BioPilot: Data-Intensive Computing for Complex Biological Systems

Presented by

Nagiza Samatova

Computer Science Research Group
Computer Science and Mathematics Division

Sponsored by

the Department of Energy's Office of Science
Advanced Scientific Computing Research
Scientific Discovery through Advanced Computing

Goals for enabling data-intensive computing in biology

Biological research is becoming a high-throughput, data-intensive science, requiring development and evaluation of new methods and algorithms for large-scale computational analysis, modeling, and simulation

Computational algorithms for proteomics

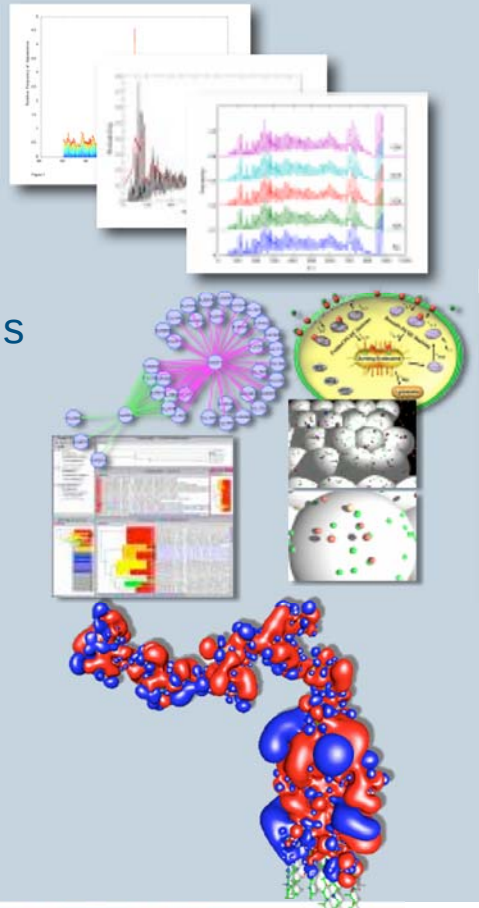
- Improve the efficiency and reliability of protein identification and quantification
- Peptide identification using MS/MS using pre-computed spectral databases
- Combinatorial peptide search with mutations, post- translational modifications and cross-linked constructs

Inference, modeling, and simulation of biological networks

- Reconstruction of cellular network topologies using high-throughput biological data
- Stochastic and differential equation-based simulations of the dynamics of cellular networks

Integration of bioinformatics and biomolecular modeling and simulation

- Structure, function, and dynamics of complex biomolecular system in appropriate environments
- Comparative analysis of large-scale molecular simulation trajectories
- Event recognition in biomolecular simulations
- Multi-level modeling of protein models and their conformational space



Pacific Northwest National Laboratory

OAK RIDGE National Laboratory

Comparative molecular trajectory analysis with BioSimGrid

BioSimGrid
Comparative Trajectory Analysis Tools

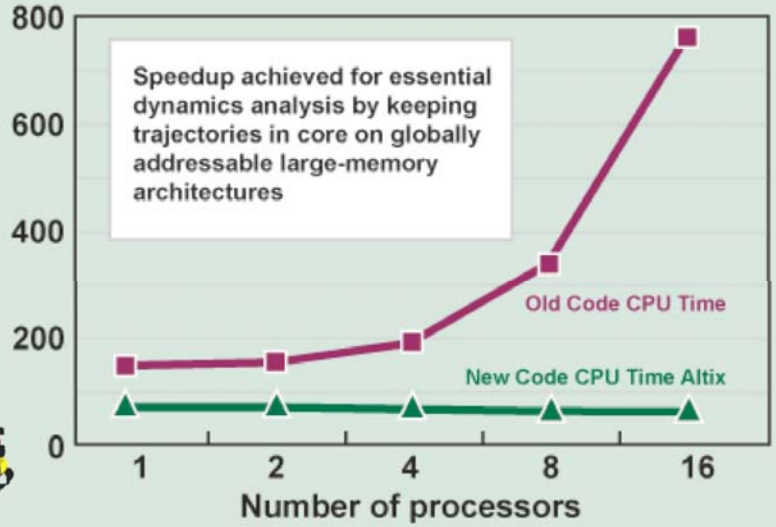
The scientific challenge:

