte)

Pete Smiatana, Ph.D.
Senior Staff Software Engineer,
Bioinformatics
Ciphergen

Peter Karp, Ph.D.
Scientific Fellow
Pangea Systems

Rick Bott
X-ray crystallographer
Genencor

Julie Rice
Computational Chemist
IBM-Almaden

Eric Martin
Sr. Scientist Small Molecule Discovery
Chiron

LBNL: Gilbert, Head-Gordon, Holbrook, Mian, Rekhsar, Simon, Spengler, Zorn

We want to listen to Biotech industry perspective on Computational Biology white paper

Is there strong objection to any of the content?
NO, very supportive
Are there other areas to be included, stronger emphasis placed?
Will be a new chapter on databases: integrating, querying, visualization
Technical input: contribute a "vignette" on important Comp. Bio. application
Parke-Davis, Chiron, Zymogenetics, Pangea

Supercomputing 99-Portland

Center for Integrative Physiome Analysis (CIPhA)

NCRR submitted 2/1/99

P.I.: Adam Arkin

Cell cycle, asymmetric division and differentiation in Caulobacter crescentus
Analysis of developmental pathways in C. Elegans
Analysis of databases of two-hybrid interactions
The role of cyto mechanical and nuclear structure in mammary gland transformation

Interrrelationships among the various tools and databases used and developed by the Center. Blue rectangles are databases built by the Center (with the exception of Interact 1.0 which is provided courtesy of Roger Brent, Molecular Sciences Institute). Green boxes are off-site database.

Hexagons are tools to be developed by this Center.

Adam Arkin, Mina Bissell, Roger Brent, Silvia Crivelli, Tarek Elaydi, Teresa Head-Gordon, Stephen Holbrook, Stuart Kim,
Casimir Kulikowski, Harley McAdams, Saira Mian, Ilya Muchnik, Lucy Shapiro, NERSC

Supercomputing 99-Portland
Deinococcus Radiodurans (DR: Strange Berry That Withstands Radiation)

Bacteria isolated from tins of spoiled meat given “sterilizing” doses of γ radiation.
3x10⁶ base pairs, or ~3000 protein products
fully sequenced by TIGR under DOE/BER sponsorship

Three components to DR’s successful DNA repair strategy
- specifies of the DNA repair mechanism
- the fact that it is multi-genomic
- coupling of repair, replication, export of damaged DNA from intracellular medium.

Propose to construct molecular models of key components of the DNA repair system:
- Damaged DNA
- Multigenomic repair intermediates such as Holliday junctions
- Proteins known are yet to be discovered to be involved in DNA repair
- Protein-protein or protein-nucleic acids that couple repair, replication, transport.

Developing better fold recognition, comparative modeling, and ab initio prediction methods, and docking methods to describe macromolecular complexes.
Application of methodologies will be to fully and completely annotate the DR genome
Learn underlying components of highly honed strategies for DNA repair in DR.

Involves significant portions of community white paper on high end computing needs

---

The Need for Advanced Computing for Computational Biology

Computational Complexity arises from inherent factors:
- 100,000 gene products just from human; genes from many other organisms
- Experimental data is accumulating rapidly
- N³, N⁴, N⁵, etc. interactions between gene products
- Combinatorial libraries of potential drugs/ligands
- New materials that elaborate on native gene products from many organisms

Algorithmic Issues to make it tractable
- Objective Functions
- Optimization
- Treatment of Long-ranged Interactions
- Overcoming Size and Time scale bottlenecks
- Statistics
Acknowledgements for Community White Paper in Computational Biology

The First Step Beyond the Genome Project: High-Throughput Genome Assembly, Modeling, and Annotation
P. LiCascio, R. Mural, J. Snoddy, E. Uberbacher, ORNL
S. Mian, F. Olken, S. Spengler, M. Zorn: LBNL
David States, Washington University

From Genome Annotation to Protein Fold:
Comparative Modeling and Fold Assignment
D. Eisenberg, UCLA
A. Ladner, LANL
A. Sali, Rockefeller University
B. Hong, Columbia University

Low Resolution Folds to High Resolution Protein Structure and Dynamics
C. Brooks, Scripps Research Institute
P. Kollman & Y. Daan, UCSF
A. McCammon & V. Heims, UCSD
G. Martyna, Indiana University
D. Tobias, UCI
T. Heat-Gordon, LBNL

Biotechnology Advances from Computational Structural Genomics: In Silico Drug Design and Mechanistic Enzymology
S. Abagyan, NYU, Skirball Institute
P. Bash, ANL
J. Blaney, Metaphorics, Inc.
F. Cohen, UCSF
M. Colvin, LLNL
I. Kuntz, UCSF

Linking Structural Genomics to Systems Modeling:
Modeling the Cellular Program
A. Arkin & D. Wolf, LBNL
P. Karp, PangeoS. Subramaniam, U Illinois Urbana

