Hi Technical gaps: improved models (implicit solvent, coarse-grained force fields, hybrid classical/quantum models)

Nathan Baker (baker@biochem.wustl.edu)
Washington University in St. Louis

I. Adaptive and transferable coarse-graining methods

• Current computational applications challenged by length and time scales
• Current coarse-grained models are not accurate enough for some questions
• Need adaptive (truly multiscale) methods
• Existing multiscaling methods are very limited and very problem-specific

I.A. What new theory or algorithms need to be developed to achieve it?

• Goal driven adaptive-refinement methods
• Goal/objective definitions (structural, thermodynamic, kinetic)
• Multiscale-savvy sampling methods

I.B. What computational capabilities will be required to achieve it?

• Parallel numerical methods
• Existing computational hardware infrastructure

I.C. What is the scientific impact of achieving it?

• Ability to examine site-specific behavior in large systems:
  o Site binding to membranes
  o Ligand binding sites in large proteins or nucleic acid assemblies
  o Defect behavior in bio-inspired materials
• Ability to perform more detailed titration state or QM calculations in large systems

II. Robust hybrid models for a variety of environments

• Can be cast as a more specialized version of Challenge I.
• Current computational applications challenged by length and time scales
• Continuum is not accurate enough for some questions
• Need hybrid methods
• However, many hybrid methods are specific to exposed aqueous environments with little conformational change

II.A. What new theory or algorithms need to be developed to achieve it?

• Algorithms/theories to adapt explicit region with conformational change
• Algorithms to describe heterogeneous dielectric surroundings
• Metrics and validation for assessing performance, explicit region size requirements, etc.

II.B. What computational capabilities will be required to achieve it?

• Faster partial differential equation solvers
• Existing computational hardware/infrastructure
II.C. **What is the scientific impact of achieving it?**

- Ability to examine site-specific behavior in large systems:
  - Site binding to membranes
  - Ligand binding sites in large proteins or nucleic acid assemblies
  - Defect behavior in bio-inspired materials
- Ability to perform more detailed titration state or QM calculations in large systems

III. **On-the-fly titration state sampling**

- Biomolecular titration states are not static – they change with ligand binding, folding, conformational changes, and with thermal fluctuations
- However, most molecular simulations assume static titration states
- Those simulations which include variable titration states are extremely slow to converge

III.A. **What new theory or algorithms need to be developed to achieve it?**

- The basic theory is in place; however, the implementations are unsatisfactory
- Need enhanced sampling methods appropriate to this application
- Need reliable explicit solvent force fields for titration state changes

III.B. **What computational capabilities will be required to achieve it?**

- Simulations of this type will likely be more computationally demanding than traditional molecular dynamics simulations and will therefore require more advanced computational hardware infrastructure than is currently routinely available.

III.C. **What is the scientific impact of achieving it?**

- Ability to more accurately describe:
  - Ligand binding
  - Biomolecular folding
  - Biomolecular conformational changes
- Ability to predict:
  - Titration states for understanding enzymatic mechanism
  - pH profiles for engineering protein stability

IV. **Common frameworks for development and validation**

- Progress in this field is hampered by
  - Several siloed software projects with differing implementations and abilities
  - The lack of clear validation sets which can be applied across software packages, force fields, etc. to assess accuracy and precision of simulations
The proliferation of inexpensive commodity computing hardware has changed the landscape of biomolecular simulation. Molecular dynamics (MD) and other simulation techniques were once the exclusive domain of supercomputers. While the supercomputing sites remain vital for studies of large complexes or high-throughput screening and simulation of many molecules (e.g. Dynameomics\textsuperscript{1}), many labs have small clusters and even desktops capable of running single molecule simulations, small-scale drug and complex docking searches, and other molecular mechanics calculations. In addition, the ‘free cloud’, e.g. BOINC\textsuperscript{2}, has made it possible for some groups to run very large numbers of simulations, dockings, etc.

This shift in computing has resulted in scientists generating significantly more simulation data than 10 years ago. As an example, the Dynameomics project has over 100 terabytes of coordinate vs. simulation time data (i.e. trajectories) from native state and un/folding simulations of 1000s of proteins \textsuperscript{1, 3-5}. In addition, cloud computing projects like Folding@Home\textsuperscript{6}, Rosetta@Home\textsuperscript{7}, and Docking@Home\textsuperscript{8}, generate large amounts of synthetic data. These types of very large-scale theory based projects pose significant challenges for analysis, data management, mining, etc. Funding agencies, supercomputing providers, and cloud participants have investing significant computing power in these projects and it may be that the resulting datasets are not being fully exploited due to limitations in analytical methods.

Another important shift in biomolecular modeling is that the simulations are no longer the only significant computation. Previously, simulations were conducted on a cluster or supercomputing resource then downloaded to a scientist’s local facility and studied in isolation with fairly ubiquitous algorithms, e.g. solvent accessible surface area. The new paradigm can be the reverse: a large collection of trajectories is uploaded to a supercomputing site for analysis using novel algorithms and strategies to handle the scale of 1000s of proteins\textsuperscript{3} or 10,000s\textsuperscript{9} to 100,000s of trajectories. This change necessitates computing time awards for analysis in addition to or in place of simulation.

The significant challenges for large scale, high-throughput modeling projects can be put on a scale of purely scientific to purely technical. The scientific challenges are concerned with the theory of the techniques (i.e. methods development), understanding of individual trajectories (i.e. chemistry, biochemistry, etc.), and the behavior of biomolecules in general (i.e. biophysics). The middle ground, commonly referred to as eScience, is the intersection of the science and technology. The eScience challenges tend to be centered on the implementation of algorithms for mining, informatics, data storage and retrieval, and other practical concerns related to the tractability of computations. These concerns are often shared across disciplines, and in the case of biomolecular simulation, similar to those from astrophysics and cosmology simulation. Purely technical challenges, on the other hand, embody aspects of hardware design for simulation and data intensive computing.
Some of the purely **scientific challenges** include contextualizing single molecule simulations with ensemble experiments, evolving protocols, force fields and algorithms to improve agreement with experiment, and pulling together results from many diverse trajectories to better understand underlying phenomena, e.g. folding, docking, binding. In the first example, it is routine to perform replicate simulations that are 10s to 100s of nanoseconds long for a single molecule or complex. How do microscopic properties derived from these simulations relate to ensemble behavior in vitro / in vivo? Are deviations from observations related to errors or imperfections in the modeling and simulation? As we evolve protocols, force fields, and algorithms, are we making improvements in experimental agreement? What prior science, conclusions, and experimental comparisons are invalidated or need to be re-validated when we make improvements in methods? What do we mine from a large corpus of 1000s of simulations of 1000s of systems to extract general determinants of (e.g.) protein dynamics and folding?