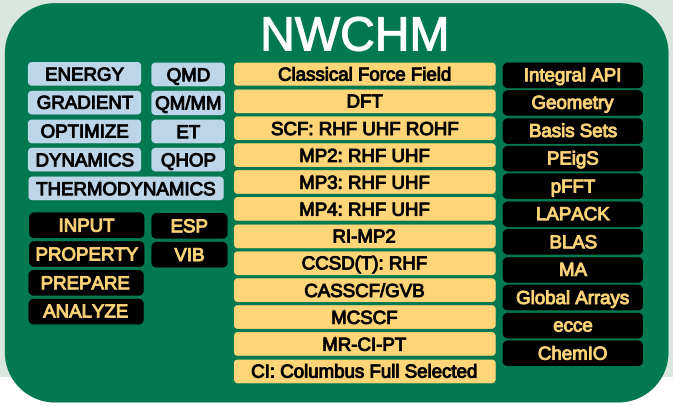
The analysis of molecular dynamics trajectories presents a large, data-intensive analysis problem:

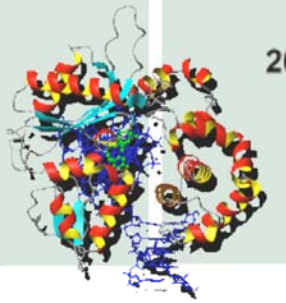
- Routine protein simulations generate TB trajectories.
- Routine analysis tools (ED) require multiple passes.
- Correlation analyses have random access patterns.
- Comparative analyses require multiple/distributed trajectory access.

Integrated molecular simulation and comparative molecular trajectory analysis:

- Enzymatic reaction mechanisms
- Molecular machines
- Molecular basis of signal and material transport







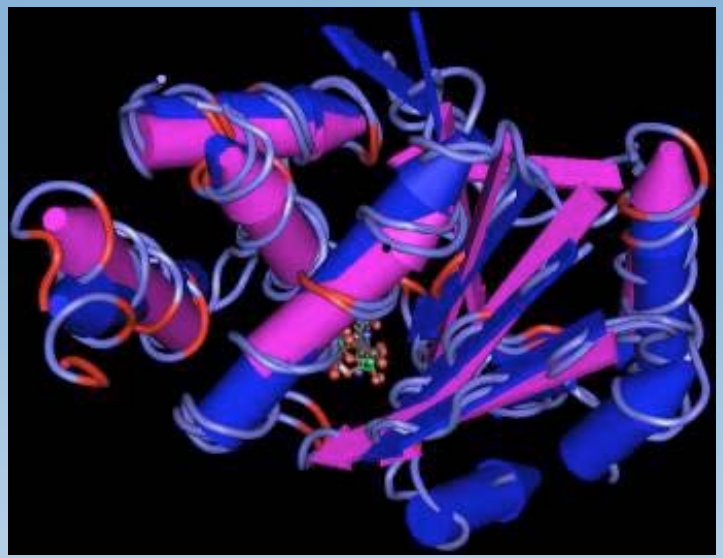
T. P. Straatsma, Computational Biology and Bioinformatics

Pacific Northwest National Laboratory

OAK RIDGE National Laboratory

Building accurate protein structures

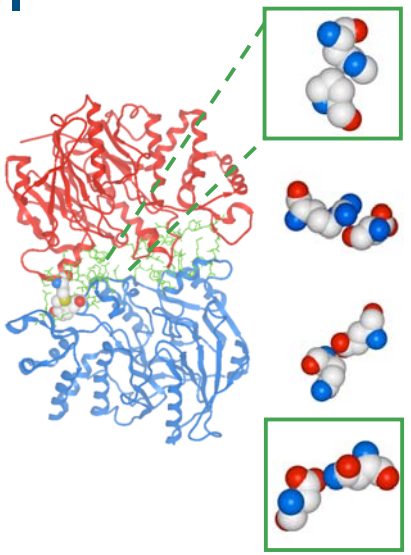
- Homology models can be built from related protein structures, but even with proper alignment, usually have about 4A RMS error–too large to permit meaningful computational simulation/molecular dynamics.
  Goal is to reduce the errors in initial models.
- We multi-align the group of neighbors using sequence and structure information to find the most stable parts of the domain.
  Extracted stable core structure is 20–30% closer in terms of RMSD to the target than to any of the original templates.
- More flexible parts of target are modeled locally by choosing most sequentially similar loops from the library of local segments found in all the homologues.
- A genetic algorithm conformation search strategy using Cray XT4 uses backbone and sidechain angles as parameters.
  Each GA step is evaluated by minimization.

