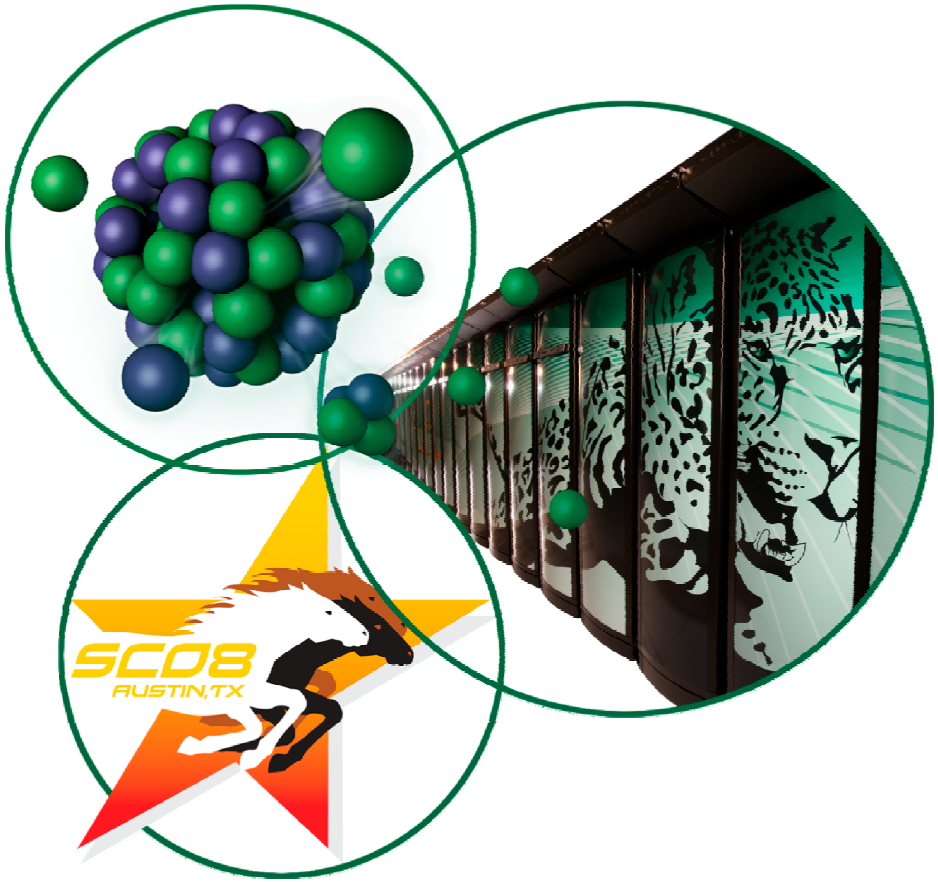
BioPilot: Data-Intensive Computing for Complex Biological Systems



Presented by

Nagiza Samatova

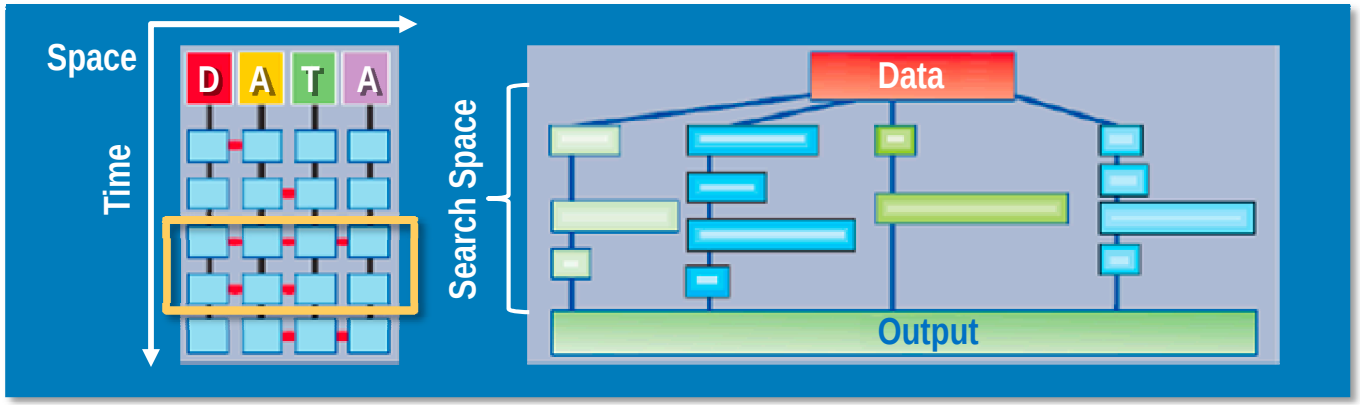
Computer Science Research Group
Computer Science and Mathematics Division

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

Distinct data-access patterns

Data-driven model building requires a different mix of memory, disk storage, and communication trade-offs

