LA-SiGMA Louisiana Alliance for Simulation-Guided Materials Applications

Many Big Jobs on Many Supercomputers

Rajib Mukherjee and Thomas C Bishop

Center for Computational Science, Tulane University, New Orleans, LA

Abstract

Molecular simulation is an effective tool to study structure function relationships in bio-molecules. Today's supercomputers contain up to 200,000 processors, too many for a single simulation of a modest size system; however, it is reasonable to simulate 10's to 100's of structures simultaneously on a single supercomputer or to distribute them on whatever computing resources may be available. Such high throughput high performance simulations require careful coordination and strategy. Our present efforts are directed toward investigation of nucleosome positioning and stability with all atom molecular dynamics simulation. Study of nucleosome positioning as a function of DNA sequence requires simulations of hundreds of nucleosomes with different sequences. In order to achieve these simulations with minimum effort and proper utilization of resources, we have utilized two scheduling tools: ManyJobs and BigJobs. Both are pilot job implementations. We are using PetaShare for data management. ManyJobs and BigJobs can be configured to run any computational task.

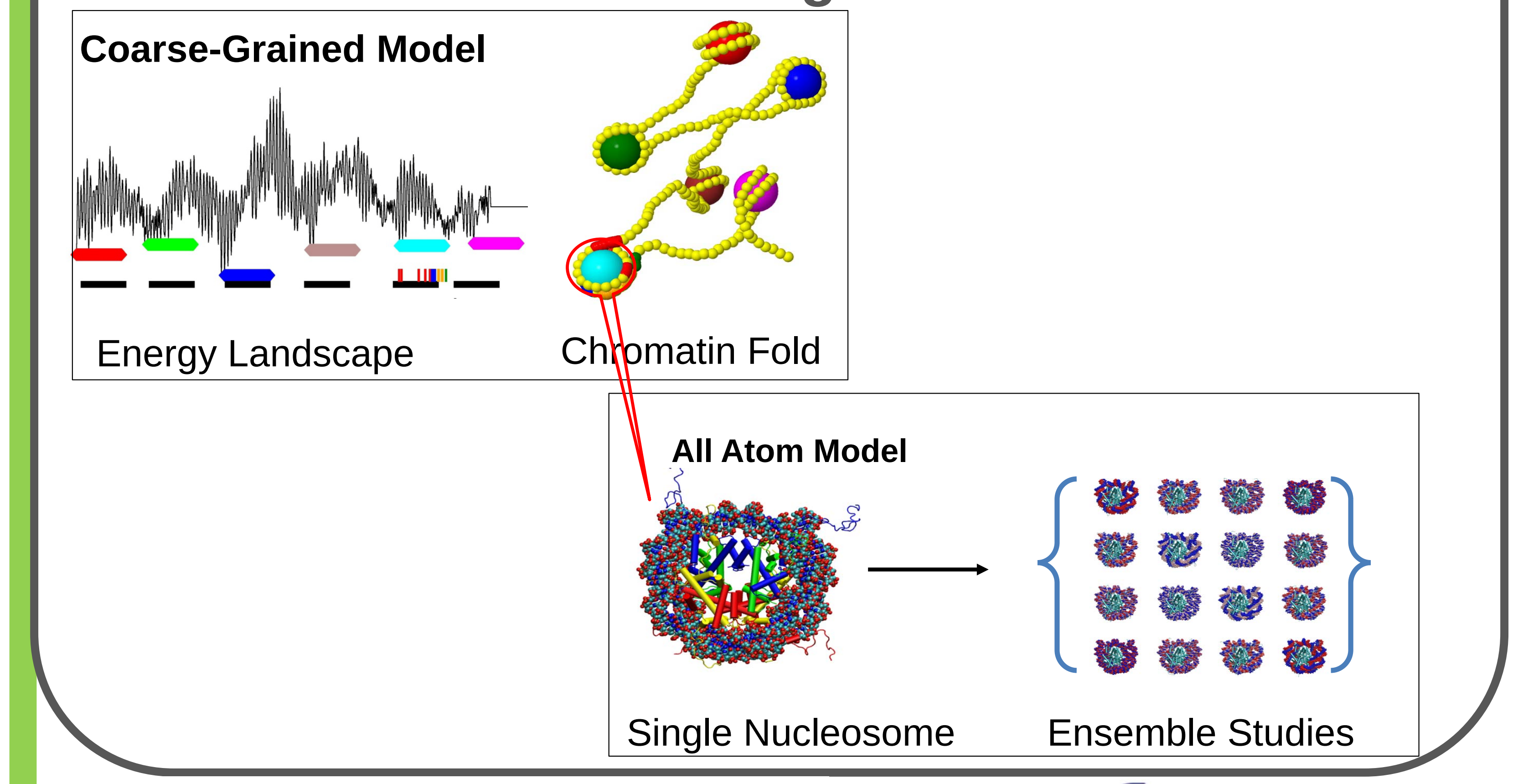
Background

We seek to understand nucleosome positioning as this plays a role in the folding of DNA into chromatin (see figure below). Coarse-grained models are very fast and allow entire genomes to be scanned and assembled into putative chromatin folds in near real time. But they do not yet provide reliable predictions. We have therefore turned to traditional all