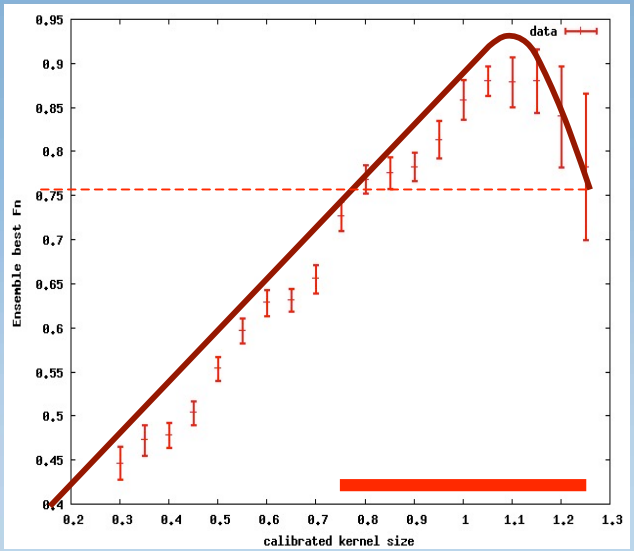


A computed template compared with the target, improved from 3.4A initially to 2.3A

Analysis of ultra-large structural ensembles

- **Ultra-scale docking.** The Bayesian potentials allow exploration of 100s and 1000s of protein complexes instead of one or two.
- **Docking from independently crystallized subunits.** We demonstrated excellent results on complexes reconstructed from ~500 independently crystallized subunits.
- **Whole genome predictions.** Developed technology has been used to predict 1000s of protein complexes on the genome-wide level in several organisms.





Best structure. The set with the largest common kernel always includes nearly the best native structure present in the ensemble of 10,000 folded structures.

Quality estimator. The size of the largest common kernel (obtained from the connectivity graph) provides an excellent estimate of how close the selected structure is to the native ones.

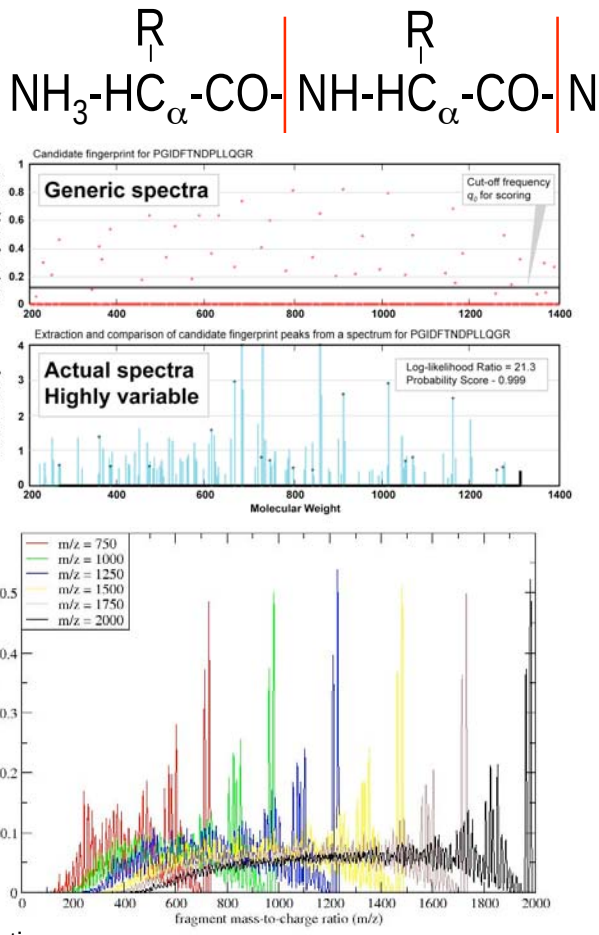
A. Gorin, Computer Science and Mathematics Division