Implicit Collaborations Across the DOE Mission Sciences
M. Colvin & C. Musiek, LLNL
T. Gaasterland, ANL (now Rockefeller)
S. Crivelli & T.Head-Gordon, LBNL
G. Martyna, Indiana University

Bioinformatics

Manfred D. Zorn
November 15, 1999
Overview

- 30 seconds of Biology
- DNA Sequencing: View from 10,000 feet
- Genome Analysis
  - Genome Projects
  - Identify a possible gene
  - Characterize a gene
- Large-scale Genome Annotation
- What's supercomputing got to do with it?
- Challenges

Biology is Special

Life is characterized by

- *Individuality*
- *Historicity*
- *Contingency*
- *High (digital) information content*
**Basic Biology**

- Rough endoplasmic reticulum
- Golgi apparatus
- DNA packs tightly into metaphase chromosomes
- Mitochondrion
- Smooth endoplasmic reticulum
- Nucleus

**Fundamental Dogma**

- DNA
- RNA
- Proteins
- Circuits
- Phenotypes
- Populations

The fundamental dogma of molecular biology is that genes determine phenotypes. Information from DNA is transcribed into RNA, which is then translated into proteins. Interactions among proteins, regulatory circuits, and metabolic pathways give rise to phenotypes. Collections of individual phenotypes constitute a population.
DNA Codes

DNA Sequencing

Read base code from storage medium!

- Read length: About 600 bases at once
- Reader capacity
  ✓ 100 lanes in parallel in about 2-5 hours
Sequencing: “bird’s eye view”

- Prepare DNA
  - about a trillion DNA molecules

- Do the sequencing reactions
  - synthesize a new strand with terminators

- Separate fragments
  - by time, length = constant

- Sequence determination
  - automatic reading with laser detection systems

---

Sequence Traces

Good quality sequence needs about 10X coverage
Human Genome Project - Goals

- Construction of a high-resolution genetic map
- Production of a variety of physical maps of all human chromosomes and of selected model organisms
- Determination of the complete sequence of human DNA and DNA of selected model organisms
- Development of capabilities for collecting, storing, distributing, and analyzing the data produced
- Creation of appropriate technologies necessary to achieve these objectives

Genome Projects

- Model organisms sequenced
  - E. coli 4.5 Mb
  - S. cerevisiae
  - C. elegans 100 Mb
  - Dozens of bacteria 1 - 6 Mb
  - D. melanogaster 140 Mb

- Human
  - 408 Mb
  - ~14% of the genome
DNA Analysis

Disassemble the base code!

- Find the genes
  - Heuristic signals
  - Inherent features
  - Intelligent methods

- Characterize each gene
  - Compare with other genes
  - Find functional components
  - Predict features
What is a Gene?

Heuristic Signals

DNA contains various recognition sites for internal machinery

- Promoter signals
- Transcription start signals
- Start Codon
- Exon, Intron boundaries
- Transcription termination signals
Inherent Features

DNA exhibits certain biases that can be exploited to locate coding regions

- Uneven distribution of bases
- Codon bias
- CpG islands
- In-phase words
- Encoded amino acid sequence
- Imperfect periodicity
- Other global patterns
Intelligent Methods

Pattern recognition methods weigh inputs and predict gene location

- Neural Networks
- Hidden Markov Models
- Stochastic Context-Free Grammar

Neural networks

- 6-mer vocabulary
- 6-mer-in-frame
- Markov
- Isochore GC Composition
- Exon GC Composition
- Size prob. profile
- Length
- Donor
- Acceptor
- Intron Vocabulary 1
- Intron Vocabulary 2

Xu 1997
Hidden Markov Models

Silent states

Production states

---

Characterize a Gene

Collect clues for potential function

- Comparison with other known genes, proteins
- Predict secondary structure
- Fold classification

- Gene Expression
- Gene Regulatory Networks
- Phylogenetic comparisons
- Metabolic pathways
Large-scale Genome Annotation

- Multi-laboratory Project
- Standard Annotation of Genomes
  - Genome Channel
  - Genome Catalog
- Comprehensive integration of
  - Analysis tools
  - Data management systems
  - Data mining
  - User services
- Extensible Framework
  - High-performance computing
  - Data integration technology
  - Artificial intelligence

Annotation Pipeline

Sequence Input → Analysis Queue → Update Agents → Data Archive → Annotation Report → Search Agents

Supercomputing 99-Portland
Gene Search - BEAUTY Results

Distribution of 29 Blast Hits on the Query Sequence

Supercomputing 99-Portland

Highlights - Data Analysis

Objects databases processes

Supercomputing 99-Portland
What's supercomputing got to do with it?

- Complexity of the information
- Amount of data
- Most applications are trivially parallel

Layers of Information

The same base sequence contains many layered instructions!