The **eScience challenges** extend the science challenges, often through the lens of pragmatism. That is, what computational approaches are tractable for algorithms, statistical / ensemble analyses, and informatics (e.g. clustering) when faced with very large (e.g. petabyte) datasets of many different simulations. Traditional serial analysis of individual trajectories is no longer viable for large repositories of data and we need to consider parallel analysis similar in style and scale of those used to perform simulations. In addition, newer computing paradigms like Map-Reduce may hold promise for unlocking large databases. Is a relational database (RDBMS) required for storing the collections of trajectories, docking results, etc? Can the use of RDBMS facilitate random access to the data necessary for some analytical techniques? How can we track provenance of the trajectories (e.g. what simulation software, what compiler, what computer, what user, what version of a force-field parameter library)?

Finally, purely **technical challenges** are sometimes the most real. What storage infrastructure is appropriate for the data and how can it be designed to deliver the petabyte dataset to computational nodes quickly and reliably? How can we design computers specifically for analyzing the data? Analytical high performance computing (HPC) presents different needs than traditional simulation driven HPC. For data-intensive analytics, we may not want to optimize for cores per node (as simulation might), instead choosing to maximize memory per core or network bandwidth per core.

As many of these challenges are cross-cutting sub-topics of macromolecular simulation, so too are the rewards. As an example, improving the fidelity of data mining techniques for large biomolecular simulation databases can improve the success rate of pharmacophore screening, help identify common dynamic modes in Dynameomics, decrease false-positives and false-negatives in structure prediction, etc. The computing challenges are no longer solved by just throwing bigger and faster hardware at the simulation problem, and instead involve components of analysis.


8. Taufer, M. et al. in First Workshop on Large-Scale, Volatile Desktop Grids (PCGrid'07), in conjunction with IPDPS'07 (Long Beach, California, USA, 2007).


What are the missing pieces in the areas of mathematical models, algorithms, software required to solve the problem?

1. Automated annotation of large-scale MD simulations:

   Improvements in computer speeds and simulation algorithms are greatly increasing the accessible size and time scale of atomistic MD simulations. It is now realistic to envision in the near future multi-microsecond simulations of biomolecular systems large enough to contain a membrane-embedded protein or multiprotein machines such as a DNA replication complex. By comparison, the methods used to analyze MD trajectories have been improved much less dramatically. Most analysis of biomolecular MD simulations involves a combination of calculating structural and dynamical properties averaged over the trajectory, combined with viewing movies of the results while monitoring structural properties (such as specific interatomic distances). Such analysis works well when the research question can be easily related to a specific structural change and the simulation is short enough that the trajectory is relatively homogeneous in time and space. In contrast, for multi-microsecond simulations of complex biomolecular systems, many changes in the structure, dynamics and state of different components can occur. Such changes could be due to errors or inaccuracies in the simulations (e.g. decoupling of solvent and solute temperatures or anomalous precipitation of counter ions). Alternatively, such changes might reveal true, but unanticipated properties of biomolecular systems (e.g. residual structure in unfolded proteins or changes in solvent properties near solutes). With current approaches, identifying such changes requires extensive manual effort, and may not be noticed at all.

   The ideal tool for analyzing long-timescale MD trajectories of complex biomolecular systems would automatically partition the system into its different chemical components (solvent, counter ions, biomolecules) and then annotate the trajectory in several ways. First, inherently dynamical properties of the components (e.g. solvent viscosity, counter ion residence times, and magnitudes of biomolecular motion) would be calculated and then mapped onto the trajectory, so that individual frames could be annotated with the time-dependent behavior of the system. Additionally, this analysis system would identify spatial and temporal regions where the structural and dynamical features of each component do not match reference values stored in a pre-existing knowledge base of features.

2. Algorithms for analyzing disordered proteins:

   Traditionally the focus of protein biophysics has been on the structure and interactions of the natively folded protein. There is now growing interest in the properties of denatured protein structures as related to the mechanisms of protein folding and in the function of so-called intrinsically disordered proteins. Molecular dynamics has an important role to play in describing the range of structures that comprise such unfolded protein states. However, new theories and software tools are needed to fully characterize the properties of the unfolded or disordered proteins. Most current molecular dynamics studies of such systems describe the ensemble of conformations sampled in terms of the probability distributions of standard descriptors of folded proteins, such as radius of gyration, or fractional secondary structure
composition. Some new measures of protein dynamics have also been developed, such as the structural decorrelation time and principle component analysis; however, there are currently no tools available to determine the two most fundamental descriptors of protein dynamics, the intrinsic *dimensionality* of the motion and the volume of structural space sampled by the protein. The effective dimensionality is a robust measure of the complexity of the protein motion, ranging from simple hinge opening (low dimensional) to conformational collapse from an extended to a compact structure (high dimensional). Once the dimensionality of the motion has been calculated, it is possible to robustly estimate the volume of conformational space adopted by an unfolded or intrinsically disordered protein during a molecular dynamics simulation and the rate at which this space is sampled. For example, the widely-held "free energy funnel" model proposes that protein folding involves the protein moving in an increasingly restricted conformational space until it converges on a single folded structure. Once completed, these tools will be broadly useful molecular dynamics practitioners in two ways. First, they will provide robust new metrics for concisely describing the dynamics of unfolded or natively unstructured proteins. Second, these tools will provide diagnostics for designing molecular dynamics studies by measuring how efficiently different simulation lengths and numbers of simulation replicates sample conformational space.
Algorithmic and Methodological Challenges in Performing Long Time Scale Simulations

Robert S. Germain

August 11, 2009

1 Probing Long Timescales

In the absence of any broadly accepted algorithmic method for accelerating the simulated dynamics of a biomolecular system (including the possible coarse graining of the simulation), the only most straightforward option is to speed up the brute force molecular dynamics calculation through some means including improved (serial) algorithms and/or implementations[21, 1, 19, 18], specialized hardware[20, 15], or strongly scaled parallelism[11, 6].

