

# **Basic Biomedical Peta(flops/ops) Challenges**

**Jacob V. Maizel, Jr**

**Chief, Laboratory of Experimental  
and Computational Biology**

**NCI**

**[jmaizel@ncifcrf.gov](mailto:jmaizel@ncifcrf.gov) [www.lecb.ncifcrf.gov](http://www.lecb.ncifcrf.gov)  
[www-fcrdc.ncifcrf.gov](http://www-fcrdc.ncifcrf.gov)**

**Examples from structural biology**

**Genome Sequence Analysis**

**Molecular Structure**

**3-D Heart Model**

# Why Structural Biology?

**A major goal is to understand how the information of the genes is expressed in the molecules of an organism.**

**Knowledge of molecular structure will provide major insights into functional behavior.**

**This may enable beneficial manipulation, as in therapies, or biodesign, or remediation, etc.**

**Computable data are accumulating at a growing rate.**

# Analysis of gene sequences

## Comparative analysis -

query vs db finds similar, known genes  
db vs db classifies the db

## Algorithms and characteristics -

String (4-letter for nucleic acids, 20-letter for proteins)  
Dynamic programming with weighted scores for similarity and gaps

## Complexity -

$O(N \times M)$ ,  $O(N^2)$  where  $N$ ,  $M$  are lengths of sequences

# Analysis of gene sequences(cont)

## Comparison of genomes - brute force approach

Pairwise (e.g., human vs mouse),  $\sim 10^{19}$  elements present machines do 1 elemental comparison / 10-100 clocks would need  $10^5 - 10^6$  petaop-sec.

Multiple simultaneous comparisons become very hard ( $N^k$ ) where N is length of each sequence, k is number of different sequences.

But -

Strong heuristics are available and allow acceptable results on high end workstation/servers.

Problem shrinks if simplifying classifications are found.

# Structure prediction from sequence (a.k.a. the folding problem)

Present experimental methods are limited by availability of crystals, size and solution properties, etc., although shotgun crystallization, hi-intensity x-ray sources, and direct solving may give great speedup.

Of >100,000 expected unique structures and with 1000s of person-years of experimental effort -

Few high resolution RNA structures from experiment are available

About 300-400 unique protein structures exist in a highly redundant database of about ~10000 X-ray and NMR structures

Improved computer and algorithmic power by many orders of magnitude are needed for data analysis, discovery of principles and prediction.

# RNA structure prediction

Simple rules exist for secondary structure prediction based on so-called complementary interactions, calibrated from thermodynamic studies on model compounds

These structures are useful for interpretation of experimental results and as a basis for higher dimensional prediction

Good parallel dynamic programming methods exist that use  $O(N^2)$  memory and  $O(\sim N^3)$  time on SMVPs, and  $O(\sim N^2)$  on SIMD MPPs.

Foldings of HIV RNA (9218 bases) takes 6 Gops-hr

## RNA structure prediction (cont.)

**Petaops would enable larger and more detailed simulations needed to develop new algorithms for secondary structure, including MC and GA that can accommodate more realistic structures**

**2-D models are valuable starting points for construction of 3-D models that in turn need refinement by molecular mechanical approaches**

# Protein structure prediction

This is considered one of the most important challenges -

For exhaustive conformational search of a small protein of 100 amino acids each with ~10 possible local conformations there are  $\sim 10^{100}$  global shapes. Nature doesn't do this. This is called the Levinthal paradox, for Cyrus Levinthal who first pointed it out.

Lattice models have allowed development of exhaustive searches for small proteins and MC searches using rules on compactness and potentials of mean force to evaluate and constrain searches have some success.

Kollman and colleagues (UCSF) have done a small protein by molecular mechanics with some promising results.

Petaops will be an invaluable tool for exploring algorithms, potentials and other parameters

**At this point I showed the equations used in protein  
molecular mechanics algorithms from  
<http://www.amber.ucsf.edu/amber/eqn.txt>**

**and the paper by Duan and Kollman on microsecond  
dynamics folding of villin fragment at  
<http://www.amber.ucsf.edu/members/yduan/983445.html>**

# Value Added Computation on Experimentally Determined Protein Structures

Computation offers the potential to augment experimental data with -

Dynamics on static x-ray structures.

Placement of H atoms.

Electronic structure to understand catalytic mechanisms and unstable intermediates.

Docking of drugs with enzymes and prediction of interactions with hypothetical drugs.

## Proteins(cont.)

Examples of the need -

For HIV protease and potential drug(i.e., inhibitor) 100 Gflop-hours is needed for  $\sim 10^{-9}$  sec of real time.

We would like milliseconds, requiring  $10^{12}$ -fold more processing speed than at present, but petaflops would put us in the range of doing critical simulations in hours.

Consume significant fractions of existing high end cycles.

The algorithms -

- Scale as  $O(N^2)$ , need floating point, high precision

- Involve updating a global list of atoms having non-bonded interactions.

- Are most popular in dusty-deck versions. Have been parallelized only moderately in forms that don't scale to many processors .

**At this point I introduced some views of web pages and molecular graphics.**

**HIV protease 8hvp.pdb downloaded from [www.pdb.bnl.gov](http://www.pdb.bnl.gov)  
viewed by Rasmol available at  
<http://www.pdb.bnl.gov/pdb-docs/molgraph.html>**

**Electron micrographs of influenza virus downloaded from  
<http://www.uct.ac.za/depts/mmi/stannard/fluvirus.html>,  
and rasmol display of neuraminidase with drug in 1bji.pdb.**

# Electronic structure

These calculations are central to understanding chemical reactions, and are proving fruitful for enzyme mechanisms.

More detailed ones proceed from atomic to molecular orbital descriptions and are often used to parameterize more approximate methods such as molecular mechanics.

These detailed forms are  $O(\sim N^4)$  on the number of electrons.

These calculations are using more of our existing high-end time than molecular mechanics.

The wish is for petaflops at high precision with large memory, and to use existing, commercial codes, which are only moderately parallel on SMPs, then to combine MM and QM.

## Realistic 3-D Heart model

**Peskin and McQueen at Courant Institute have a model based on a fiber representation of the anatomy of the heart and lattice-based Navier-Stokes fluid dynamics for the blood.**

**Scales as  $\sim N^3$  in memory,  $\sim N^4$  in time.**

**One beat requires one Cray C90-week and 50MW using a 128x128x128 lattice for CFD.**

**Petaflops would enable more realistic heart modeling, refinement of parameters and introduction of features such as electrical activity.**

**At this point I showed the web page at  
<http://www.psc.edu/research/graphics/gallery/heart.html>**

# Closing Remarks

**Biomedical computing could use petaflops now in several areas of structural research.**

**Other areas can be imagined to benefit( e.g., genetic simulations, regulatory pathway modeling, real-time medical applications, 3-D imaging, etc.)**

**Only a minority of the small population of computational biologists have a high enough pain threshold to undertake recoding of algorithms for parallel architectures.**

# Petaop(flop) asides (not presented at conference)

**Biocomputing may help us achieve petaops by inspiring radical devices, architectures and programming models.**

**DNA computing by Adelman.**

**Schneider, in the LECB, has patented a molecular flip-flop based on DNA-protein interactions.**

**Protein based memories.**

**Connectionist, and genetics-inspired programs.**

**We have a proof of concept of a petaops biological device -  
The human brain has  $\sim 10^{14}$  synapses that can fire  
at  $>10/\text{sec}$ , and incidentally uses  $<10$  watts in about a 1-2 liter box  
and is  $\sim 90\%$  water.**