Predictive proteome analysis using advanced computation and physical models

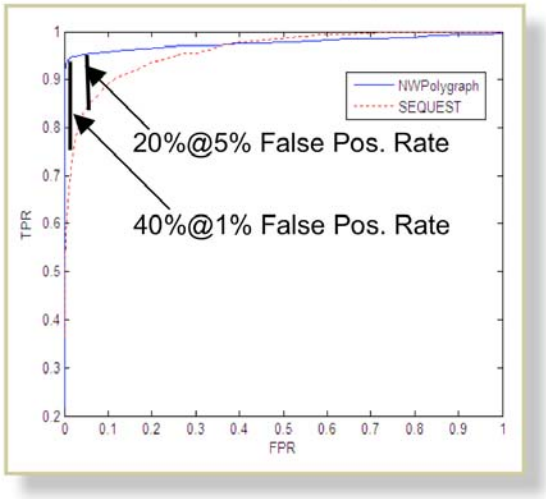
Scientific challenge:

Data sizes: 30,000 to 200,000 spectra from a single experiment to be compared with as many as millions of theoretical spectra

- Computational tools identify only 10–20% of spectra from proteomics analysis.
- Fragmentation patterns are highly sequence-specific.
- Statistical characterization of noise essential.
- Identification rate increased by 23–40%.



Increase in identified peptides



W. R. Cannon, Computational Biology and Bioinformatics

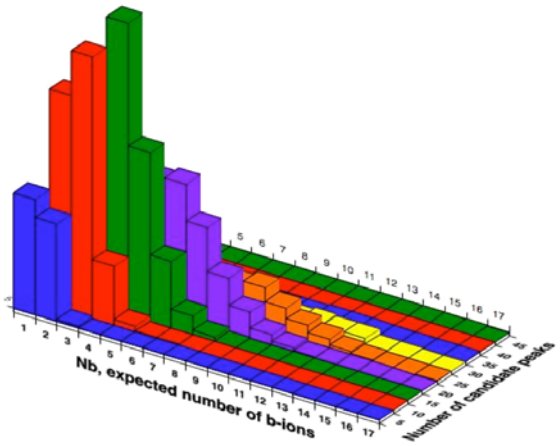
Pacific Northwest National Laboratory

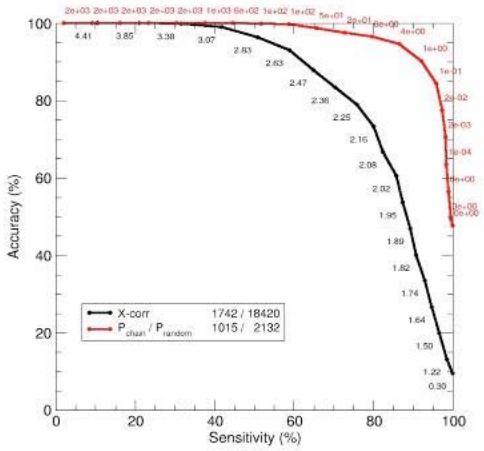
OAK RIDGE National Laboratory

In-depth analysis of MS proteomics data flows

Mass-spectrometry-based proteomics is one of the richest sources of the biological information, but the data flows are enormous (100,000s of samples each consisting of 100,000s of spectra).

Computationally, both advanced graph algorithms and memory-based indexes (> 1 TB are required for in-depth analysis of the spectra).