It is now broadly accepted that increased machine performance over the next few years will be achieved through increased parallelism (larger numbers of processing elements such as hardware thread contexts, CPUs, and nodes). Devising parallel decompositions and algorithms that can take advantage of larger numbers of processing elements (improve strong scaling) for a single trajectory is an important challenge for the simulation community. In this area, there have been a number of proposed algorithms and actual implementations that can distribute the computation of finite-ranged pairwise forces over very large numbers of nodes[11, 17, 14, 7, 6], but the true challenge lies in dealing with the contribution of the long range electrostatic forces that imposes a global data dependency on the positions of every charge in the system. Since most codes use some form of particle mesh Ewald technique[1, 2] to evaluate the long range electrostatic forces, this global data dependency appears via a convolution between the meshed charge distribution and an appropriately chosen kernel (evaluated using a 3D-FFT). Implementations of a 3D-FFT have been demonstrated that continue to speed up out to the limits of concurrency for a row-column evaluation scheme where each processing element evaluates a single 1D-FFT during the three compute phases that are separated by global transpose operations[3]. However, unless full bisectional bandwidth is maintained as the machine size increases, the transposes required by the 3D-FFT will become the rate limiting factor (even if latency effects can be neglected). Given the reality that the computation rate is likely to increase more quickly and less expensively than communication bandwidth on future generations of machines, it is worth pursuing the investigation of algorithmic alternatives for the evaluation of long range electrostatic forces such as fast multipole[9], multigrid[13], and even equivalent pair interactions using the Ewald technique or Lekner sums[10] used in a force decomposition[12] scheme.
2 Pitfalls in Benchmarking Molecular Dynamics or What Makes a Valid Simulation?

One of the most difficult issues in evaluating algorithmic alternatives is that different problems may call for differing approaches, e.g. an integration scheme or simulation parameters that may be suitable for sampling may not be right for probing the long time dynamics of a system. This also influences how various groups choose to benchmark their codes and performance comparisons are often made between codes that use different choices of parameters on nominally the same molecular system.

In particular, the level of energy conservation observed in a given simulation can be extremely sensitive to the same parameters that affect performance such as choice of time step size and other integration parameters (particularly for multiple time-stepping schemes). Especially as very long (microsecond or longer) simulations become more common, the biomolecular simulation community needs to achieve a better understanding about what makes a given simulation valid[5]. There has been some research done on how to monitor simulations[16] and some characterization of long term energy drift as a function of time step size for Verlet integration[4, 8]. See Figure 1 for a demonstration of the sensitivity of long term energy conservation to time step size.

References


(a) Drift in total energy as a function of time for various time step sizes.

(b) Slope of drift as a function of time step size.

Figure 1: These figures illustrate the strong dependence of long term energy drift on choice of time step size. These simulations were run on a rhodopsin molecular system embedded in a lipid bilayer (about 44,000 atoms total) in an NVE ensemble with all heavy atom to hydrogen bonds constrained and with tight tolerances on RATTLE.


Opportunities in Extreme Computing for Biology Workshop

Zan Luthey-Schulten, University of Illinois at Urbana-Champign

- What specific problem could be attacked and solved with the application of sustained multiple petaflops of computing power? What progress could be obtained on the problem at roughly the 10, 100, and 1000 petaflops levels of sustained performance?

One area of biology that is poised for rapid growth is the modeling of dynamics in cell-scale systems. Systems biology methods couple genomic and other “omics” (e.g. transcriptomics, proteomics, metabolomics, and interactomics) data to map cellular networks and predict their functional states, but they lack dynamic resolution which is important if one wishes to be able to predict and control the time-varying responses of living cells. To analyze the dynamics of complex living systems, analytic model are not presently sufficient and one must turn to computational techniques. All-atom molecular dynamics simulations address biology from the bottom up, and researchers have made impressive methodological advances that allow modeling of the largest assemblies in the cell (e.g. ribosome, membranes and membrane proteins, chromatophores in photosynthetic bacteria, small viruses, ...) for short times, but they are unlikely to reach the length scale of an entire cell (even a small bacteria) for the relevant times (minutes to hours) even at a computational performance of 1000 petaflops. To address these length- and time-scales, one must turn to coarse-grained models of the chemical kinetics within the cell.

For a point of reference, my group is developing a methodology for simulating reaction-diffusion kinetics inside a full-scale model of an E. coli cell, packed with an approximate in vivo environment derived from proteomics and imaging data. The goal is to be able to analyze the stochastic effects of spatial heterogeneity and molecular crowding inside of cells. We are using graphic processing units (GPUs) to perform these simulations because of their currently high performance/price ratio, but the general methodology is compatible with any computational architecture. At 1 teraflop (1GPU=1 teraflop), we are able to calculate up to a few coupled biochemical pathways for on the order of 30 seconds for each 24 hours of computing time at a resolution of 8 nanometers. This model includes approximately 20,000 ribosomes together with approximately 1 million other protein:protein and protein:nucleic acid complexes.

Taking the above performance as a baseline (and naively assuming linear speedup) one can estimate the simulation potential of these sorts of reaction-diffusion models at various performance levels:

10 petaflops: a small biochemical network within a full-scale model of a bacterial cell at 4 nm resolution for 60 minutes per day; alternatively one could consider increasing the size of the network accounted for dynamically (such as by including the TCA cycle) at the expense of simulation time.

100 petaflops: a small biochemical network within a full-scale model of a eukaryotic
cell at 4 nm resolution for 60 minutes per day; a large biochemical network within a full-scale bacterial cell for 60 minutes per day.

1000 petaflops: a large biochemical network within a full-scale model of a eukaryotic cell at 4 nm resolution for 60 minutes per day; a large biochemical network within a full-scale bacterial cell for 10 hours per day.

• Is the problem one of the “top 10” problems for the scientific discipline, independent of computing? Who would constitute the community of scientists and/or engineers that would enthusiastically address the problem? What would be the degree of international potential participation?

