# Molecular Dynamics simulations of Biological Systems

Fatemeh Khalili-Araghi
Department of Physics
University of Illinois at Chicago

Molecular Dynamics (MD) simulation has been used successfully to study biological processes with atomic details. The simulations study the time evolution of atomic coordinates by numerically solving classical equations of motion. Atomic interactions are usually described by finely tuned parameters, or forcefields, that have evolved tremendously over the past decade. The simulations provide realistic representation of the biological systems by including all the environmental elements atom by atom.

Recent advances in MD software and computing power has made it possible to carry out simulations of single protein systems on the local computing clusters. More and more universities are investing in high performance computing clusters (HPC) as shared resources across the institution, making this technology accessible to a larger pool of researchers at a much lower cost. Researchers can invest as little as a few nodes in these clusters and have access to the entire system for a fraction of time.

However, biological systems usually consist of a complex of several proteins. Simulation of these real life machineries consisting of millions of atoms is not yet possible on the local HPCs. These simulations are only possible on leadership computational facilities such as Blue Waters or Mira that provide thousands of nodes at the same time. As the size of the simulation system increases, so does the natural time scale of biological events occurring in these systems, requiring further computational cycles to simulate events of interest. In most cases, these simulations are well beyond the capabilities of the local HPCs or even medium size computing resources available through XSEDE.

While leadership facilities funded by NSF or DOE provide a valuable resource to carry out such large simulations, there is an inherent problem in accessing longer time scales that cannot be addressed by simply distributing the simulations over an increased number of nodes. MD simulations integrate classical equations of motion with a finite time step. One would assume that increasing the integration time step would make longer time accessible in the simulations. However, to capture the fastest vibrational frequencies of the system, in this case vibration of the covalent bonds between hydrogen atoms and the rest of the system, the simulation time step has to be ~ 1fs. This limit imposed by the forcefield parameters describing the vibrational frequencies poses a hard limit on the actual "speed" of the simulation. Considering that real life events in biological processes occur within several hundreds of microseconds to milliseconds, each simulation trajectory requires $10^{10}$-$10^{12}$ integration cycles, which are two to three orders of magnitude longer than what can currently be achieved.

Currently, the only solution for the above problem is to use advanced physical algorithms that allow extraction of physical and biological properties of the system from hundreds or thousands of short trajectories. These simulations can be distributed over thousands of node to run simultaneously. The algorithms allow simulation of processes that occur over microseconds or millisecond timescales in a fraction of time by increasing the probability of rare events occurrence. Physical properties of the event such as the transition pathways, transition times, or even the free energy surface can then be extracted by combining individual trajectories. The most famous algorithm is the replica exchange method that has been developed more than a decade ago. However, until recently, its use has been limited to toy systems or very small proteins. The main hindrance for widespread use of the algorithm is the fact that the number of replicas required for proper sampling of the system increases with the number of atoms or degrees of freedom of the system. So, successful application of this method to a single molecule system requires a few hundreds copies of the simulation to run simultaneously as the algorithm needs frequent communication between the replicas. These simulations are now possible only on leadership computational facilities. However, practical considerations have prevented a concerted effort in implementing these methods into current MD software as well as its implementation on various platforms.

Recently, there has been a collective effort to implement previously proposed methods and algorithms, mostly developed by mathematicians or theoretical physicists or chemists, into available MD packages and make it possible to run on leadership computational facilities. However, the process is still in its early stages. It takes several years from initiation of such projects to make the code available, and it requires continuous collaboration and feedback among researchers at universities and staff scientists at supercomputer centers. Several leadership facilities have already been very successful in providing the knowledge and support to make this possible.

The current computational resources, such as Blue Waters or Mira, have opened up a new door for studying biological events and processes that were not possible before. These investigations require further development and implementation of new computational methodologies that have already been described theoretically, but have never found practical grounds in previous generations of supercomputers.