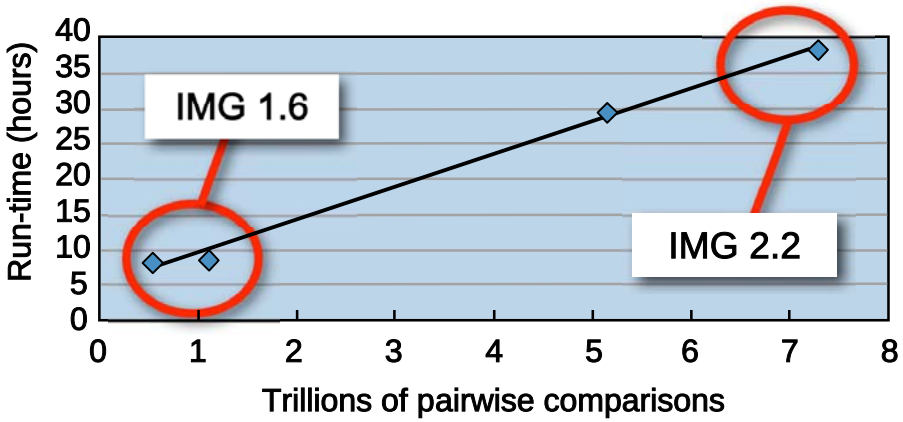
Two principally new capabilities are demonstrated:
- Quantity of identified peptides is 50% increased.
- Among highly reliable identifications, increase is many-fold.

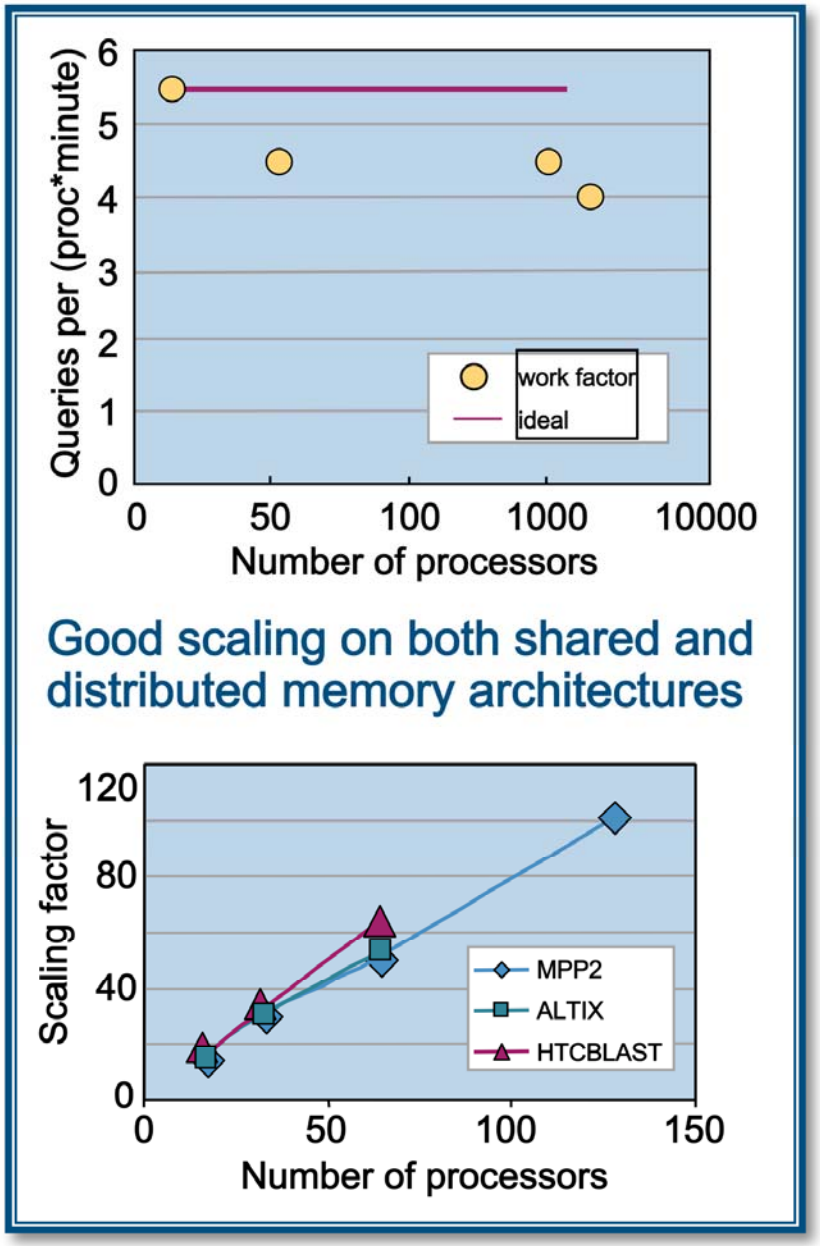
A. Gorin, Computer Science and Mathematics Division

ScalaBLAST

Scientific challenge:

- Standalone BLAST application would take >1 year for IMG (>1.6 million proteins) vs IMG and >3 years for IMG vs nr.
- "PERL script" approach breaks even modest clusters because of poor memory management.