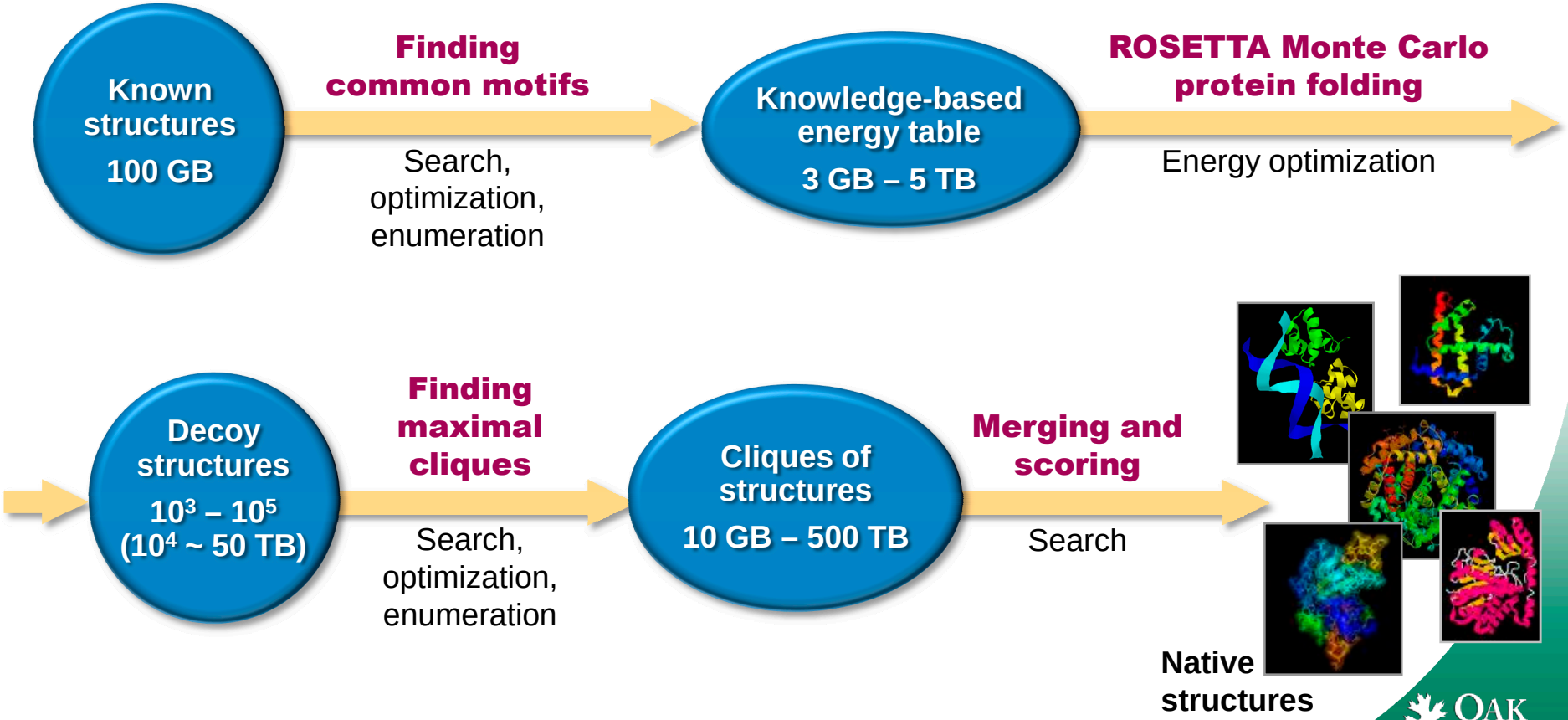
	Simulation	Search	Enumeration
Input	Medium	Huge	Medium
Memory Access	Local (2 Time Steps)	Global (Entire Database)	Exponential Irregular
Output to Disk	Iterative	Irregular	Irregular, Huge
Communication	Intensive	For Scoring	Load Balancing
Arithmetic	Float	Integer (Float)	Integer (Float)



Scientific data flow in structure modeling

Each step is a combinatorial optimization problem with different data-access patterns

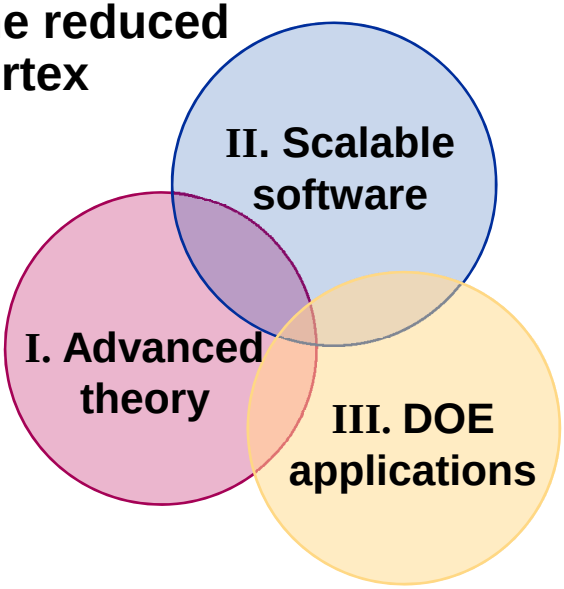
Pipeline: ab initio prediction of Protein 3-D structure



Network modeling: Accomplishments

Aim 1: To advance the theory of search space reduction

- Asymptotically best current algorithm for maximal clique search (MCS) O*((m+1)n)
- Asymptotic maximal clique enumeration (MCE) time reduced to $O^*(3^{c/3})$ from $O^*(3^{n/3})$ where c is the size of the vertex cover (VC)
- VC-constrained MCS of $O^*(3^{m/3}(m+1)^c)$
- Introduced multi-MCS; showed it is W[1]-complete