- Chromosome structure and function
  - Telomers, centromers
- Gene Regulatory information
  - Enancers, promoters
- Instructions for gene structure
- Instructions for protein
- Instructions for protein post-processing and localization
**Moore’s Law and Genomics**

![Graph showing Spec95 Integer Performance vs. Genbank Search]

**The Shape of the Wave**

- **1999**
  - JGI releases 150 Mbases draft
  - Celera releases the sequence of Drosophila (140 Mb)
  - Public “draft” effort reaches halfway point (1,500 Mb)
  - 20 more Microbial genomes completed (80 Mb but 60,000 genes)
  - First release of Celera “shotgun” (9,000 Mb)

- **2000**
  - JGI releases 150 Mbases draft
  - Public “draft” completed (1,500 Mb)
  - Mouse “draft” begins (500 Mb - comparisons with human)
  - Two more Celera shotgun releases (18,000 Mb)
  - 40 more Microbial genomes sequenced (160 Mb - 120,000 genes)
CPU Requirements

- Current annotation
  - 250 Mbases DNA yield ~125 Gbytes of data
  - It takes ~ 7.5 days on 20 workstations ~3,600 nhr

- Celera Data
  - 9 Gbases (36x) in small pieces every 3 months ~2,000 hr.
  - Analysis time approx. quadratic (1300x)
  - 1,300 x 3,600 nhr / 2,000 hr. = 2,340 nodes

- Celera Sequencing
  - Assembly of 1.7 Million reads in 25 hrs
  - Annotation 8-10 Mbases per months with 6 FTE
  - Assembly of Human Genome: expected ~ 3 months

Projected Base Pairs

- The amount of digital data necessary to store 10^16 bases of DNA is only a fraction of the data necessary to describe the world's microbial biodiversity at one square meter resolution.

Projected size of the sequence database, indicated as the number of base pairs per individual medical record in the US.
Sequence Assembly

- Complexity
  - Adding a day’s read of 100 Mb to a billion base pairs of contig would require 100 Pops operations
  - A 1 Tops machine would take about one day to process 100 Mbases

Assembly / Integration / Modeling

- BAC end integration
  - JGI draft (1st half) = 300 Pops
  - First Celera release requires = 3,000 Pops
- Draft and whole genome shotgun integration
  - JGI draft (1st half) + Celera first release = 1,300 Pops
- Gene modeling
  - Celera first release (9Gbases) - 1 day of Paragon time
- Placing STSs
  - JGI draft requires = 9 Pops
  - Celera first release = 90 Pops
Data Transfer

<table>
<thead>
<tr>
<th>Mbytes/sec</th>
<th>year</th>
<th>month</th>
<th>week</th>
<th>day</th>
<th>12 hours</th>
<th>1 hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.03</td>
<td>0.39</td>
<td>1.65</td>
<td>11.60</td>
<td>23.10</td>
<td>27.70</td>
<td></td>
</tr>
</tbody>
</table>

Expect 1999 rate = 100Mbase/day  
100 Mbase = 800 Gbytes

Challenges

- Discovering new biology
- Lack of software integration
- Beginning to build high-performance applications
- Shortage of personnel
Comparative Genome Analysis

aful
aful
bbur
bsub
ecoli
hinf
hpy/28695
mgen
mjan

Supercomputing 99-Portland

Alternatively Spliced?

Supercomputing 99-Portland
One Gene - Many Proteins

As many as 30% of human genes, in particular structural genes, may be alternatively spliced.

9p21 Gene Cluster is a Nexus of the Rb and p53 Pathways

- Same partial nucleotide sequence
- Different amino acid sequence

Cell Cycle Progression
Credits

- NERSC / LBNL
  - John Conboy
  - Donn Davy
  - Inna Dubchak
  - Sylvia Spengler
  - Denise Wolf
  - Eric P. Xing
  - Manfred Zorn

- ORNL
  - Ed Uberbacher
  - Richard Mural
  - Phil LoCascio
  - Sergey Petrov
  - Manesh Shah
  - Morey Parang

Protein Fold Recognition, Structure Prediction, and Folding

Teresa Head-Gordon
Physical Biosciences and Life Sciences Divisions
Lawrence Berkeley National Laboratory

November 15, 1999
Protein Fold Recognition, Structure Prediction, and Folding

(1) Drawing analogies with known protein structures
   - Sequence homology, Structural Homology
   - Inverse Folding, Threading
(2) Ab initio folding: the ability to follow kinetics, mechanism
   - Robust objective function
   - Severe time-scale problem
   - Proper treatment of long-ranged interactions
(3) Ab initio prediction: the ability to extrapolate to unknown folds
   - Multiple minima problem
   - Robust objective function
   - Stochastic Perturbation and Soft Constraints
(4) Simplified Models that Capture the Essence of Real Proteins
   - Lattice and Off-Lattice Simulations
   - Off-Lattice Model that Connect to Experiments: Whole Genomes?