atom molecular dynamics simulations to understand the energy landscape in greater detail. However, this requires that we conduct simulation ensembles. Each simulation in the ensemble is itself a high performance computing event that requires nearly 8,000 SU and generates ~25Gb of data.

As part of the larger LA-SiGMA effort, we are developing tools that manage 100's of simulation tasks and the associated data on supercomputing resources distributed across LONI and the TeraGrid.

We are developing these tools with NAMD as our compute engine, but tools are not specific to any simulation methodology. Thus they can run NAMD, AMBER, GROMACS, Charm, Gaussian or any other generic computational task for that matter. The purpose of these tools is to manage compute jobs and the associated data.

Bio-molecular Modeling on Different Scales

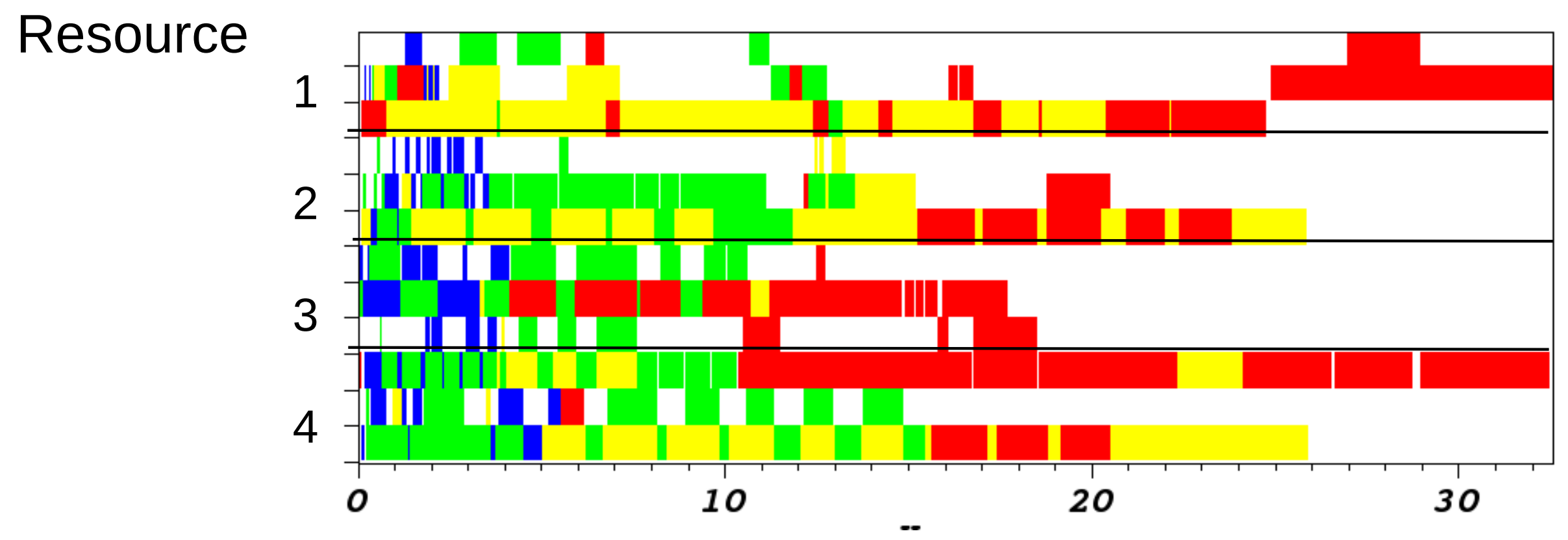


Progress to Date

We are developing two methods for running many big jobs on many supercomputers.

1) A light weight portable collection of python scripts call ***ManyJobs***.
2) A more robust implementation based on SAGA called ***BigJobs***.

Both allow us to rise above the various high level services (schedulers, queues …) and low level services (compilers, network topologies…) associated with today's computing, data and network resources and merely run our application.



As indicated in the above graphic of resource utilization versus time, ManyJobs and BigJobs distributes our simulation tasks to whatever resource is available at a given time. Both approaches manage dependencies between tasks (colored threads in the graphic) and the distribution of concurrent, yet independent, threads to available resources.