I believe that uncovering the fundamental properties of how cells dynamically respond to environmental fluctuations (changing resources, community signaling, morphogenic gradients, etc) is one of the top, if not the #1, problem in computational biology. Virtually all computational biologist would be enthusiastic about developing methodologies for the integration of the massive data and computational techniques that would be necessary to realistically simulate an entire cell. As the GPU cell simulations incorporate spatial distributions obtained from available cryo-electron tomography reconstructions, single molecule experiments, and proteomics data, such simulations complement the experimental efforts by synthesizing a more “unified” picture. In addition integration of results from the stochastic dynamics simulations as constraints into systems biology networks would enable the modeling community to begin constructing integrated whole-cell-scale models.

• How is the use of petascale computational modeling and simulation irreplaceable in answering this question? Does it augment existing techniques or replace them? Is there history of large-scale computation being the preferred approach for this problem?

As discussed above, it seems unlikely that an analytic theory will be developed in the near or mid-term future that will be able to predict the cellular response to various environmental fluctuations. The underlying biochemical networks are likely too complex for the emergent phenomena on which life is based to be predicted without explicit simulation. That being said, fully simulating these biochemical networks (including their spatial heterogeneity within a cell) is a very compute intensive application and could serve as a bridge between all-atom simulations of macromolecular complexes and systems level models. In particular, information gained from all-atom simulations could be used to build up mesoscale kinetic models.

• Why are the other techniques (e.g., experiments/observation, more traditional theory) that could answer these questions not satisfactory? Is it even feasible to consider other techniques?

No other techniques can address the problem directly. Indeed it is more appropriate to think of whole-cell simulations as a complimentary to other computational biology
methods.

• **What is the current status of the computing tools for the work being proposed:** mathematical models, algorithms, software, and data analysis tools? What is the largest scale to date that codes have been run? (e.g. 1,000, 10,000, 100,000 cores) Are there existing code teams working on codes for this problem area, or is this a new area that would need seed investments?

Current models have focused on either accuracy (single core) or GPU (many core) based systems. Little has yet been done on massive parallelization of reaction-diffusion algorithms. As such, much work will be required to efficient utilize petascale computing systems and beyond.

One should also not underestimate the time required to develop the appropriate software that can exploit petascale computing.

• **What experimental and observational data is there available to validate the codes? Is the validation method well established?**

Many experimental observations (particularly single molecule experiments) have been performed recently which provide the information necessary to begin to construct whole-cell models. Much more experimental data will be necessary to improve and verify the models. Fortunately, single molecule techniques are becoming more commonplace and physical biology investigators are interested in the same problems as whole-cell simulations would address.

• **What are the missing pieces in the areas of mathematical models, algorithms, software required to solve the problem? How would you rank them in terms of importance, cost, and risk?**

Algorithms: algorithms for efficiently decomposing three-dimensions problems in such a way as to be amenable to petascale computing area lacking.

Software: the most important missing software are the tools necessary to build and evaluate petascale scientific codes. Until the software development barrier to these systems is brought within reach of more scientists, scientific progress to which petascale systems contribute will be limited.
Proteins are involved in most of life's processes. Understanding protein interactions has a vast array of applications in chemistry, biology, medicine and bioenergy. Proteins are the link between genetics and biological function. However, understanding how proteins achieve their function is challenging. Like most chemical reactions, protein interactions occur when the physical and chemical forces between the participating molecules are energetically favorable. Proteins have specific arrangements of chemical and physical properties that allow them to recognize only the molecules that they should interact with, but we do not yet entirely understand how this occurs.

A century-old metaphor for recognition is a “lock and key,” as each protein must fit together appropriately with its binding partners. Just as the shape of a key creates an appropriate set of forces to open a particular lock, the ``shape'' of a protein (the arrangement of its various physical and chemical properties) must be appropriate to act on its partners.

Protein function is often too difficult to predict from first principles. Instead, the primary tool used to study uncharacterized proteins is pairwise and multiple sequence alignment via tools such as BLAST/Psi-BLAST and MUSCLE. The sequence of a newly discovered protein is compared with a database of known proteins in order to determine close matches that can elucidate the protein’s function based on the function of its closest relatives. However, while sequence homology generally does imply a similar function, the reverse is not true. That is, two proteins can have a similar function even when their sequences are dissimilar.

Going further, one can place the backbone structure of a protein into a family of related folds using databases such as SCOP and Dali. This type of alignment can identify more distant relatives of a protein, but since proteins with different functions can have a common structural scaffold, this type of alignment is not always suited to the exploration of protein function.

As structural genomics projects rapidly progress in their goals, we are beginning to see many proteins (or putative proteins) for which a structure is known in addition to the sequence, but for which no function has been assigned. Additionally, those working in protein docking are well aware that even when a protein’s function is known, along with that of a binding partner, identification of the protein functional surface remains nontrivial in many cases.

The wealth of protein structure data and the capability of today’s massively parallel computer systems have finally led us to the point where we can consider doing for protein functional surfaces what has been done for protein sequence and backbone fold. That is, computational biology is poised to introduce a working model for functional surface homology. Unlike sequence or fold data, which can be represented in a one-dimensional form, functional surfaces are two-dimensional and rely on physical forces that reside in three dimensions. In addition, their topography can be complex and cannot be easily represented using a simple angular representation in the way that backbone fold can be.
With many tens of thousands of solved crystal structures in the Protein Data Bank, including an increasing number of cocrystal structures with explicit detail of the protein functional surfaces, there is sufficient data with which to begin a working model of functional surface homology. There do, however, remain challenges both algorithmically and scientifically.

**Goal**

A key advance in the study of protein interactions will come through the development of a working model of protein surface homology. Such a model would be able to take the structure of a protein and identify regions that are a match to functional surfaces of well-characterized proteins. Such a tool has the potential both to identify binding sites as well as suggest their function.

**Key Challenges**

1. How do we represent the surface geometry of a protein in a way that allows it to be efficiently matched against a large database of functional surfaces?

2. What surface characteristics are most crucial to the identification of binding sites and/or function? (i.e., electrostatics, desolvation, hydrogen bond donors and acceptors, hot spot motifs?)

3. What is the best way to collapse the 3D features above onto a 2D functional surface, or is full matching of features in three dimensions the best approach?

4. How do we deal with the multiscale aspects of the problem? At what level of resolution are we able to match some functionally relevant details. At what level can we begin to distinguish the fine-scale aspects of function? Key examples in this respect are DNA-binding proteins, which bind many sequences non-specifically and some with very high affinity.