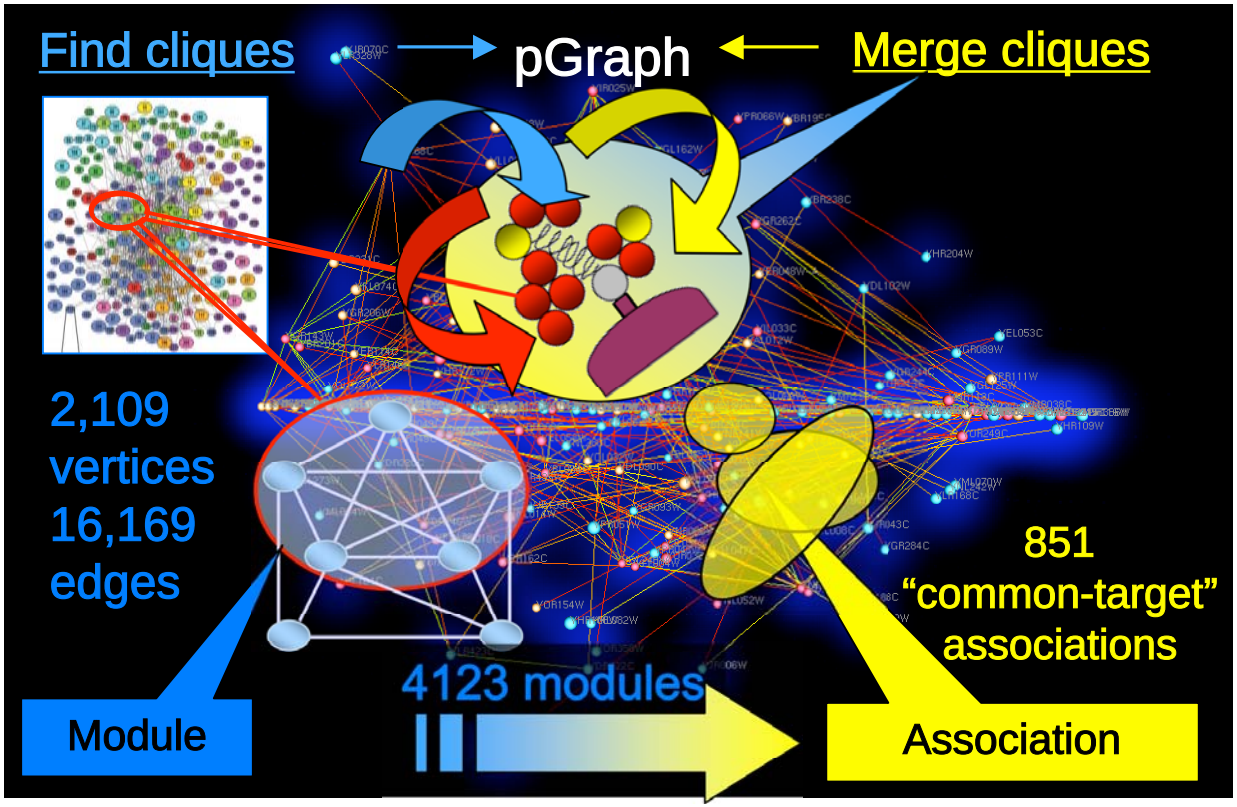


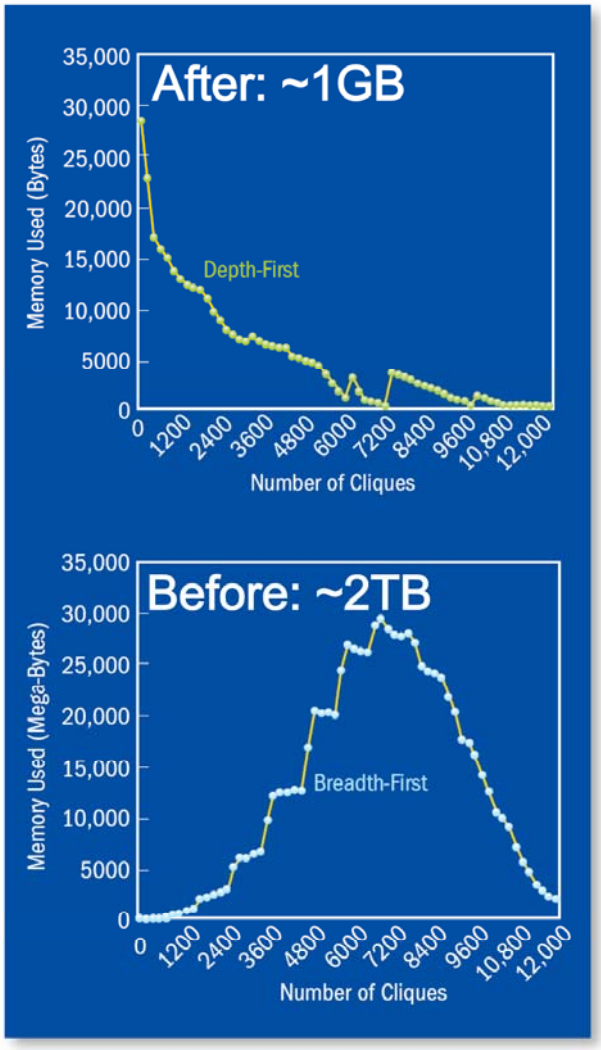
C. S. Oehmen, Computational Biology and Bioinformatics

pGraph: Parallel Graph Library for analysis of biological networks

Major features:

- Memory reduction by 1000 times
- Scalability to 100s of processors
- FPT-based search space reduction theory