What is a protein?

A biopolymer which is distinct from a heteropolymer in one very important way
- It’s 3-D structure is uniquely tailored to perform a specific function

NMR, X-ray and electron crystallography solve structures slowly (1/2-3 yrs.)
The "Beads" are Chemically Complex Structures

Leucine (LA Ala)

Glutamic Acid (GAQA)

Glutamic Acid (GAQA)

Protein Fold Recognition: Threading

Sequence Assignments to Protein Fold Topology (David Eisenberg, UCLA)

Take a sequence with unknown structure and align onto structural template of a given fold

Score how compatible that sequence is based on empirical knowledge of protein structure

Right now 25-30% of new sequences can be assigned with high confidence to fold class

100,000's of sequences and 10,000's of structures (each of order 10^3-10^4 amino acids long)
Protein Fold Recognition: Threading

Computational Approach:

*Dynamic programming:* capable of finding optimal alignments if optimal alignments of subsequences can be extended to optimal alignments of whole objective functions that are one-dimensional $E=V_i + \Sigma V_{gap}$

*Complexity:* all to all comparison of sequence to structure scales as $L^2$ Whole human genome: $10^{15}$ flops

Improve Objective function:

*Take into account structural environment*

3D → 1D: dynamic programming, $L^2$

*Build pairwise or multi-body objective function*

NP-hard if: variable-length gaps and model nonlocal effects such as distance dependence
Recursive dynamic programming, Hidden markov models, stochastic grammers

*Complexity:* all to all comparison of sequence to structure scales as $L^3$
Whole human genome: $\sim 10^{16}$ flops

---

Computational Protein Folding

One microsecondsimulation of a fragment of the protein, Villin. (Duan & Kollman, Science 1998)

(1) robust objective function ✓
all atom simulation with molecular water present: some structure present

(2) severe time-scale problem ✓
required $10^8$ energy and force evaluations: parallelization (spatial decomposition)

(3) proper treatment of long-ranged interactions X
 cut-off interactions at 8Å, poor by known simulation standards

(4) Statistics (1 trajectory is anecdotal) X
Many trajectories required to characterize kinetics and thermodynamics

---

Supercomputing 99-Portland
Computational Protein Folding

(1) Size-scaling bottlenecks: Depends on complexity of energy function, V

Empirical (less accurate): CN^2; ab initio (more accurate): CN^3 or worse; e^<<C

empirical force field used

"long-ranged interactions" truncated so eM^2 scaling; M < N

spatial decomposition, linked lists

(2) Time-Scale of motions bottlenecks (Δt)

\[ r_i(t + Δt) = 2r_i(t) - r_i(t - Δt) + \frac{f(t)}{m_i} Δt^2 + O(Δt^3) \]

\[ f_i = m_i \dot{r}_i = -\nabla V(r_1, r_2, ..., r_N) \]

Use timestep commensurate with fastest timescale in your system

bond vibrations: 0.01Å amplitude: 10^{-15} seconds (1fs)

Shake/Rattle bonds (2fs)

Multiple timescale algorithms (~5fs) (not used here)

1 Microsecond simulation of Villin Headpiece in Water

Generate 10^9 steps; Assume 1 teraflop machine; 1000 Flops per energy/force evaluation

N^2 evaluation of energy & forces

N evaluation of energy & forces

Ewald Sums:

\[ q_i = \sum_{j \neq i} \left( \sum_{k \in \mathbb{Z}} q_j \sum_{n} \exp \left( -k_j^2 / 4k^2 \right) \cos \left( k \cdot r_{ij} \right) \right) + V_{\text{self}} \]

- Particle Mesh Ewald (N)
- Spatial Decomposition in r-space; Parallelization of FFT's in k-space
- Evaluate full Ewald sum in r-space using FMM techniques
Ab Initio Protein Structure Prediction

Primary Sequence and an Energy function → Tertiary structure

Empirical energy functions:

(1) Detailed Atomic description: leads to enormous difficulties!

\[ V_{MM} = \sum_i k_i (\theta_i - \theta_0)^2 + \sum_i k_i (\phi_i - \phi_0)^2 + \sum_i k_i (\tau_i - \tau_0)^2 + \]

\[ \sum_i k_i [1 + \cos(\alpha_i + \delta)] + \sum_i \sum_{j<i} \left( \frac{q_i q_j}{r_{ij}} + \frac{C_i}{r_{ij}^{12}} - \frac{\sigma_i}{r_{ij}} - \frac{\sigma_i}{r_{ij}^{12}} + \Delta \alpha \right) \]

(1) Multiple minima problem is fierce

Find a way to effectively overcome the multiple minima problem

(2) Objective Functions: Replaceable algorithmic component?