5. How do we address the question of conformational change? This is an issue that continues to plague studies of protein interactions, and whether and how to deform the query protein surface in order perform a functional surface alignment is a question must be addressed in the course of studying this problem.

The large amount and complexity of data makes this a problem well-suited to leadership class computing systems. Ideally, the solution can become ultimately tractable in the way that BLAST and Dali searches are presently, but even beginning to explore the question of functional surface homology will require access to vast computational power.
Suggested Template/Questions for Participant Contributions to the "Opportunities in Extreme Computing for Biology Workshop"

What specific problem could be attacked and solved with the application of sustained multiple petaflops of computing power? What progress could be obtained on the problem at roughly the 10, 100, and 1000 petaflops levels of sustained performance?

To give some perspective, our MD simulations of the the Kv1.2 channel system (360,000 atoms) on BG/P generates about 14ns/day with 1 rack. Scaling up to BG/P (40 racks) is already achieved in the context of replica-based strategies. This means that a 557 TF (~0.5 PT) machine is required to run the 40 replicas of Kv1.2. But to map free energy landscapes of large and complex conformational transitions, we often need more like 500 to 1000 replicas, corresponding to 7 to 14 TF. If the molecular system is bigger, then, there is about linear scaling in the need to simulation the basic unit. For example, molecular machines such as the P-type ion pumps or the ribosome are 5-10 times larger (up to 3.6 million atoms), and the need to simulate large conformational transitions in such system would require easily 35 to 70 TF.

Is the problem is one of the “top 10" problems for the scientific discipline, independent of computing? Who would constitute the community of scientists and/or engineers that would enthusiastically address the problem? What would be the degree of international potential participation?

In biology, the biggest impact of MD is still largely at the level of a strong validation of the results, but some room for new discovery. The latter are believed only if there is sufficient prior validation. Ass an example, the rate between all the steps of the duty-cycle of a P-type ion pump are known experimentally, and producing a quantitative movie of all those steps with quantitative agreement with experiment would have a great impact. Doing all this would require usage of 35 to 75 TF for several months.

How is the use of petascale computational modeling and simulation irreplaceable in answering this question? Does it augment existing techniques or replace them? Is there history of large-scale computation being the preferred approach for this problem?

No experimental measurement can produce the information that the computation would produce.

Why are the other techniques (e.g., experiments/observation, more traditional theory) that could answer these questions not satisfactory? Is it even feasible to consider other techniques?

In the best scenario, atomistic computations and experiments in biology are complementary. In some case, the computations can provide numbers (e.g., binding free energies) that can be measured. Another example is the effect of point mutations on the stability and function of a protein. Most importantly, the computations can provide detailed structural information that are not accessible by experiments. Even if binding free energy can be both calculated and measured, only the computations offer the possibility to dissect the atomic interactions and reveal the origin of binding (or the cause for the lack of binding). Another key example concerns the transitions between the stable states separated by a large free energy barrier. Because the time spent at the “top of the energy barrier” is much much shorter than the time spent at the bottom of the energy wells, it is nearly impossible to capture such transient states experimentally. Only the computations offer the possibility to elucidate such details by using virtual (in silico) models.
What is the current status of the computing tools for the work being proposed: mathematical models, algorithms, software, and data analysis tools? What is the largest scale to date that codes have been run? (e.g., 1,000, 10,000, 100,000 cores) Are there existing code teams working on codes for this problem area, or is this a new area that would need seed investments?

The best MD codes for large platforms are NAMD, Gromacs, and Desmond. The other codes such as Amber and CHARMM have other functionalities but are not designed for high performance on these platforms. Scalability depends on the size of the molecular systems, e.g., NAMD performs well up to about 50 atoms per core.

What experimental and observational data is there available to validate the codes? Is the validation method well established?

There is plenty of quantitative experimental data to validate the atomic models in the case of systems that are either simple or of moderate complexity. Examples are the solvation free energy of small molecules (few thousands of molecules), conformational propensity of short poly-peptides (< 20 residues). All this information has not yet been used to validate the current models, and potentially improve those models. In the case of larger biological systems, quantitative rates and relative population are less well known. Sometimes, even the mechanism are unknown, and computations can help to converge toward the elucidation of the correct mechanism by proposing a smaller number of possibilities that need to be tested.

What are the missing pieces in the areas of mathematical models, algorithms, software required to solve the problem? How would you rank them in terms of importance, cost, and risk?

The effect of atomic motion is best treated by the correct classical physics, F=MA. No heuristic knowledge-based method is going to replace this to address issues at the atomic level. The key issues are (1) development and validation of accurate atomic models, (2) development and validation of computational strategies and framework going beyond simple brute force MD.

Classical simulations based on potential functions for modeling molecular systems have a long history. Nonetheless, systematic efforts to develop and optimize such potential functions, also called “force fields”, have been relatively scarce over the years. The approximation of the Born-Oppenheimer surface by classical functions can indeed be valid, as long as no bond-breaking processes are taking place. On the other hand, it is easy to see that bad or inaccurate force fields can have a very significant impact on the utility and accuracy of simulations. In some cases, the inadequacies of force fields have been identified only after a significant amount of computational resources invested in simulations had been wasted. An additional problem is that of lack of versatility. For example, in material science projects, specific molecular moieties needed to address the issues are often missing from the currently available force fields. Standard force fields become rapidly insufficient in the context of material science, where one may be interested in the impact of novel chemical groups. This either leads to sloppy models (scrambling to put together some moieties quickly), or to unwanted delays as researchers attempt to parameterize the functionalities that they need prior to tackle the real problems of interest. Mostly, each project are considered independently from one another, leading to much uncoordinated efforts as researchers attempt to test and validate new force fields. For all these reasons, there has been a general skepticism and lack of confidence about the general usefulness of atomistic
simulations in material design. This problem is significantly holding back atomistic simulations and preventing them to achieve their real promises.

Although it is complex, the problem is amenable to a solution. First and foremost, one must have a solid understanding of the basic molecular physics that is involved (this is outlined below). Then, one must cast this problem in an algorithmic perspective. Once a functional form has been adopted to correctly incorporate the physics of the microscopic interactions, the issue of determining the parameters of the force field is then primarily an optimization problem requiring a good search strategy. However, this is a costly optimization problem, which requires a distributed computing strategy. An important advantage to solve the problem of determining force fields using an automated procedure will be to associate the physical expertise to the protocol, thus yielding a better-defined method (rather than in the parameters themselves). As progress will be made, the protocol will be refined, allowing a continuous improvement of methodology that will be readily available to all. The system will insure a unified view of molecular forces such that future simulations on a wide range of physical and chemical systems are accurate.

