Present and Future Computing Requirements for Computational Prediction of Protein-DNA Binding

Mohammed AlQuraishi

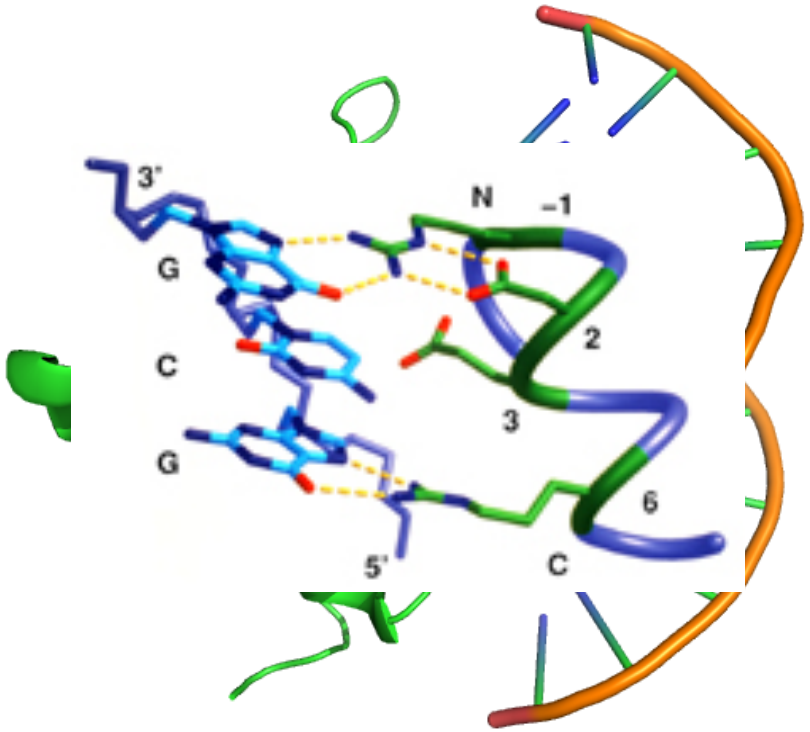
Stanford University, Harvard University

NERSC BER Requirements for 2017
September 11-12, 2012
Rockville, MD

1. Project Description

Harley McAdams, Stanford University

- Computational prediction of biomolecular interactions
  - Given atomic structures of molecules, predict binding affinity

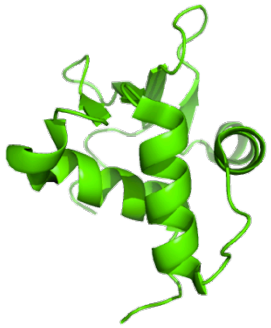


1. Project Description

Harley McAdams, Stanford University

- Computational prediction of biomolecular interactions
  - Given atomic structures of molecules, predict binding affinity