Global energy minimum should be native structure, misfolds higher in energy

The Objective (Energy) Function

Empirical Protein Force Fields: AMBER, CHARMM, ECEPP

“gas phase”


α-helical sequence/β-sheet structure β-sheet sequence/α-helical structure

Energies the same! Makes energy minimization difficult!

Add penalty for exposing hydrophobic surface: favors more compact structures

\[ E_{native} < E_{misfold} \text{ for a few test cases} \]

Solvent accessible surface area functions: Numerically difficult to use in optimization
Hydration Forces from Experiment/Simulation and Optimization

Find model $g_c(r)$ that best reproduces excess experimental signal, $I_{cc}(Q)$

$W(r)$ is "potential of mean force" between two hydrophobic solutes

(Feature Article, J. Phys. Chem., 1999)

$V = \text{AMBER} + (\text{predicted helices fixed}) + W(r) \text{ like that from experiment}$

Global optimization can find no lower energy structures than crystal structures

1pou (72 aa), 3icb (77 aa), 2utg_A (70 aa), 3cln (145)

Neural Networks for $2^\circ$ Structure Prediction

- Input units represent amino acid sequence
- Hidden units map sequence to structure
- Output Units represent secondary structure class (helix, sheet, coil)
- Weights are optimizable variables that are trained on database of proteins

Poorly designed networks result in overfitting, inadequate generalization to test set

Neural network design
- input and output representation
- number of hidden neurons
- weight connection patterns that detect structural features
Neural Network Results

No sequence homology through multiple alignments

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total predicted correctly = 66%</td>
<td>Total predicted correctly = 62.5%</td>
</tr>
<tr>
<td>Helix: 51% $C_\alpha=0.42$</td>
<td>Helix: 48% $C_\alpha=0.38$</td>
</tr>
<tr>
<td>Sheet: 38% $C_\beta=0.39$</td>
<td>Sheet: 28% $C_\beta=0.31$</td>
</tr>
<tr>
<td>Coil: 82% $C_\epsilon=0.36$</td>
<td>Coil: 84% $C_\epsilon=0.35$</td>
</tr>
</tbody>
</table>

Network with Design: Yu and Head-Gordon, Phys. Rev. E 1995

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total predicted correctly = 67%</td>
<td>Total predicted correctly = 66.5%</td>
</tr>
<tr>
<td>Helix: 66% $C_\alpha=0.52$</td>
<td>Helix: 64% $C_\alpha=0.48$</td>
</tr>
<tr>
<td>Sheet: 63% $C_\beta=0.46$</td>
<td>Sheet: 53% $C_\beta=0.43$</td>
</tr>
<tr>
<td>Coil: 69% $C_\epsilon=0.43$</td>
<td>Coil: 73% $C_\epsilon=0.44$</td>
</tr>
</tbody>
</table>

Combine networks of Yu and Head-Gordon with multiple alignments

Supercomputing 99-Portland

Neural Network Predictions As Soft Constraints In Local Optimization

Make neural network prediction of 2<sup>n</sup> structure for each amino acid

Network Output: Helix ($P_\alpha$), Sheet ($P_\beta$), Coil ($P_\gamma$)

$P_\alpha$ = probability of being helix
$P_\beta$ = probability of being sheet

Optimize on following energy surface:

$$\text{Bias} = V_{MM} + V_{\psi\psi} + V_{HR}$$

$$\psi = k_\phi [1 - \cos(\phi - \phi_p)] + k_\psi [1 - \cos(y - y_p)]$$

$V_{HR} = q_{ij} r_{ij}$

$\phi_p$ and $y_p$ define perfect helix values
predictions define $k_\phi$, $k_\psi$, and $q_i$

Using optimized structure from $V_{bias}$

optimize on $V_{MM}$ (AMBER: unbiased objective function)

Supercomputing 99-Portland
Neural Networks Used To Guide Global Optimization Methods

Generate expanded tree of configurations
Predicted coil residues: generate random, dissimilar sets of $\phi_0$ and $\psi_0$

Explore tree configuration in depth:
Global Optimization in sub-space of coil residues: walk through barriers, move downhill

Neural Networks Used To Guide Stochastic Perturbation Algorithm

Stochastic/perturbation in sub-space of dihedral angles predicted to be coil
(1) Local minimization of a set of start points in sub-space
(2) Define a critical radius
$$r_k = \left[ \frac{1}{\pi} \left( 1 + \frac{n}{2} V \sigma \log \rho \right) \right]^{1/n}$$
a measure of whether a point is within a basis of attraction
(3) Generate many sample points in sub-space volume, $V$
(4) Evaluate r.m.s. between new sample points and minimizers of (1)
   If (r.m.s. $< r_k$) ignore this sample point
(5) Minimize sample points not in any critical distance and merge into (1)