Aim 2: To develop scalable software

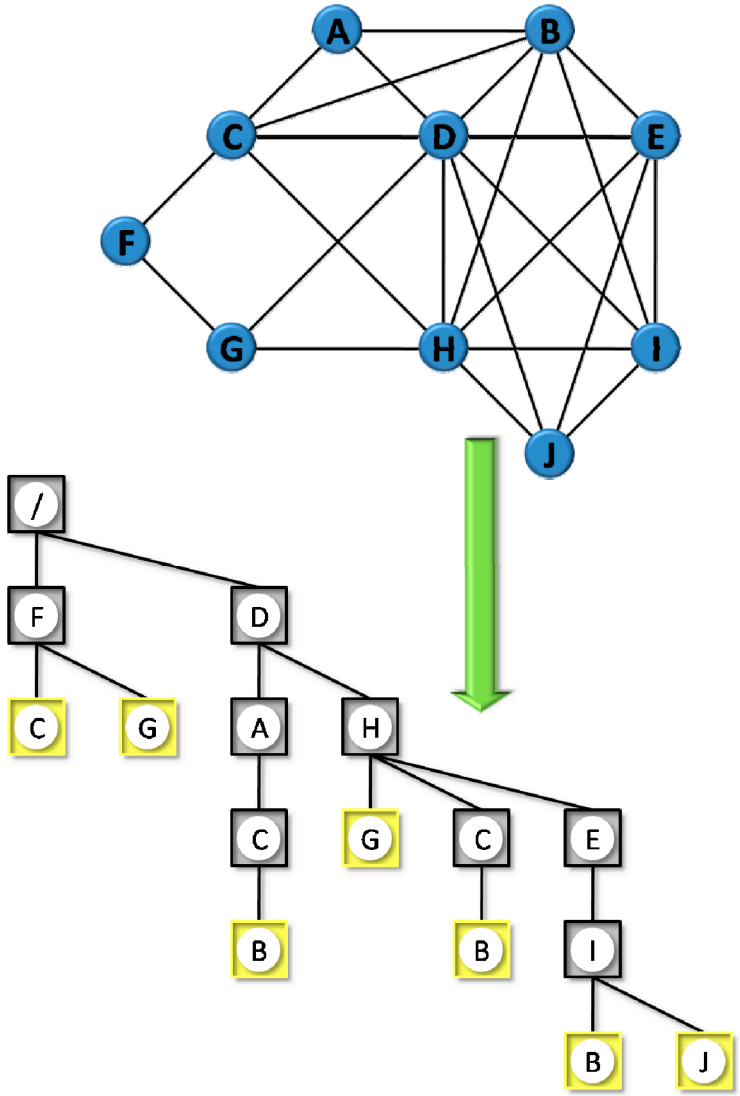
- *FPT-based* pGraph library
- Scaled to thousands of processors
- Reduced memory by thousands

Aim 3: To apply to large-scale DOE-mission problems

- **M. Buchanan:** *Genome-scale reconstruction of* R. palustris *interactome*
- **C. Harwood:** *Aromatic compound degradation pathways in* R. palustris
- **J. Banfield:** *Characterization of acid mine drainage microbial community*

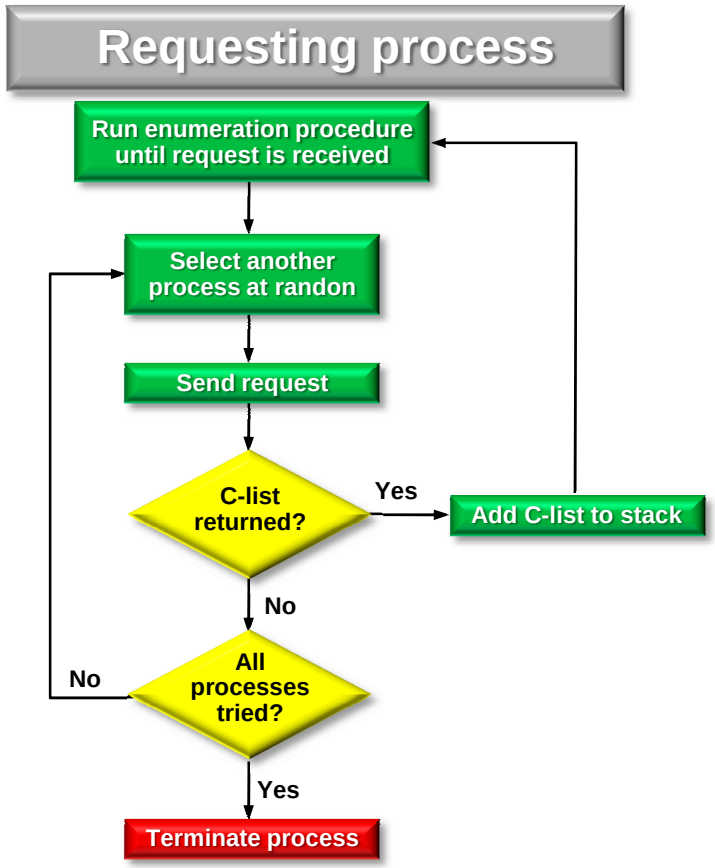
Computational challenges for large networks