Input	Output
☐ Protein Structure	☐ Position Weight Matrix



1. Project Description

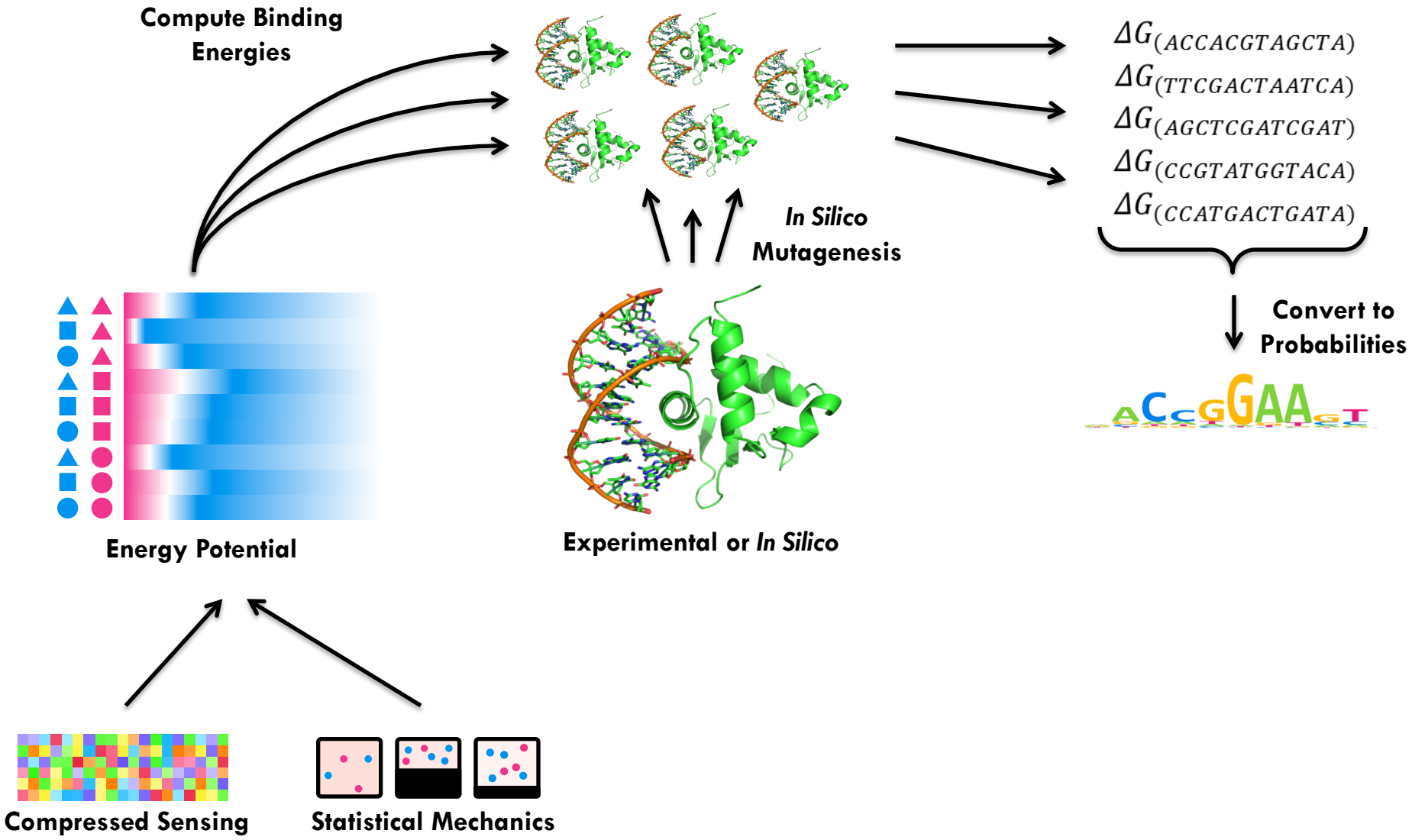
Harley McAdams, Stanford University

- Computational prediction of biomolecular interactions
  - Given atomic structures of molecules, predict binding affinity

- Present focus
  - Protein-DNA binding (bacterial families)
  - Limited binding modalities

- By 2017:
  - Protein-DNA binding (all families)
  - Protein-protein binding
  - Very diverse binding modalities

2. Computational Strategies



2. Computational Strategies

- Three high-level computational challenges
  - Inference of energy potential
  - Computation of binding energy
  - Identification of putative binding sites (future)

- Code is a mix of off-the-shelf and custom-built:
  - Standard solvers for convex optimization
  - Custom-built for Monte Carlo
  - Mathematica, Matlab, R for analysis

- Main algorithms
  - Convex optimization
  - Monte Carlo runs
  - Geometric algorithms

2. Computational Strategies

- Biggest computational challenges
  - Large memory requirements (convex optimization)
  - Long sequential runs (Monte Carlo)