Choose new set of dihedral angles and repeat

Probabilistic theoretical guarantees of global optimum in sub-spaces
Global optimization by solving a successive series of global optimum in sub-spaces?
Hierarchical Parallel Implementation of Global Optimization Algorithm

Static vs. Dynamic Load Balancing of Tasks

Central Processor

\[ \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \]

\[ W_{1,1} \rightarrow W_{1,11} \quad W_{2,1} \rightarrow W_{2,11} \quad W_{3,1} \rightarrow W_{3,11} \quad W_{4,1} \rightarrow W_{4,11} \quad W_{5,1} \rightarrow W_{5,11} \]

Central Processor: Assigns starting coordinates to GOPT's

Task time is highly variable

GOPT's: Divide up sub-space into \( N \) regions for global search

Task time is variable

Workers: Generate sample points; find best minimizer in region

(Number of workers depends on sub-space)

Dynamical load balancing of tasks: reassigning GOPT/workers to GOPT/workers

Gain in efficiency of a factor of 5-10

---

Global Optimization Predictions of \( \alpha \)-Helical Proteins

Crystal (left), Prediction (right)

R.M.S. 7.0Å

\[ \Rightarrow \]

1pou: 72 aa DNA binding protein

\[ \Rightarrow \]

2utg_A: 70aa \( \alpha \)-chain of uteroglobin:

Prediction (left) and crystal (right)

R.M.S. 6.3Å

Still have not reached crystal energy yet!

---

Supercomputing 99-Portland
Simplified Models for Simulating Protein Folding

Simplifies the "real" energy surface topology sufficiently that you can do

1. Statistics ✓
   - Can do many trajectories to converge kinetics and thermodynamics
2. Severe time-scale problem ✓
   - Characterize full folding pathway: mechanism, kinetics, thermodynamics
3. Proper treatment of long-ranged interactions ✓
   - All interactions are evaluated; no explicit electrostatics
4. Robust objective function?
   - Good comparison to experiments

α/β Protein Model Resembling IgG-binding Proteins L and G

- Folding is highly cooperative, chain collapse accompanying folding.
- Two parallel folding pathways:
  - One pathway contains an intermediate—protein G
  - One pathway contains no intermediates—protein L.
- Sequence mutations affecting secondary structure propensities
  - Similar to mutational experiments on Protein G & L
- Same Hamiltonian can model all-β (SIH3) and all-α proteins (four helix bundles)

Computational Complexity of Simplified Models for Protein Folding

Thermodynamics of the folding process are characterized using multi-histogram method: complexity increases with multiple order parameters constant-temperature Langevin simulations
Folding kinetics are characterized by tabulating mean-first passage times, and temperature scans
One week using two Compaq/Dec EV10000 (~50 specfp95) per protein sequence 100,000 sequences for Human Genome; Ample mutational study data

Acknowledgements for THG Research in Computational Biology at LBNL

Silvia Crivelli, Physical Biosciences and NERSC Divisions, LBNL.
Jon M. Sorensen, NSF Graduate Fellow, Dept. Chemistry UCB
Greg Hura, Graduate Group in Biophysics, UCB
Alan K. Soper, Rutherford Appleton Laboratory, UK
Alexander Pertsemlidis, Dept. of Biochemistry, U. Texas Southwestern Medical Center
Robert M. Glaser, Mol. & Cell Biology, UCB and Life Sciences Division, LBNL

Funding Sources:
AFOSR, DOE (MICS), DOE/LDRD (LBNL), NIH, NERSC for cycles
Structure-Based Drug Discovery

Brian K. Shoichet, Ph.D
Northwestern University, Dept of MPBC
303 E. Chicago Ave, Chicago, IL 60611-3008
Nov 15, 1999

Problems in Structure-Based Inhibitor Discovery & Design

- Balance of forces in binding
  - Energies in condensed phases
    - interaction energies
    - desolvation
  - Problem scales badly with degrees of freedom
    - Configuration
      - configs $\alpha$ (prot-features)$^d$ X (lig-features)$^d$
    - Conformation
      - Ligand & Protein, confs $\alpha$ 3burds X 3burds
- Sampling chemical space (scales very badly)
- Defining binding sites
The Pros & Cons of Proteins

18 - Crown-6

sulfate binding protein

Conserved Residues, Ordered Structure, Function Unknown

Supercomputing 99-Portland
Inhibitor Discovery or Design?

- **Design ligands**
  - Ludi (Bohm)
  - Grow (Moon & Howe)
  - Builder (Roe & Kuntz)
  - MCSS-Hook (Miranker & Karplus)
  - SMOG (DeWitte & Shakhnovitch)
  - Others...

- **Discover Ligands**
  - DOCK (Kuntz, et al., Shoichet)
  - CAVEAT (Bartlett)
  - Monte Carlo (Hart & Read)
  - AutoDock (Goodsell & Olson)
  - SPECTTOPE (Kuhn et al)
  - Others...