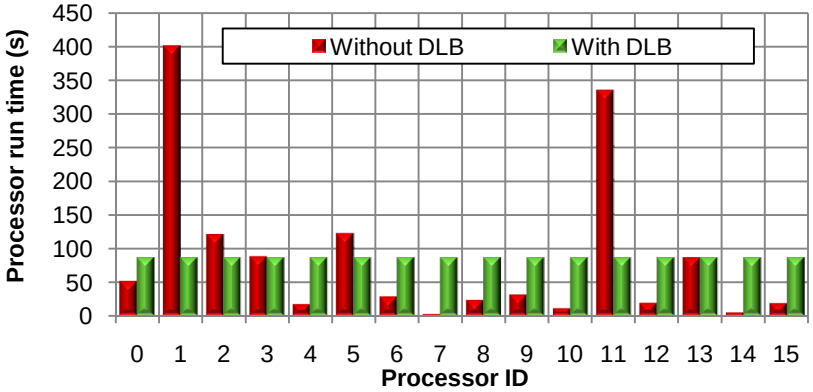
- **Need for parallel MCE**
  - Even the most efficient serial MCE algorithms require inordinate amounts of time to enumerate the maximal cliques in biological networks
- **Parallelization challenges**
  - Data-intensive nature of MCE problem
  - Unpredictable, irregular search tree

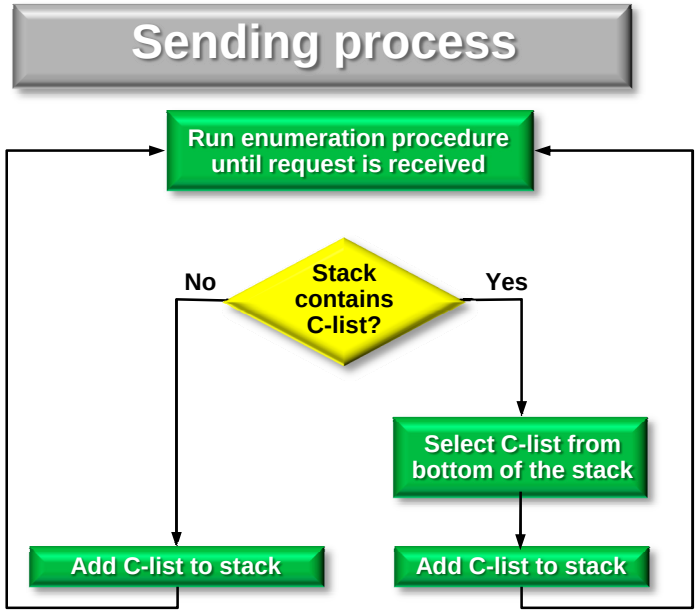


Dynamic load balancing

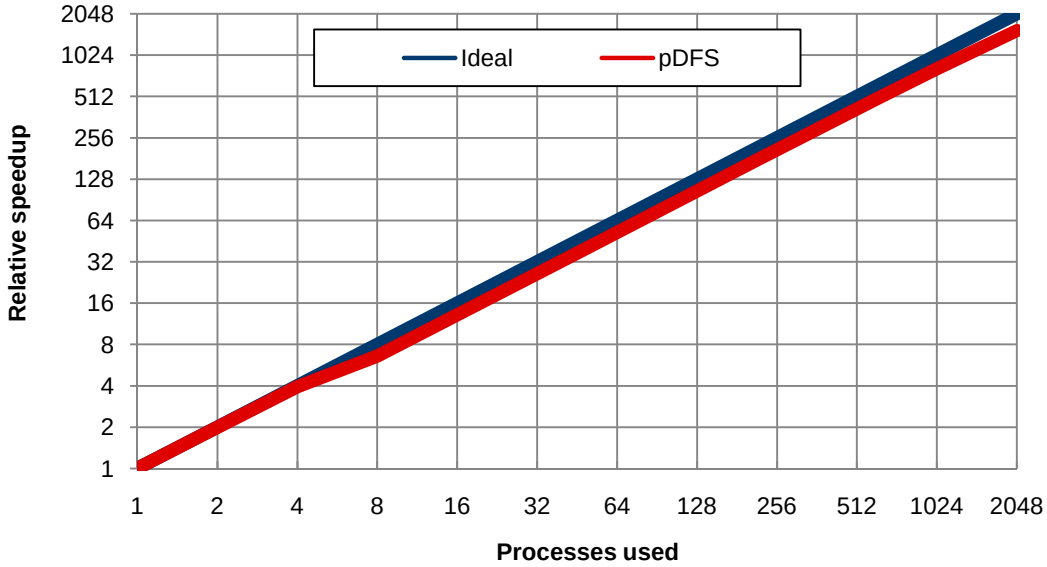
The dynamic load-balancing processes keeps the execution time for different computing nodes from differing greatly