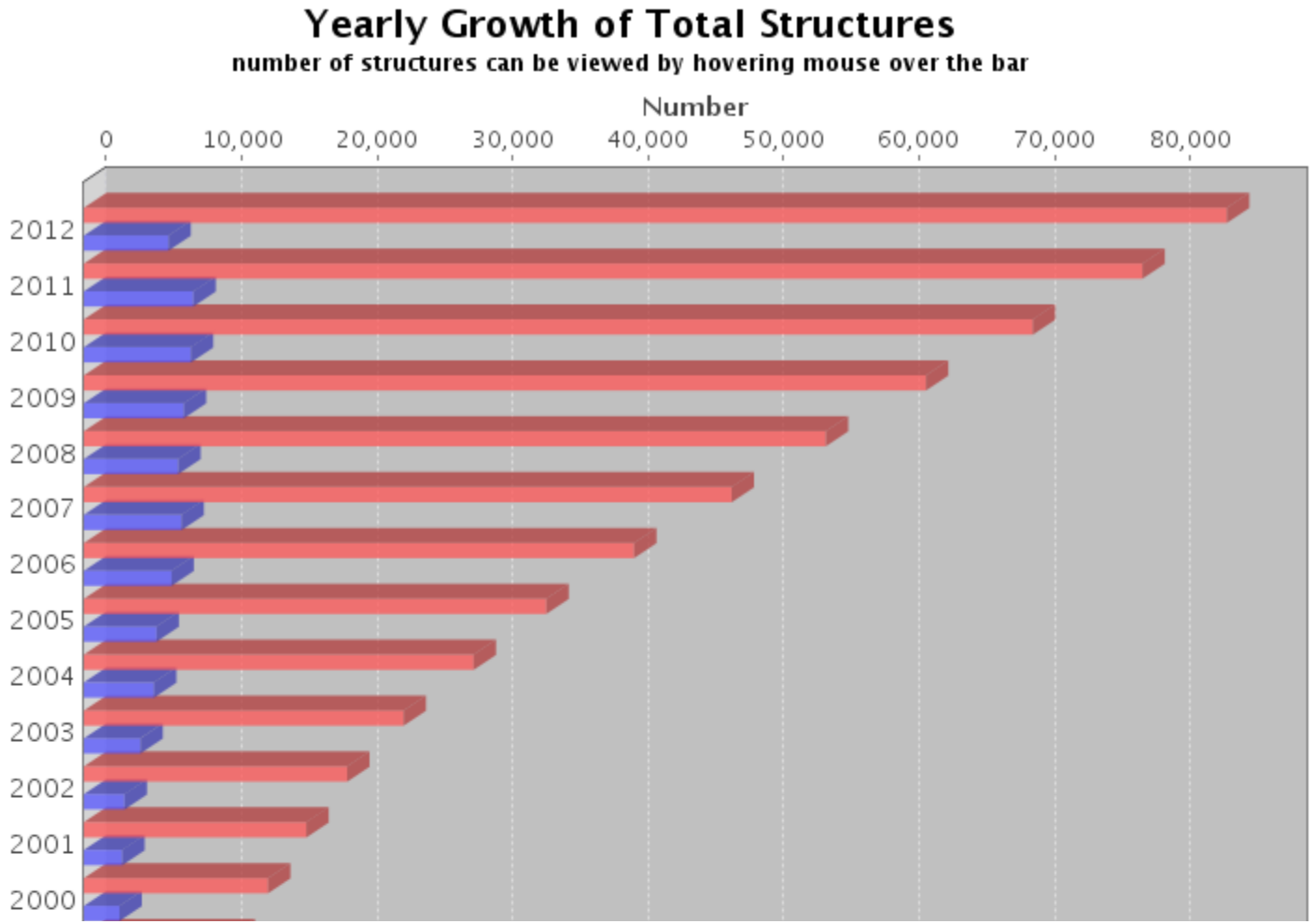
- Parallel scaling limitations
  - Most computations scale (MPI), except:
  - Monte Carlo runs, depends on how well ergodicity is satisfied

- We do not expect computational approach to change significantly by 2017, except for possible use of GPUs.

3. Current HPC Usage

- Hours used in 2012 (list different facilities):
  - Carver ~0.5M Hrs
- Typical parallel concurrency and run time, number of runs per year:
  - 384-512 cores / run, several hundred runs per year
- Data read/written per run:
  - ~50GB
- Memory used per (node | core | globally)
  - ~2GB/core for Monte Carlo, ~200GB/node for convex optimization
- Necessary software, services or infrastructure
  - Mathematica, Matlab, R
  - Built-in checkpointing would be great

4. HPC Requirements for 2017



4. HPC Requirements for 2017

- Compute hours needed (in units of Hopper hours)
  - 5M hrs – 10 M hrs, based on expected increase in available data
- Changes to parallel concurrency, run time, number of runs per year
  - Number of independent runs will increase
- Changes to data read/written
  - Increase by 2-3X
- Changes to memory needed
  - Per node
    - Monte Carlo runs will be largely unchanged
    - Convex optimization may require >1TB memory per node
  - Globally
    - Monte Carlo runs will require ~2-3X memory

5. Strategies for New Architectures

- Our strategy for running on new many-core architectures (GPUs or MIC) is:
  - Standard solvers should get updated automatically
  - For custom-code, still in exploration stage.

5. Summary

- With enough computing power…
  - Aim to solve molecular recognition problem
  - Applications in
    - Understanding / simulating biological pathways
    - Protein engineering
    - Reengineering of metabolic networks

- Wh[illegible]eded
  - More m[illegible] node[illegible]ger Euclid)
  - Bet[illegible]quential runs
  - Easier acces[illegible]g environments like Mather[illegible]ca, R,