Screening Databases by Molecular Docking

- **Dock into site**
- **Calculate energies**
- **Test highscoring molecules**
- **Structure determination**
- **Now inhibitor design**

© Chemistry & Biology, 1996
Database Screening Using DOCK

Each molecule is fit into the binding site in multiple orientations. Multiple conformations of each ligand are considered. Each orientation is evaluated for complementarity, using van der Waals and electrostatic interaction energies. Solvation energies are subtracted.

The inhibition constants of the best fitting molecules are established in an enzyme assay.

Inhibitor-receptor complex structures are determined. New interactions with the enzyme are targeted.

Novel Ligand Discovery Using Molecular Docking

<table>
<thead>
<tr>
<th>Receptor</th>
<th>Lead from molecular docking</th>
<th>Receptor</th>
<th>Lead from molecular docking</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV protease</td>
<td><img src="image1" alt="HIV protease" /></td>
<td>HOXPRTase</td>
<td><img src="image2" alt="HOXPRTase" /></td>
</tr>
<tr>
<td>Dihydrofolate synthase</td>
<td><img src="image3" alt="Dihydrofolate synthase" /></td>
<td>RNA</td>
<td><img src="image4" alt="RNA" /></td>
</tr>
<tr>
<td>Neurolysin</td>
<td><img src="image5" alt="Neurolysin" /></td>
<td>Zn β-lactamase</td>
<td><img src="image6" alt="Zn β-lactamase" /></td>
</tr>
<tr>
<td>Cerebral elastase</td>
<td><img src="image7" alt="Cerebral elastase" /></td>
<td>Thrombin</td>
<td><img src="image8" alt="Thrombin" /></td>
</tr>
<tr>
<td>Neutrophil elastase</td>
<td><img src="image9" alt="Neutrophil elastase" /></td>
<td>AmpC β-lactamase</td>
<td><img src="image10" alt="AmpC β-lactamase" /></td>
</tr>
<tr>
<td>Neutrophil protease</td>
<td><img src="image11" alt="Neutrophil protease" /></td>
<td>Thymidylate synthase</td>
<td><img src="image12" alt="Thymidylate synthase" /></td>
</tr>
<tr>
<td>CD4 peptide</td>
<td>unpublished</td>
<td>HOXPRTase</td>
<td>unpublished</td>
</tr>
</tbody>
</table>
de Novo Structure Prediction: blip/tem-1

Ligand Flexibility:
Conformational Ensembles

Generate an ensemble
dock it into the site
## Conformational Ensembles vs. Brute Force

![Graph showing comparison between single, multiple, and ensemble methods for various proteins.](image)

## Database Docking

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Number of Conf.</th>
<th>Time (hrs.)</th>
<th>Score</th>
<th>RMS (Å)</th>
<th>Rank in Database</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single Conformation Database</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complexed DHFR</td>
<td>5,761</td>
<td>0.58</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Uncomplexed DHFR</td>
<td>5,761</td>
<td>1.40</td>
<td>91.9</td>
<td>8.32</td>
<td>16.09%</td>
</tr>
<tr>
<td>Complexed TS</td>
<td>281</td>
<td>0.31</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Uncomplexed TS</td>
<td>281</td>
<td>0.51</td>
<td>-8.3</td>
<td>3.67</td>
<td>97.15%</td>
</tr>
</tbody>
</table>

| **Multi Conformation Database** |                 |             |         |         |                  |
| Complexed DHFR  | 867,822         | 0.94        | -12.5   | 1.20    | 99.33%           |
| Uncomplexed DHFR| 867,822         | 2.96        | -7.4    | 1.34    | 98.83%           |
| Complexed TS    | 88,487          | 0.27        | -89.2   | 0.77    | 99.62%           |
| Uncomplexed TS  | 88,487          | 0.18        | -31.5   | 2.71    | 99.24%           |

| **Full Multi Conformation Database** |                 |             |         |         |                  |
| Complexed DHFR  | 33,717,639      | 26.50       | -12.5   | 1.20    | 99.72%           |
| Complexed TS    | 33,715,748      | 80.90       | -89.2   | 0.77    | 99.93%           |
Hierarchical Docking

Flexible docking: Hierarchical docking:
27 confs 27 confs
3 atoms 3C + 3A + 9B
81 atom positions 15 atom positions

Correcting for Ligand Solvation Energies

$$\Delta G_{\text{bind}} = \Delta G_{\text{interact}} - \Delta G_{\text{solv, L}} - \Delta G_{\text{solv, R}}$$

$$\Delta G_{\text{interact}} = \sum (q_i P_i + v_i P_v)$$

$$\Delta G_{\text{elec,solv}} = (q^2/2r) \left(1/D_0 - 1/D_w\right)$$

$$= (1/D_0 - 1/D_w)/2 \sum \sum Q_i \delta q_j$$

$$\bar{M}_{hp} = -621.48 - 25.890 \times \text{area}$$
Solvation Corrections: Benzene Cavity Screen

