# Massively Parallel Simulations of Large DNA Molecules Using the IBM Blue Gene Supercomputer

Amanda Peters[1], Greg Lakatos[2], Maria Fyta[3], Simone Melchionna[1], & Efthimios Kaxiras[1,2]

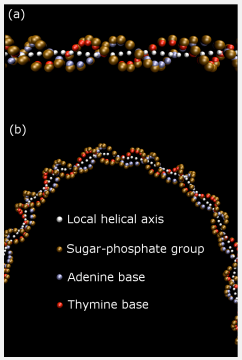
[1]School of Engineering and Applied Sciences, Harvard University. [2]Physics Department, Harvard University, [3]Physics Department, Technical University of Munich, Germany

## Summary

The conformational variability of DNA remains a problem of critical importance, especially in view of recent experimental studies of DNA translocation through nanopores and DNA interaction with nanostructures such as carbon nanotubes. Naturally occurring DNA molecules are typically very large, containing thousands to billions of individual chemical units. Due to this size and complexity, simulating systems involving natural DNA molecules requires the development of an effective simplified molecular model, and a massively parallel molecular dynamics implementation of the model.

Our results show a 97% overall time reduction and 98% efficiency at 512 processors on Harvard's Blue Gene/L supercomputer. Most importantly, this framework enables simulation of large DNA molecules that were previously impracticable. This model and implementation represent a significant new advance in the development multi-scale simulation tools capable of dealing with large DNA molecules. Together these tools will provide insights into the behavior of DNA, both in living cells, and in a new generation of novel DNA sequencing and sorting devices.

## DNA Potential




We have constructed a coarse-grained potential to model the dynamics of DNA at large length scales, using as basic structural units the individual bases and the sugar-phosphate backbone. The interactions between these units were derived from extensive first-principles calculations, and fitted to simple functional forms. The potential was validated by reproducing the correct persistence length as a function of temperature and salt concentration.

The figure above shows some representative structures from MD simulations,indicating that DNA has the correct helical structure, in both (a) long straightsegments, as well as in (b) segments with very large curvature.

## Parallel Techniques

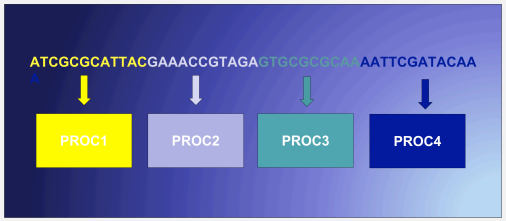
The goals of a parallel implementation were to reduce overall computation time, enable the modelling of much longer DNA sequences, and allow for a greater number of steps to be carried out by the simulation in a tractable amount of time. The code was implemented in parallel using the standard MPI library. The following were the key points of focus in parallelizing the application:

- Contiguous 3D arrays
  - Each dimension individually allocated
  - Addresses set explicitly
- Memory Management
- I/O Reduction
- Static Partitioning

The largest amount of time was spent calculating the force involved in the energetics. In this code there are thousands of calls to the two force functions making any improvement there aggregate to a much larger impact in large scale simulations.

Effort to speed up the code were focused on these two functions using the following methods:

- Distribute ranges of base pairs to different nodes
- Create a temporary 3D array to track force updates
- Gather the data and add the updates to the force values




In this manner, the work was spread across the DNA nucleotides allowing a larger strand of DNA to be simulated. Time steps were inherently serial but minimizing runtime of each step further allowed more time steps to be handled in a reasonable amount of time.

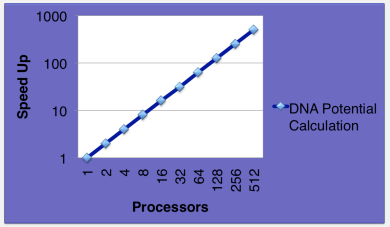
The data was statically partitioned across the processors with each maintaining local force information. Calculation itself relied only on nearest neighbor data from the previous time step allowing communication to be minimized.

Following each time step, all local data would be reduced to all processors therefore initializing the next step.

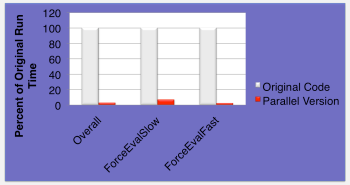
## Results

The evaluation of the parallel version of the DNA potential code was carried on using a medium sized time setting and the DNA sequence consisting of 1280 base pairs. The temperature was set to 300, Debye Length to 10.0, and the time step to .000002.

The system size was grown to fit the DNA length to ensure that all nodes would remain busy while still reducing the overall time.




We saw an overall time reduction of **97%** and a scaling efficiency of **98%** up to 512 nodes.




## Future Work

In future work, the code is now poised to handle much longer DNA sequences and a higher number of time steps in a feasible time frame. It has shown a strong ability to scale well and a likelihood that this will continue to grow as the problem size is increased. As the code is scaled out we may need to investigate:

- Master/Worker and Static Partitioning Hybrid for further parallelization
- Single node optimization
- SIMD utilization

## Reference

Fyta, Maria, Greg Lakatos, Simone Melchionna, and Efthimios Kaxiras. "Coarse grained interactions between DNA nucleotides from first-principles density-functional-theory based calculations." In progress.

**Acknowledgements:** We'd like to thank Seppo Sahrakorpi and the rest of the Crimson Grid support team for continued assistance with the system and providing us with dedicated time and access.