ManyJobs and BigJobs effectively port the jobs to where ever the resources are available. This provides the shortest time to completion for us and optimizes use of shared computing infrastructures. Not being committed to one computing resource, allows us to take advantage of otherwise wasted compute cycles on idle machines. A win-win for both us and the supercomputing community.

We are currently using both BigJobs and ManyJobs in production mode on LONI and TeraGrid to make progress on 336 separate simulations being run for 20 ns as 6,720 Tasks. Each Task is a 1ns simulation utilizing from 32 to 192 processor cores. Current objectives are to determine which parameters provide the greatest throughput and shortest time to completion for our project.

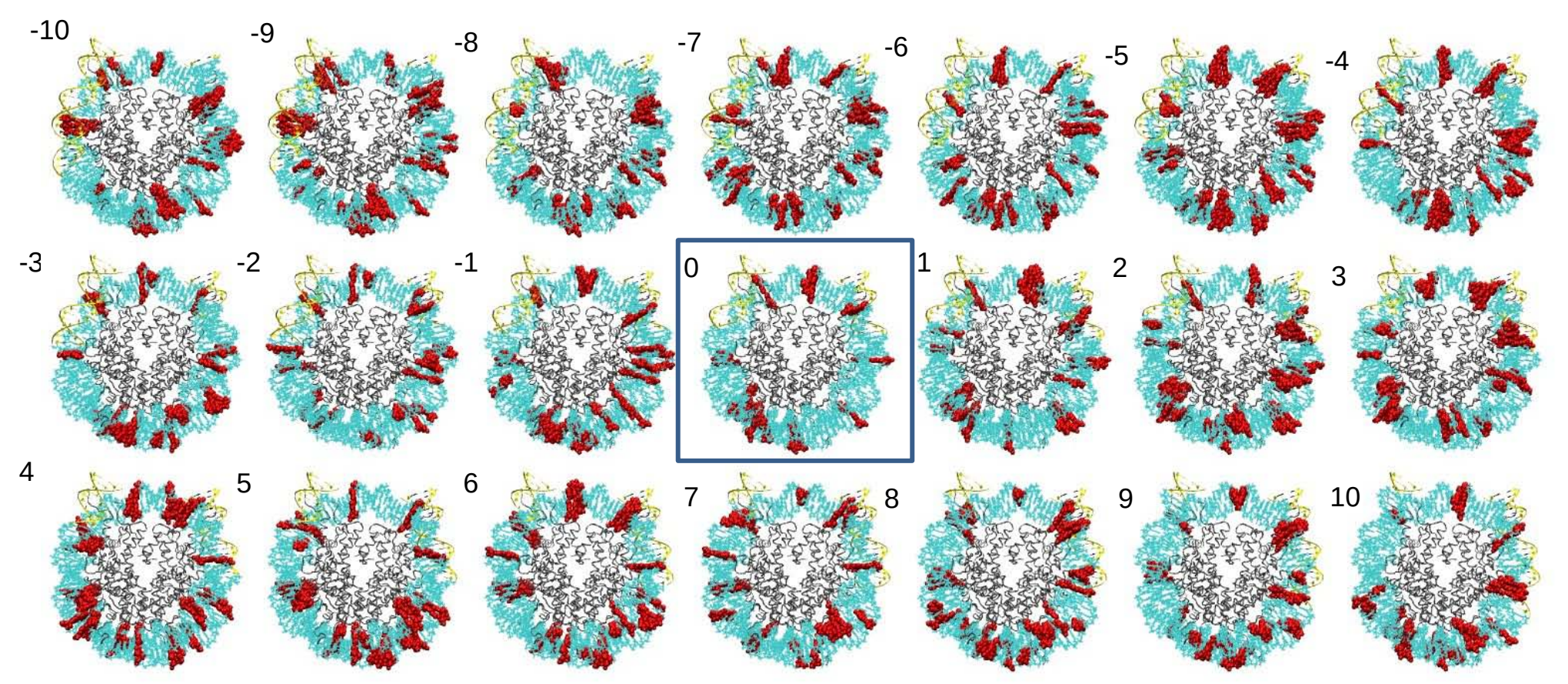
Comparison of ManyJobs and BigJobs

ManyJobs	BigJobs
Completely Python Based	SAGA Based with Python Extensions
SSH or GSISSH authentication	SAGA authentication and simulation management
Validated on:	Validated on:
CCS cluster	
LONI (all machines)	LONI (all machines)
TeraGrid : abe(defunct), ranger, lonestar, kraken, queenbee	TeraGrid: abe(defunct), ranger, lonestar, kraken, queenbee
Pilot Job Based	Pilot Job Based
One Task per Job Assignment with many concurrent jobs	bundles multiple Tasks in single big job
Availability:	Availability:
http:/dna.ccs.tulane.edu/ManyJobs	http://saga.cct.lsu.edu/