Hit Rates

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Hit Rate</th>
<th>IC_{50} for</th>
<th>Compounds</th>
<th>Random Hit Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>AmpC (E. coli)</td>
<td>50%</td>
<td>&lt;10 μM</td>
<td>20</td>
<td>??</td>
</tr>
<tr>
<td>HXPR Tase (T. cruzi)</td>
<td>60%</td>
<td>&lt;12 μM</td>
<td>22</td>
<td>??</td>
</tr>
<tr>
<td>Corporate (homology modeled)</td>
<td>2%</td>
<td>???</td>
<td>895</td>
<td>0.04% (per 102,000)</td>
</tr>
</tbody>
</table>
Unmet Challenges

- Better Scoring
  - context dependent desolvation
  - receptor desolvation
  - better force-fields
- Receptor Flexibility
- Combinatorial Chemistry

This work supported by the NIH, Genetics Institute, and Procter & Gamble

Cellular Network Analysis

Adam Arkin
Physical Biosciences
Lawrence Berkeley National Laboratory
Bioengineering and Chemistry
University of California, Berkeley
11/15/99
Asynchronous Digital Telephone Switching Circuit

Full knowledge of parts list
Full knowledge of "device physics"
Full knowledge of interactions

No one fully understands how this circuit works!!
Its just too complicated.

Designed and prototyped on a computer (SPICE analysis)
Experimental implementation fault tested on computer

Asynchronous Analog Biological Switching Circuit

Partial knowledge of parts list
Partial knowledge of "device physics"
Partial knowledge of interactions

No one fully understands how this circuit works!!
Its just too complicated.

We need a SPICE-like analysis for biological systems

Analysis of Cell Function

The challenge is to integrate data from all levels to produce a description of cellular function.

There are challenges in:

- Systematization and structuring of data
- Serving and query this data
- Representing the data
- Building multiscale, multiresolution models
- Dynamic and static analysis of these models

Pay-off in

- Industrial bioengineering
- Rational pharmaceutical design
- Basic biological understanding
Spatiotemporally resolved pictures of developmental processes take up Gigs of storage.

Analyses take days-weeks.

Models are in early days.

Each of those little bright spots contains networks vastly more complicated than those on the last slide!
Heterogeneity of Data

Data are:
1) Qualitative->Quantitative
2) Collected at many levels
3) Of heterogeneous structure
4) Of heterogeneous availability

Challenge:
Optimal use of available data to make predictions about cell function and failure.

Tools for “multilevel” analysis

Cellular networks
Physical properties
Finding Parts
Why now?

- Genome projects are providing a large (but partial) list of parts
- New measurement technologies are helping to identify further components, their interactions, and timings
  - Gene microarrays
  - Two-Hybrid library screens
  - High-throughput capillary electrophoresis arrays for DNA, proteins and metabolites
  - Fluorescent confocal imaging of live biological specimens
  - High-throughput protein structure determination
- Data is being compiled, systematized, and served at an unprecedented rate
  - Growth of GenBank and PDB > polynomial
  - Proliferation of databases of everything from sequence to confocal images to literature
- The tools for analyzing these various sorts of data are also multiplying at an astounding rate

SPICE Tools for Biology?

BioSpice: A Web-Servable, Biologist-Friendly, database, analysis and simulation interface was developed into a true beta product.

- Interfaces to ReactDB, MechDB, and ParamDB.
- With Kernel, performs basic: flux-balance analysis, stochastic and deterministic kinetics, Scientific Visualization of results.
- Notebook/Kernel design optimized for distributed computing.
Components of Bio/Spice

An Example of "Device Physics"

Simulation methodology for full-up simulation of chemical Markov-Process scales exponentially with number of reactions
Complexities of Cellular Function

This is approximately 1/3 of just the initiation of the sporulation program from *Bacillus subtilis*.

There are over 100 proteins, 40 genes, 300 reactions for which data is available.

The total data on just this process is a tens of Gigs and it is incomplete. Microarray and microscope data are added 100 Megs per week. Model builders need to query this data and arrange it for simulation. Simulations must be run under many different condition and hypotheses.

The Need for Advanced Computing

**Data Handling:**
- The total data necessary for network analysis is huge.
- By nature it will be distributed and heterogeneous
- We need:
  - Database standard and new query types
  - Means of secure, fast transmission of information
  - Means of quality control on data input

**Tool integration:**
- Centralization of computational biology tools and standards
- Ability to use tools together to generate good network hypotheses
- Good quality ratings on Tool outputs

**Advanced Simulation Tools:**
- Fast, distributed algorithms for dynamical simulation
- Mixed mode systems (differential, Markov, algebraic, logical)
- Spatially distributed systems