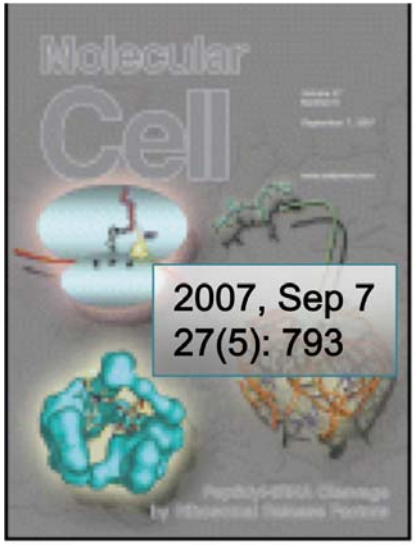


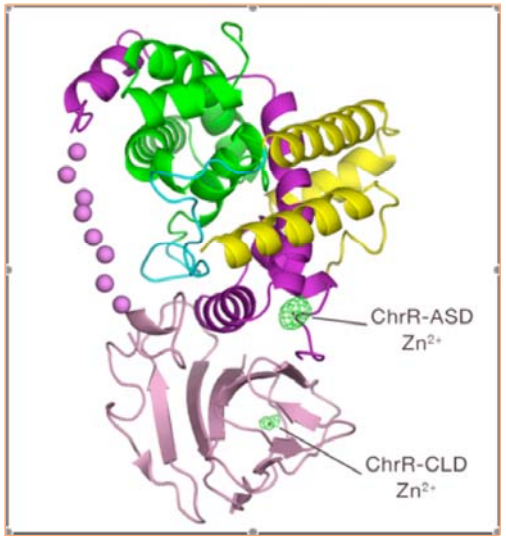


Nagiza Samatova, Computer Science and Mathematics Division

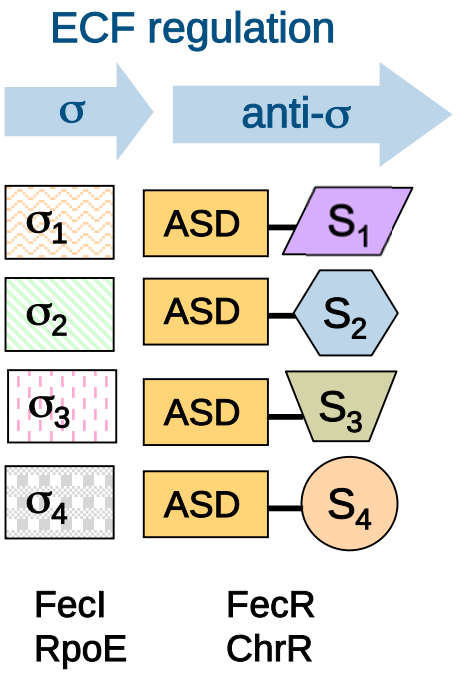
Discovering cell networks with structure and genome context

The structure of the ChrR anti-sigma factor from *Rhodobacter sphaeroides* reveals domains which help explain the combinatoric complexity of gene regulation in response to environmental conditions. An advanced toolkit for graph, genome context, and visual analytics was used to study ECF sigma factor regulation.





3D structure of ChrR anti-sigma regulator



The conserved ASD domain links ~30% of ECF sigma and anti-sigma pairs in the signal transduction cascade, e.g., response to iron and $^1O_2$ with FecIR, RpoE-ChrR proteins.

Many combinations of different sigma (σ) and sensor (S) domains exist. Two levels of positional clustering, (1) domain and (2) gene neighbor, generate vast permutations between protein pairs.