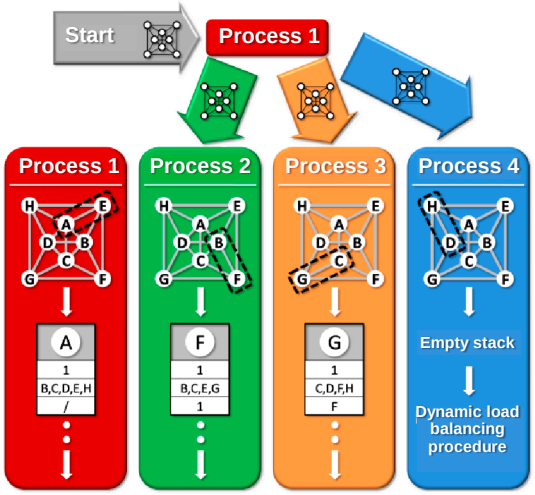




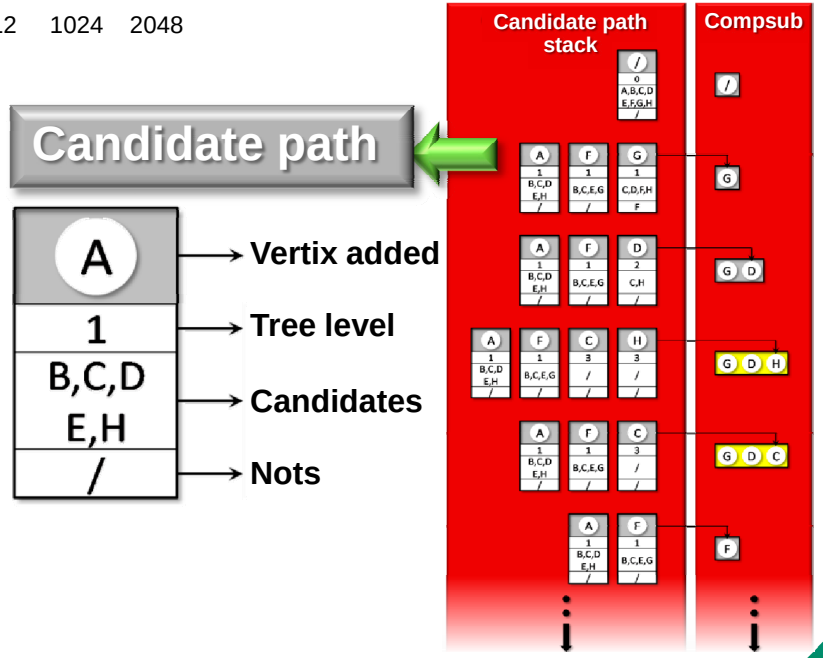


Linear speedup for parallel MCE





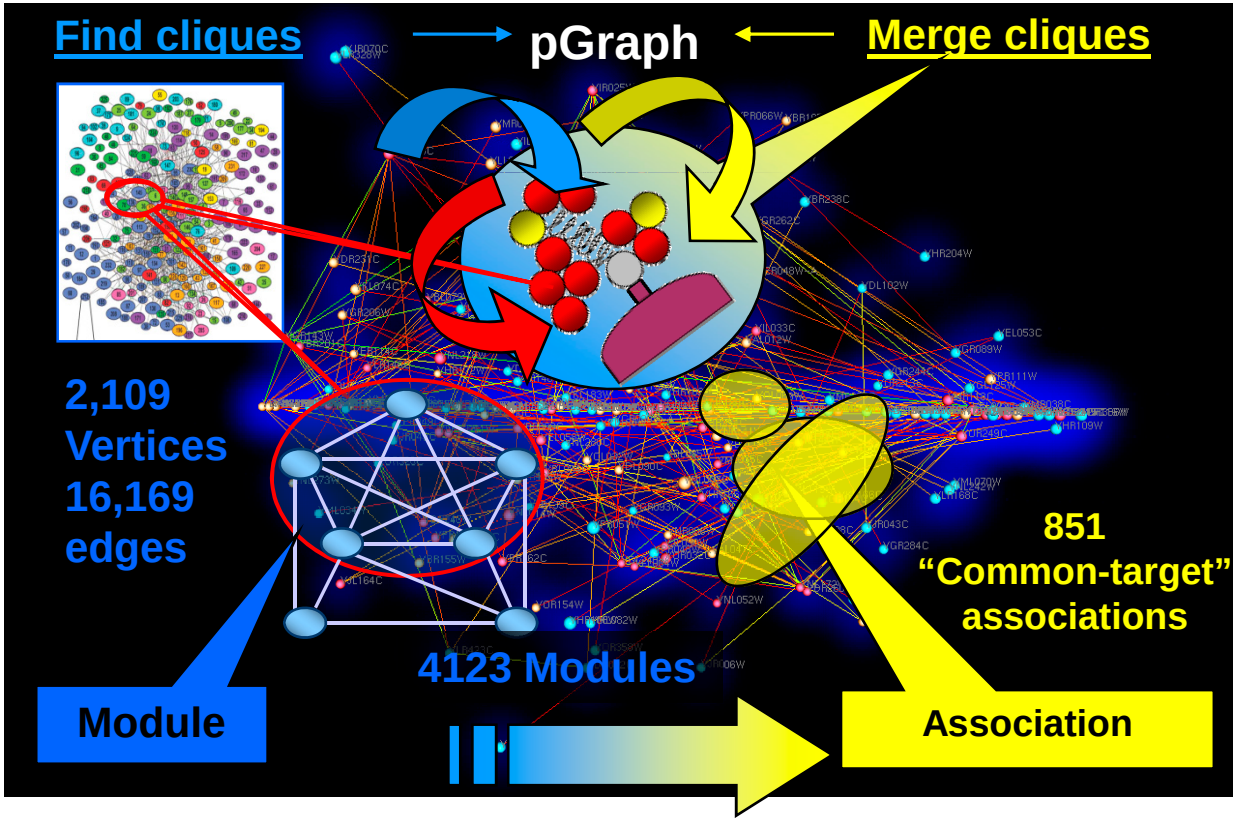
Combining dynamic load balancing with memory-efficient task decomposition strategy provides a linear speedup for thousands of processors

