PREDICTING PROTEIN STRUCTURES WITH PHYSICAL MOLECULAR SIMULATIONS

Allocation: NSF PRAC/5.00 Mnh
PI: Ken Dill[1]
Co-PI: Alberto Perez[1]
Collaborators: Emiliano Brini[1], Joseph Morrone[1], Lane Votapka[1], and Cong Liu[1]

[1]SUNY Stony Brook University

EXECUTIVE SUMMARY

Ab initio protein folding has been a computational challenge for the last 50 years. We have developed a highly efficient platform called MELD, running on graphics processing units (GPUs) that allow the folding of protein structures in weeks of simulation time. We are using Blue Waters in a **worldwide** blind protein folding event involving ~200 scientific groups independently to assess the value of the methodology (CASP, Critical Assessment of Structure Prediction). CASP operates under very strict timelines; for four months, protein sequences are released each day, for which we have three weeks to predict the 3D structure of the protein. We are currently testing the capacity to fold proteins up to 200 residues, twice as large as has been previously possible *ab initio*.

INTRODUCTION

A long-standing grand challenge in computational biology is determining if we can use computers to find out the native structure of proteins given their sequence. This lead to IBM's effort in the 1990s with BlueGene and more recently DE Shaw's Anton supercomputer [1], which produced some of the fastest folding proteins. This approach is extremely computationally demanding and does not scale to larger and slower folding proteins.

We have developed MELD (Modeling Employing Limited Data) as a Hamiltonian and temperature replica exchange method and plugin to the program OpenMM. It has many biological applications (folding, docking, mechanisms, etc.) and runs very efficiently on GPUs. MELD's differentiating factor is the ability to incorporate noisy, sparse and ambiguous data through a Bayesian inference approach—increasing the performance to obtain protein structures by five orders of magnitude. We use coarse physical insights (e.g. proteins have hydrophobic cores) with very low signal to noise ratio. We have already produced three high accuracy structure predictions—but much more is needed to shift the field from a purely bioinformatics approach to physics-based simulations. We are participating in CASP, a blind competition extending four months, with close to 200 participating groups. This competition is the perfect scenario to witness the real life performance of MELD in the context of all the state-of-the-art methods. However, MELD as a platform goes beyond folding, and we are concurrently using Blue Waters to calculate relative binding free energies of peptides folding upon binding and to identify the most favorable oligomerization conformation for protein dimers. Our goal is to breach several milestones: (1) Fold longer proteins than previously possible *ab initio*, (2) Obtain binding free energies of peptides to proteins using MELD to do flexible ligand/protein binding and (3) Obtain binding conformations for protein dimers.

The second grand challenge is comparing the binding of several peptides to the MDM2 gene involved in cancer therapy. Traditional methods cannot obtain relative binding free energies because they take a very long time to converge. We run simulations in which two peptides are competing to bind a protein—enforcing that at any time one peptide should be in a reference state far from the protein and the other near it.

METHODS & RESULTS

We are using MELD to carry out Hamiltonian and replica exchange molecular dynamics simulations. In MELD we introduce data to guide simulations which limit the conformational space accessible to simulations, and inside those regions compatible with the data, it is physics that drive the sampling.

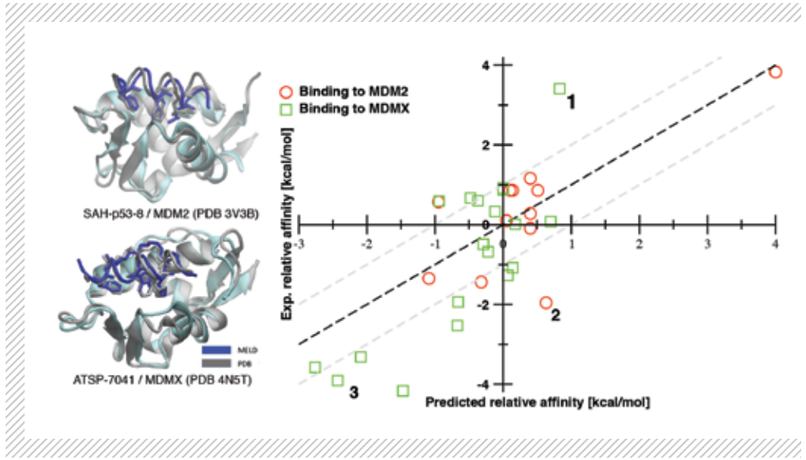

FIGURE 1: Comparison of computed vs experimental relative free energies for the binding of different peptides to target proteins MDM2 and MDMX involved in cancer pathways. The left panel shows the predictions coming from MELD on top of the experimental structure.

The biggest advantage of MELD is that the user specifies what data should be trusted [2]. The simulations then optimize two different problems: finding structures that are most compatible with the physics and the data. Solving these two problems together gives a five orders of magnitude speedup over using physics alone.

We have produced over 95 atomistic detailed protein models for proteins ranging from 80 to 230 amino acids within CASP [4]—far beyond the standard 100 residue limit for *ab initio* modeling. The CASP event continues, and we are projecting over 500 predictions by the end of the summer. The biggest advantage over database methods is that MELD provides meaningful populations which allow us to identify native conformations. We are the only physics-based group in this competition as the tight CASP deadlines are not possible with MELD's efficient sampling protocols and GPU computing.

To approach the challenge of comparing the binding of peptides to the MDM2 gene, we observe binding of one or the other at different times using MELD simulations. This protocol yields relative binding free energies in agreement with experiments (Fig. 1). This flexible receptor-flexible docking is needed for developing new drugs in which flexibility is important. In doing these two studies together (folding and binding) we are showing a pipeline that can go all the way from genomic sequencing to folding to binding, unveiling new possibilities for drug design.

WHY BLUE WATERS

With Blue Waters, we can tackle simulations of proteins that would require years of simulation time in weeks—producing high accuracy atomistic detailed protein models. In addition, we are running a time-sensitive blind prediction, in which multiple proteins need to be simulated independently. This would be impossible without Blue Waters resources.

NEXT GENERATION WORK

Our goal through the increased power of the next-generation Track-1 system is to reach the average protein size in the human body, which is 300 residues long for *ab initio* structure prediction—3 times longer than the current state of the art. Membrane proteins, the target for most pharmaceutical drugs, remain outside of our scope. We are developing a new methodology to tackle them but will need extreme supercomputer capability. We are preparing the proof of concept with Blue Waters and expect to do membrane structure prediction on a future Track-1 system.