Sofia Heidi, Computational Biology and Bioinformatics

Stochastic simulations of biological networks

Scientific challenge:

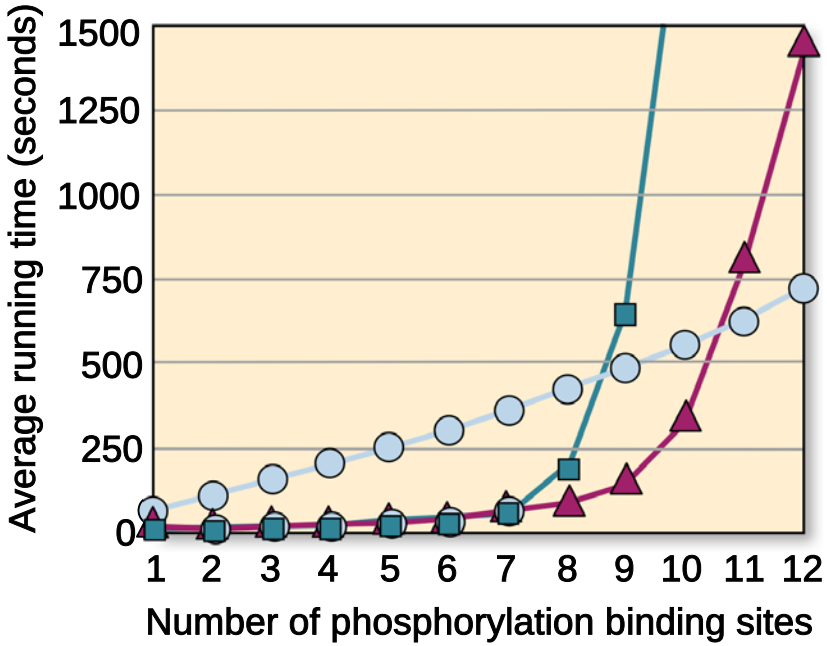
- To model microbial communities, the smallest realistic model would include >109 cells x 100 reactions ~1011 reactions.
- Currently can handle ~104–106 reactions on a single-processor machine.

Probability-weighted dynamic Monte Carlo method…

- *J. Phys. Chem. B* 105, 11026, 2001
- Speedup over exact Gillespie SSA ~25

Multinomial Tau-leaping method:

- Speedups ~40–200



Contacts

Tjerk Straatsma

Principal Investigator
Pacific Northwest National Laboratory
tps@pnl.gov

Nagiza Samatova

Principle Investigator
Oak Ridge National Laboratory
samatovan@ornl.gov

Christine Chalk

Program Manager
U.S. Department of Energy
Office of Science
Advanced Scientific Computing Research, SciDAC
christine.chalk@science.doe.gov

OAK RIDGE National Laboratory