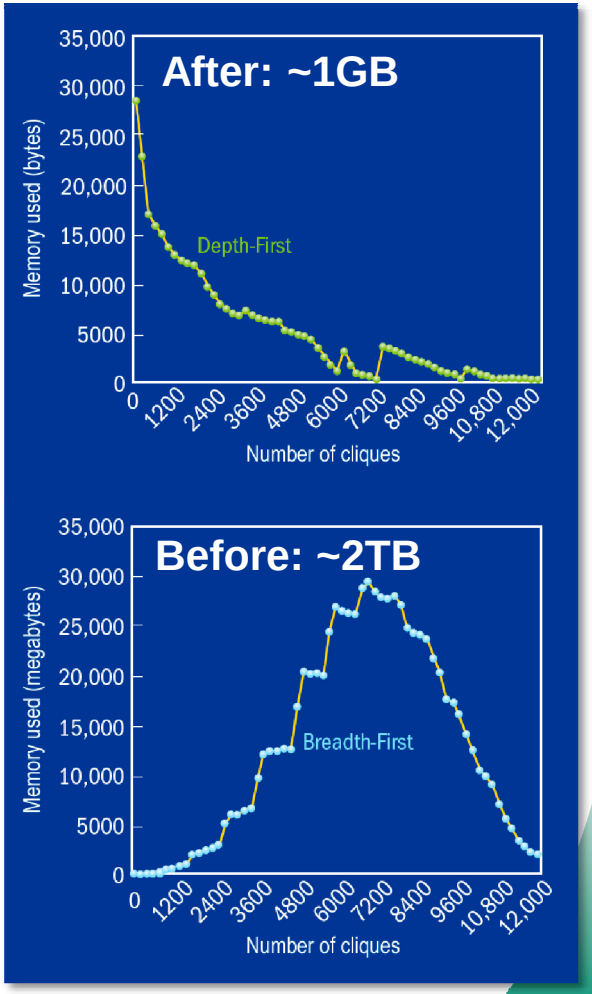


pGraph: Parallel Graph Library for analysis of biological networks

Major features

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

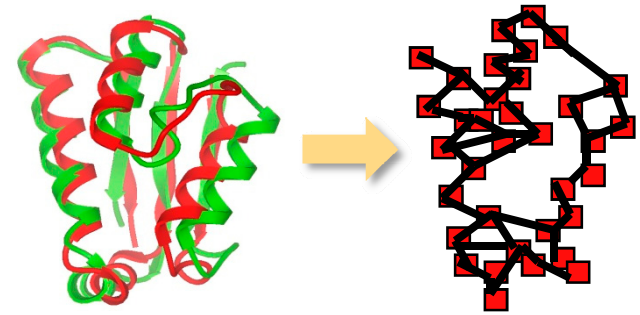


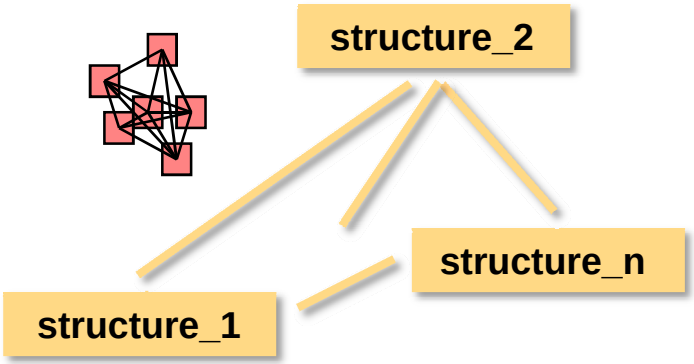


Analysis of ultra-large structural ensembles

Libraries for ultra-large structural ensembles (10^6–10^8 structures) provide unique capabilities for protein-folding and protein-docking applications

- **Ensembles of molecular structures can be represented as graphs—some of the largest naturally occurring graphs in nature; they require hundreds and thousands of processor applications for analysis**
- **Dramatically improved scalability for a number of fundamental graph analysis problems (MCE, VC) provides foundation for some of the most important problems in the analysis of biostructural ensembles**
- **Common Kernel Method determines natively folded protein domains with high precision leading, for the first time, to applicability of protein folding in system biology**





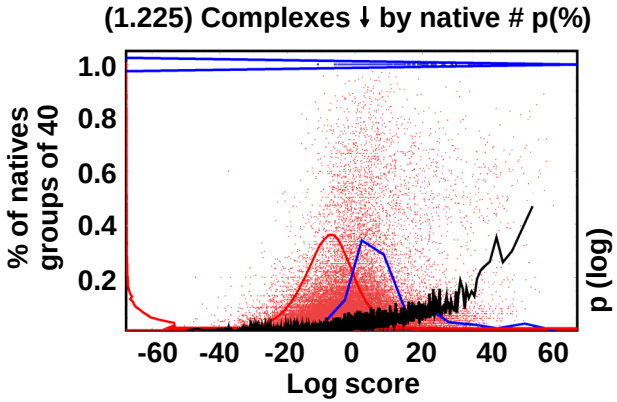
Common Kernel is the number of the common contact matrix elements shared by all members of the group