But in final analysis, one must realize that in order to make validation of the atomic models and quantitative predictions possible, it is unavoidable that one must resort to some mathematical framework that goes beyond simple brute force application of F=MA. To tackle the computations of free energies, transition rates, statistical populations of states, it is necessary and useful to seek mathematical formulations that break down the computations into smaller parts that are only weakly coupled. Such frameworks already exist but they have not yet been implemented on the most powerful machines.
Protein Complexes

Kevin Sanbonmatsu, Los Alamos Nat. Lab.

"Opportunities in Extreme Computing for Biology Workshop"

- What specific problem could be attacked and solved with the application of sustained multiple petaflops of computing power? A good grand challenge would be simulating a large biomolecular complex such as the ribosome on physiologically meaningful timescales. This complex is central to every living system since it manufactures every protein in the cell and is the object of much of cellular metabolism. The ribosome is one of the largest asymmetric structures solved to date. With 0.1 petaflops, we can simulate ~ 1 microsecond. Therefore with 10 PF, we could simulate ~ 100 microseconds. With 100 PF, we could simulate ~ 1 ms and with 1000 PF we may simulate ~ 10 ms. This time scale is close to experimental time scales (1ms – 1 s).

- Is the problem one of the “top 10” problems for the scientific discipline, independent of computing? Yes, simulations of the ribosome for 10 ms would allow the simulation of spontaneous conformations, allowing us to map the energy landscape. The community of sciences that would address this problem would be molecular biologists, biochemists, structural biologists, computer scientists and computational physicists. Many groups in the US, Europe and Asia are interested in this problem.

- How is the use of petascale computational modeling and simulation irreplaceable in answering this question? Understanding this problem in atomic detail requires computer simulation. Current experimental techniques do not have either (1) sufficient time resolution or (2) sufficient spatial resolution. A combination of high time resolution and high spatial resolution is required to understand the mechanism of conformational change. Current compute power is insufficient: 1 microsecond simulations are at least 1000 fold from experimental time scales. Does it augment existing techniques or replace them? Initially this will augment existing techniques. Eventually, if force fields are accurate, the simulations could play a primary role with experiment verifying predictions made by simulations. Is there history of large-scale computation being the preferred approach for this problem? Large-scale computation is just now catching on in the community although a few groups have been doing large scale studies for some time.

- Why are the other techniques (e.g., experiments/observation, more traditional theory) that could answer these questions not satisfactory? Is it even feasible to
consider other techniques? It is not feasible to consider other techniques. X-ray crystallography and cryo-EM give static snap shots. Single molecule FRET have insufficient spatial resolution.

• What is the current status of the computing tools for the work being proposed: mathematical models, algorithms, software, and data analysis tools? What is the largest scale to date that codes have been run? (e.g. 1,000, 10,000, 100,000 cores) Are there existing code teams working on codes for this problem area, or is this a new area that would need seed investments? The longest published simulation of the ribosome run to date has been ~20 ns.

• What experimental and observational data is there available to validate the codes? Is the validation method well established? The are x-ray crystallography structures of the ribosome in one conformation at 3 A resolution. There are cryo-EM reconstructions of the ribosome in several functional states at 8 A resolution. There single molecule FRET measurement of the ribosome undergoing conformational changes. These measurements show the distance between two components of the ribosome as a function of time. There are many studies of mutations and modifications of the ribosome and their effect on activity.

• What are the missing pieces in the areas of mathematical models, algorithms, software required to solve the problem? How would you rank them in terms of importance, cost, and risk? The missing pieces are (1) force field accuracy and (2) simulation time. (1) is currently being addressed by various force field groups – with increases in compute power, force fields are being re-tooled for accuracy in longer simulations. (2) will be addressed by 10, 100 and 1000 PF simulations.
Unraveling Regulatory Biomolecular Networks by Novel Mesoscale Models and Dynamical Approaches

Tamar Schlick

Current Research Efforts

My group’s interdisciplinary research focuses on the development and application of novel modeling and dynamical computational techniques to regulatory and long-time processes in DNA, RNA, and proteins, specifically protein/nucleic acid interactions involved in DNA synthesis and repair, chromatin folding, and RNA design.

The structural details of biomolecules are of great importance because of the crucial biological link between structure and function. Computer modeling and simulations help link the structural information on proteins and nucleic acids obtained by X-ray crystallography and nuclear magnetic resonance (NMR) with the wide range of dynamic behavior in the cell. Currently, macromolecular modeling serves not only as an important link between sequence and function but also as a vehicle for directing structural and functional initiatives, predicting biological phenomena, and pursuing medical and technological applications of the underlying biological systems.

Because of inherent practical limitations that biomolecular models face in dealing with complex, chaotic, hierarchical, and multiscale systems, new modeling and algorithmic approaches that borrow on diverse fields of mathematics (like topology and geometry), as well as computing and numerical analysis (e.g., for coarse-graining and enhanced sampling) [1, 2], are continuously needed to enhance the reliability of macromolecular simulations to address targeted biological problems (see Fig. 1).

Our interdisciplinary group is applying efficient dynamics techniques and novel models to systems of biological interest to bridge microscopic structures with macroscopic functional observations. Specifically, we focus on DNA/protein interactions in fundamental regulatory processes that occur over a range of spatial and temporal scales; such interactions control gene expression, genome packaging, replication, repair, transcription, and recombination, etc.

Problems on the atomic level include the synthesis and repair mechanisms of DNA polymerases (e.g., delineation of the conformational and chemical pathways in polymerase catalytic cycles, interpretation of varying efficiency and fidelity behavior across various polymerases). See Figs. 2–6 for studies of key aspects of the conformational and chemical pathways for DNA polymerases including pol β [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16] and its cousins from the X-family pol λ [17, 18, 19] and pol X [20, 21], as well as the Y-family polymerase member Dpo4 [22], along with pol β and Dpo4 with the common oxidative lesion 8-oxoG [22, 23, 24].