Analysis of Nucleosome Data

For the *S. cerevisiae* derived sequences, we model 336 nucleosomes. The collection represents 16 of the most well positioned nucleosomes and their immediate neighbors in sequence space. Each neighborhood spans two turns of the DNA, one upstream turn and one downstream turn. The 20 individual neighbors contain only 147bp and are created by adding one base-pair on one end of the nucleosome and removing one basepair from the other end. We have performed structural analysis to identify DNA kinking. Results indicate the number of kinks present during a simulation correlates with positioning.

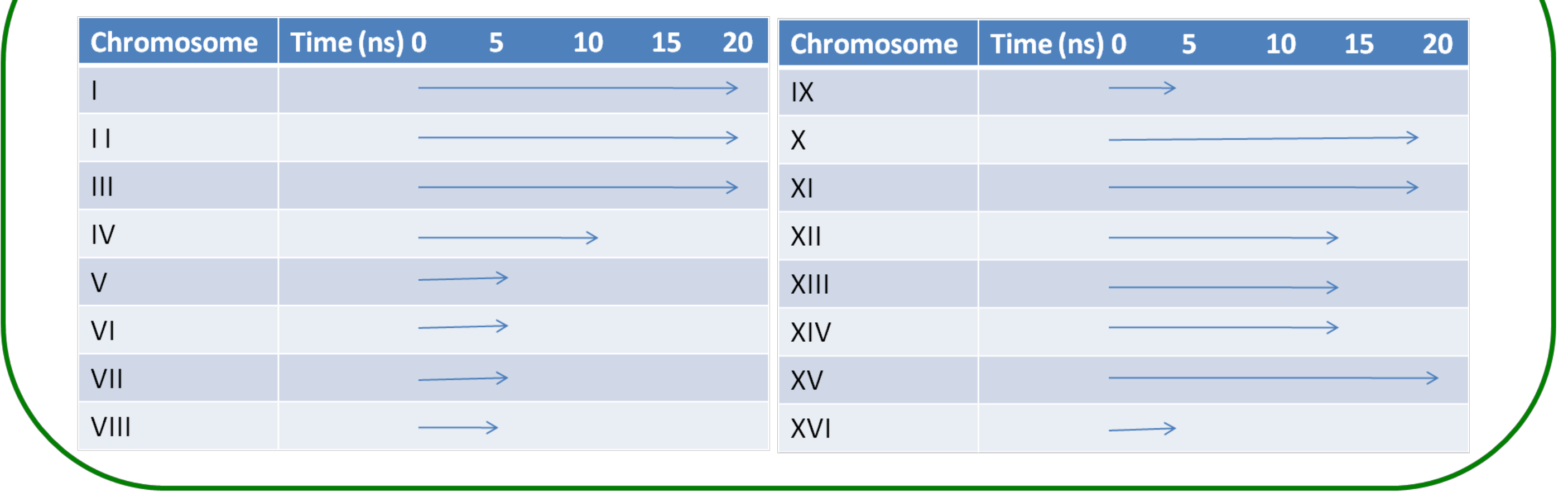
Richmond-Davey Criteria



We have applied three different criteria for defining a kink to 21 molecular dynamics simulations of the nucleosome. Kinks as determined by the Richmond and Davey criteria are shown in red, above. The 21 simulations represent the threading of a single 167nt sequence onto the histone core, thus all nucleosomes have 126 basepairs is common. The primary difference between nucleosomes is the relative position of the 126nt subsequence with respect to the histone core. This subsequence is highlighted in cyan in each image. The boxed nucleosome corresponds to the most well-positioned , well-defined nucleosome of yeast chromosome I as found experimentally.

We find that kinks occur at many positions not just the six locations previously identified experimentally by Richmond and Davey: ±36, ±48, ±58 relative to the dyad. Their definition was biased such that kinks occur where the minor groove faces towards the histone core. Our unbiased definition identifies these kinks locations and a few more.

Project Status: > 3.9μs of nucleosome dynamics



ManyJobs Future Plans

1. Expand user base to include LA-SiGMA projects
2. Improve fault tolerance of ManyJobs and BigJobs
3. Improve user interface to ManyJobs and BigJobs
4. Extend library of sample scripts/examples
5. Convert to XML as basis of task library for ManyJobs

Acknowledgements

Collaborators:
Shantenu Jha and Abhinav Thota, LSU-Center for Computation & Technology
Hideki Fujioka, LONI Computational Scientist Fellow at Tulane CCS