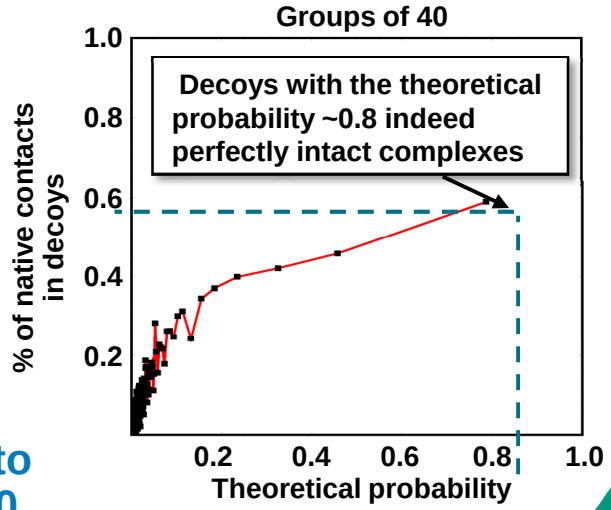
Genome sequence → M-C ensemble of 10^6–10^8 structures → → Biological function

Large-Scale docking of protein complexes

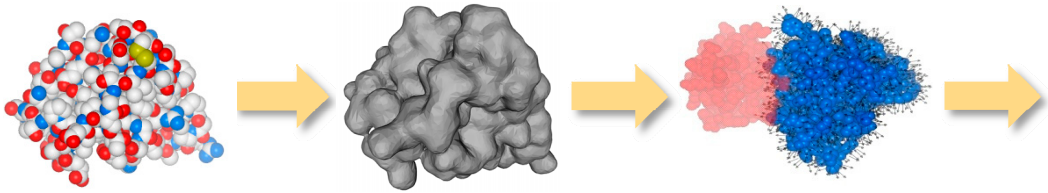
Docking pipeline powered by Bayesian energy tables shows high-quality results on uniquely large set of protein complexes

- **Novel approach for derivation of potential energy through rigorous Bayesian statistics**
- **The size of our docking sets (currently >5000!) is unprecedented in the field. High-quality results were obtained for more than 80% of all complexes, indicating high the efficiency of the method**
- **Successfully modeled protein complexes from novel organisms, including several organisms important for bioenergy program**
- **Docking calculations are scalable to thousands of processors**





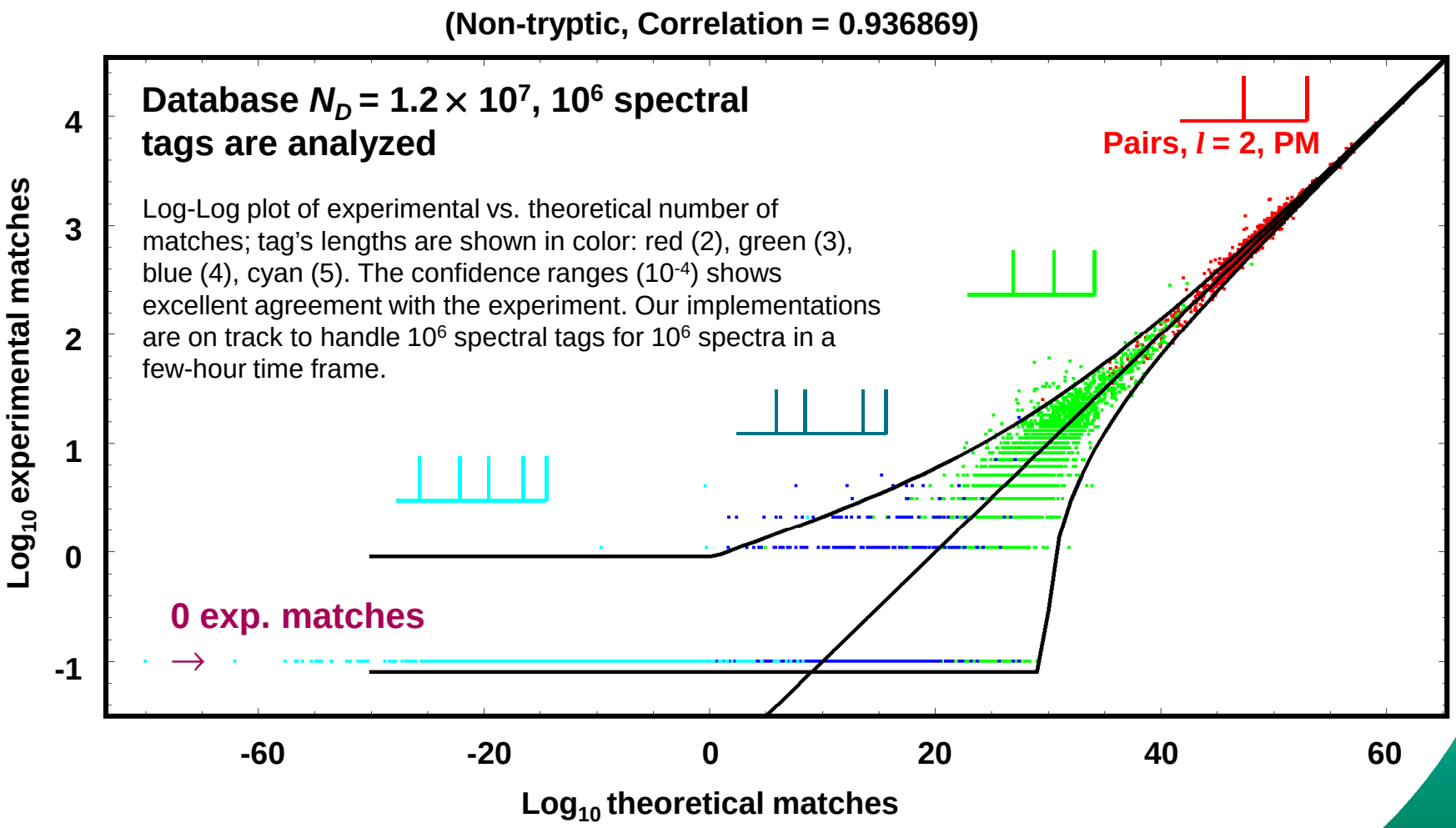
$$p_T(score) = \frac{p_{interface}}{p_{interface} + p_{decoys}}$$