On the macroscopic and mesoscopic scales, we investigate chromatin structure and function (e.g., folding/unfolding driving forces, roles of the various histone tails and linker histones in stabilizing chromatin and regulating transcription) by an integration of molecular-level components with polymer-level representations. See Figure 7 for the mesoscale model developed and applied to study chromatin folding and organization in a series of macroscopic and mesoscopic models for representing polynucleosomes with increasing resolution [25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37].

Ultimately, such mesoscale models of various components (like chromatin and signaling proteins) can be combined to study biological networks of transcription, DNA repair, and gene regulation.

Our initiative in RNA structure and function using a graph-theory approach for representing RNA secondary structures [38, 39] (Fig. 8) is focusing on RNA structure analysis and design. Works involve classifying/analyzing topological characteristics of existing RNA [40, 41] (Fig. 9); detecting structural and functional similarity among existing RNAs [42]; identifying RNA motifs of antibiotic-binding aptamers (found synthetically) in genomes [43, 44, 45]; designing; and predicting novel RNA motifs [41].
(Fig. 10); enhancing the *in-vitro* selection process of RNA discovery by systematic modeling of the process [46, 47, 48, 49] (Fig. 11); annotating tertiary motifs in RNA [50, 51, 52] (Fig. 12) and predicting tertiary motifs of RNAs from the primary sequence. Our resulting RNA topology resource RAG (RNA-As-Graphs) (http://monod.biophysics.nyu.edu/rna) has been used by the RNA community.

**Current Challenges in Biomolecular Structure and Function**

1. Understand DNA polymerase replication and replication processes on the atomic level, to establish conformational and chemical pathways for each polymerase, each polymerase family, and establish global paradigms for these vital processes.

2. Classify RNA tertiary motifs (A-minor, ribose zipper, junctions, pseudoknots, etc.) and understand their dependence on the primary sequence and on the environment (water, ions, ligands).

3. Predict tertiary structures of RNA from sequence.

4. Mimick in silico the *in vitro* selection process for novel RNA discovery to expand the complexity and diversity of resulting RNA aptamers.

5. Decipher the 30 nm structure of chromatin and its dependence on external (e.g., salt concentration) and internal (linker histones, length of linker DNA, etc.) factors.

6. Understand higher-levels of organization of the chromatin fiber beyond the 30-nm fiber, including fiber/fiber packing and interactions with regulatory proteins.

**Current Challenges in Biomolecular Modeling and Simulation**

1. Establish the validity and stability of biomolecular force fields and MD integrators over very long simulation times.

2. Develop better sampling tools for biomolecular conformations.

3. Develop methods for estimating transition rates of molecular processes that are applicable to complex biomolecular systems.

4. Develop systematic coarse-grained models and their integration with all-atom models.

5. Develop better methods for *ab initio* dynamics to compute reaction mechanisms, free energy pathways, and transition states of biomolecular reactions.

Addressing these challenges requires interdisciplinary collaborations and integration of efforts between experimentalists and theoreticians.

**References**


Figure 1: The evolution of molecular dynamics simulations with respect to simulation lengths (from Schlick text [53]).


Figure 2: Pol β’s Closing Pathway. Left: Molecular snapshots near open (left column) and closed (right column) states of pol β for four transition state regions [5]: (1) Partial thumb closing. (2) Asp-192 flip. (3) Arg-258 partial rotation. (4) Phe-272 flip. Right: Overall captured reaction kinetics profile (from TPS) for the conformational transition of pol β (for G:C) from open (state 1) to closed (state 7) forms showing free energies (in k_BT) associated with the different transition state regions [5]. The meta-stable basins (in red) along the reaction coordinate are numbered 1–7. See animated sequence of this closing on our website http://monod.biomath.nyu.edu/index/gallery.html.

Figure 3: In Silico Evidence for Pol β’s Induced Fit Mechanism. C_\alpha traces of superimposed pol β/DNA complex with dCTP (top left) and without dCTP (bottom left) for the intermediate starting structure (yellow), crystal closed (red), and crystal open (green) and the trajectory final structures (blue) [8]. Notable are the residue motions in the thumb subdomain and the 8-kDa domain. The positions of \alpha-helix N in the simulated systems are compared to the crystal structures in panels on the right (top, with dCTP, and bottom, without dCTP).
Figure 4: Pol β Kinetic Profile. Overall captured reaction kinetics profile for pol β’s closing transition followed by chemical incorporation of dNTP for G:C and G:A systems [11]. The barriers to chemistry (dashed peaks) are derived from experimentally measured kpol values. The profiles were constructed by employing reaction coordinate characterizing order parameters in conjunction with transition path sampling. The potential of mean force along each reaction coordinate is computed for each conformational event.

Figure 5: Pol β’s Chemical Synthesis Reaction. Left: Schematic drawing of the mechanism of concerted proton-hops during phosphoryl transfer in pol β. Solid arrows indicate the migration path of the proton and the dotted arrows represent the nucleophilic attack [13]. Right: Captured reaction intermediates for pol β’s phosphoryl transfer in the G:C system [13]: (a) reaction state of the closed nucleotide-bound enzyme state; (b) deprotonation of the the O3’H to water; (c,d) proton transfers to Asp192; (e) proton transfers to Asp190; (f) proton reaches the pyrophosphate unit to obtain the final product. The colors represent: cyan (D256), red (D190), blue (D192), pink (dCTP), green (CYT: terminal DNA primer), black (the O3’H-proton), yellow (the O3’ oxygen, attacking nucleophile), tan (central phosphorus), purple (leaving O3A oxygen), and orange (Magnesium). The oxygens and hydrogens of water molecules are in red and white, respectively. The arrows denote the location and direction of proton hop.


Figure 7: Chromatin Folding Revealed by Mesoscopic Models of Oligonucleosomes with Histone Tails. In this mesoscopic model, the grey nucleosome cores are modeled as irregularly shaped rigid bodies with uniformly distributed charges. Linker DNA, represented by red spheres, is treated by using the discrete elastic chain model. Histone tails, colored blue (H3), green (H4), yellow (H2A), and red (H2B), are modeled by a coarse-grained protein bead chain. For the bottom box, center, a 48-unit oligonucleosome is shown at 0.2 M salt. Surrounding the center structure are representative configurations of oligonucleosomes at 0.2 M salt. In these structures, the nucleosome cores are shown as white cylinders, the DNA is shown as red cylindrical tubes, and histone tails are omitted for clarity.
Figure 8: Graphic Representations of RNA Secondary Structures shown for two RNAs, using Tree and Dual Graphs, using the definitions at right [40, 39].

Figure 9: Segments from RAG showing some enumeration of graphs for tree and dual segments of low V numbers. The graphs are coded according to motifs found in nature (red), motifs not yet found that are RNA-like (blue), as determined by clustering analysis, and remaining motifs not yet found (black) [41]. See RAG website for details and updates.
<table>
<thead>
<tr>
<th>Graph Representation With Natural Submotif</th>
<th>RNA Secondary Structure With Natural Submofit</th>
<th>Candidates Discovered After 2004</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>single strand RNA (NDB:PR0055)</td>
<td>Mammalian CPEB3 Ribozyme (Ram: RF0022)</td>
</tr>
<tr>
<td>C2</td>
<td>bulged hairpin (Ram:CopA)</td>
<td>Purine Riboswitch (Ram: RF0167)</td>
</tr>
<tr>
<td>C3</td>
<td>DaV RNA (Ram:DaV)</td>
<td>Tymovirus RNA-like 3' UTR element (Ram: RF0233)</td>
</tr>
<tr>
<td>C4</td>
<td>bulged hairpin (Ram:CopA)</td>
<td></td>
</tr>
<tr>
<td>C5</td>
<td>single strand RNA (NDB:PR0055)</td>
<td></td>
</tr>
<tr>
<td>C6</td>
<td>bulged hairpin (Ram:CopA)</td>
<td></td>
</tr>
<tr>
<td>C7</td>
<td>DaV RNA (Ram:DaV)</td>
<td>Flavivirous DB element (Ram:RF0525)</td>
</tr>
<tr>
<td>C8</td>
<td>single strand RNA (NDB:PR0027)</td>
<td></td>
</tr>
<tr>
<td>C9</td>
<td>DaV RNA (Ram:DaV)</td>
<td></td>
</tr>
<tr>
<td>C10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 10: Ten candidates predicted to have RNA-like topologies, as determined by clustering analysis, represented as dual graphs; the red submotif occurs in natural RNAs [41]. Of those ten, four RNAs were discovered since the published predictions, as shown in the third column.
Figure 11: Scheme of experimental RNA pool synthesis, and in vitro selection (left), and theoretical modeling of pool generation using the mixing matrix, which specifies nucleotide mixtures in synthesis port or mutation rates for all nucleotide bases (right). Our design approaches will produce different mixing matrices that can be exploited by in vitro selection of RNAs.


Figure 12: Classification of RNA junctions into families according to coaxial stacking properties, perpendicular helical configurations, and flexible helical arms. Lines inside the helices represent the canonical WC bps GC, AU, and the GU wobble bp. The network symbology follows the Leontis Westhof notation [55].


"Opportunities in Extreme Computing for Biology Workshop"

Jeremy C. Smith ORNL.

Current research efforts of the ORNL Center for Molecular Biophysics are centered on the development of petascale and multiscale tools for the computer simulation of biological systems and on the use of extreme scale computer simulation in bioenergy, bioremediation and neutron scattering. We are particularly interested in simulations that can be performed only on supercomputers i.e, that require computer power not for trivially parallel applications (‘capacity’ computing) but require also communication between nodes (‘capability’ supercomputing). We are presently active with a DOE INCITE grant for simulations in cellulosic ethanol production.

State of the Art. We have recently performed the most powerful molecular dynamics benchmark simulations to date, scaling efficiently to 150k processors on the ORNL Jaguar XT5 (see ref. * below). These benchmarks achieve 20ns/day on 100M-atom systems.

Future prospects.

One tenth of a cell dimension for one microsecond. Even without any methodological advances the availability of 10, 100 and 1000 petaflop machines would allow us to simulate 1G-, 10G- and 100G-atom systems at 20ns/day. Hence, microsecond timescale simulations of systems of one micrometer dimension (100G atoms) will become accessible. Thus, we will be able to simulate the life of a system one tenth the dimension of a living cell at atomic detail from first principles for one microsecond. This will provide a myriad of opportunities for the examination of living processes in the cell.

Beyond the microsecond. Close collaboration between hardware and software design will be needed to extend to the millisecond and beyond using strong scaling. Current MD software on non-special purpose computers allows a maximum ~100ns/day. To achieve longer timescales improved network latency is advantageous. In this regard MD has special requirements relative to other high performance applications, as the sequential dependency between computationally inexpensive steps makes very frequent communication necessary. To be able to simulate ms-timescale events one integration step must be ~10µs long including all local and global communication. [The special purpose machine Anton makes this possible by supporting special multicasts, push operations and 50ns hop latency, and can simulate at ~ 14,000ns/day at ~0.5PF].

10-1000 Petaflops will also make possible the application of enhanced sampling techniques to larger systems. For example, using Replica Exchange (REMD) one can simulate a small ~100 residue protein (and explicit water) using the entire 150k cores of JaguarPF. With more computing power, REMD of bigger/more interesting systems becomes possible, and processes such as protein folding and ligand binding will become accessible even in the absence of timescale lengthening.

Further methodological improvements should allow faster simulations of bigger systems than the above to be achieved, but the performance improvement is difficult to predict. However, multiscaling, in which the results of atomistic simulations are used to derive coarse-grained models at larger scales, should provide a seamless connection with the cellular and supracellular world. No experimental techniques are likely to provide such a comprehensive description, although many will provide essential guiding input.
Processes that will become accessible to capability supercomputing.

(1) Accurate protein folding prediction.
(2) Accurate ligand binding prediction.
(3) Biofuel production: models for molecular machines involved in cellulosic hydrolysis up to the scale of the microbe:biomass interaction.
(4) Optimization of metabolic networks based on atomic-detail information.

Some Recent publications of the ORNL Group:

D. R. NUTT & J.C. SMITH, Dual function of the hydration layer around an antifreeze protein revealed by atomistic molecular dynamics simulations. JACS 130 13066-13073 (2008)
M. KRISHNAN & J.C. SMITH. Response of Small-Scale, Methyl Rotors to Protein-Ligand Association: A Simulation Analysis of Calmodulin-PEptide Binding. JACS. In press.