Up to 1000 threads

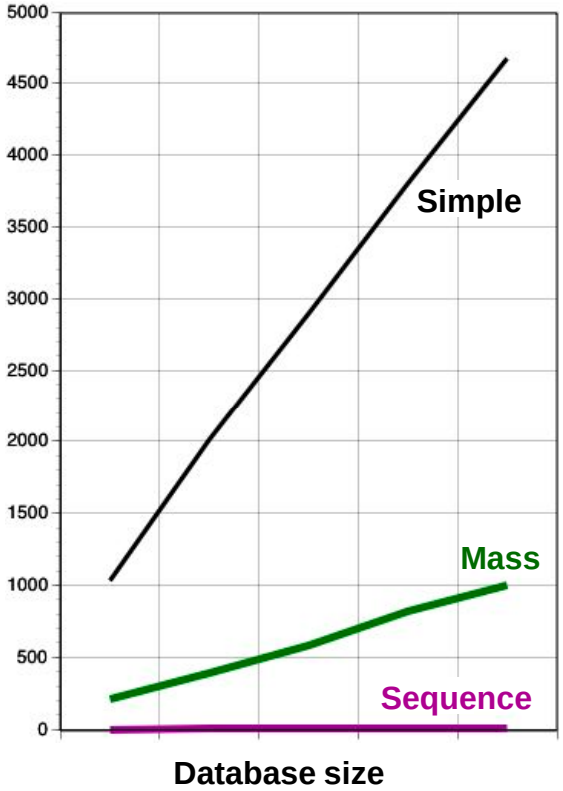
OAK RIDGE National Laboratory

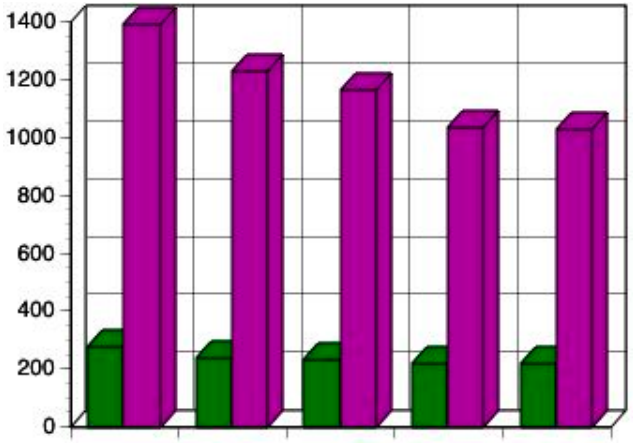
Mass-spectrometry-based proteomics require analysis of millions of spectral tags for many millions of spectra



Spectral tag search performance

Performance of the fundamental search operations has been increased by two to three orders of magnitude

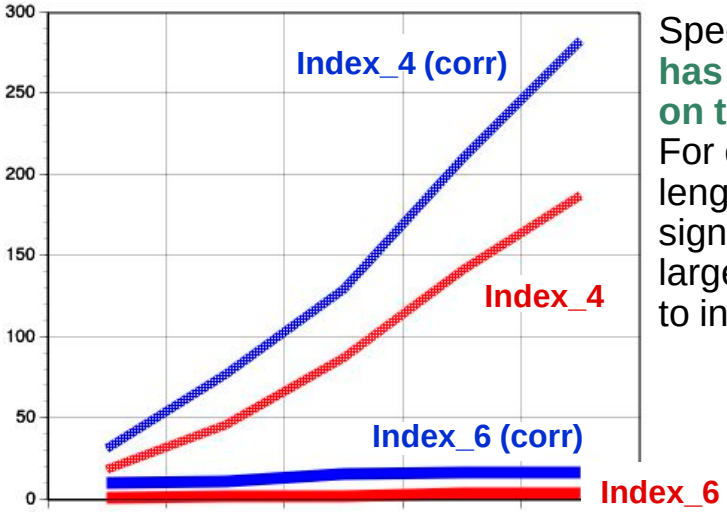




Mass/simple

Tag/simple

Computational time and performance gains to search for 10^6 tags on 5 different database sizes (1 to 5 bacterial genomes)



Specific index **architecture has a significant influence on the performance gains**; For example, indexes for length 4 start to show a significant loss of speed for larger databases compared to indexes for length 6

Contacts

Nagiza Samatova

Principle Investigator
Oak Ridge National Laboratory
samatovan@ornl.gov

Andrey Gorin

Coprincipal Investigator
Oak Ridge National Laboratory
agor@ornl.gov

Edward Uberbacher

Coprincipal Investigator
Oak Ridge National Laboratory
uberbacherec@ornl.gov

Christine Chalk

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