PREDICTING PROTEIN STRUCTURES WITH PHYSICAL PETASCALE MOLECULAR SIMULATIONS

Allocation: NSF PRAC/2,900 Knh
PI: Ken A. Dill[1]
Co-PI: Alberto Perez[1]
Collaborators: Cong Liu[1], Emiliano Brini[1], James Robertson[1], Lane Votapka[1], Roy Nassar[1]

[1]Stony Brook University

EXECUTIVE SUMMARY

Blue Waters has enabled the first use of physics-based methods to make accurate atomistic predictions of protein structure based on sequence information alone in a double-blind international competition called CASP (Critical Assessment of protein Structure Prediction). This is a step forward for the computational biophysics community and an unprecedented result in the last 24 years of CASP, where physics has not had a role in *ab initio* modeling.

Further, we have used this methodology with great success to complete the pipeline that would lead to drug discovery: from sequence to structure, to interactions with small drugs, peptides, and other proteins. We are developing new protocols in these areas and have performed well in matching experimental results for the MDM2/p53 system, a cancer target. We continue to make headway in improving the accuracy and speed of our unique technology (MELD, or Modeling Employing Limited Data), which requires the significant GPU (graphics processing unit) resources available on Blue Waters.

RESEARCH CHALLENGE

A grand challenge in the last 60 years is to use computers to predict how proteins fold from a polymer sequence into a 3D structure. Proteins are the workhorses of the cells, made from 20 different amino acids whose arrangement in a linear chain gives rise to the 3D structure. We want to learn the physical principles that lead to the 3D structure so we can design drugs to bind to these proteins and stop them from misbehaving, or to design new protein sequences that carry out new functions (e.g., digest oil spills).

Computationally cheap knowledge databases are often used to copy and mimic structures based on sequence similarity. Computationally expensive physics-based methods are necessary to push the boundary of our predictive ability.

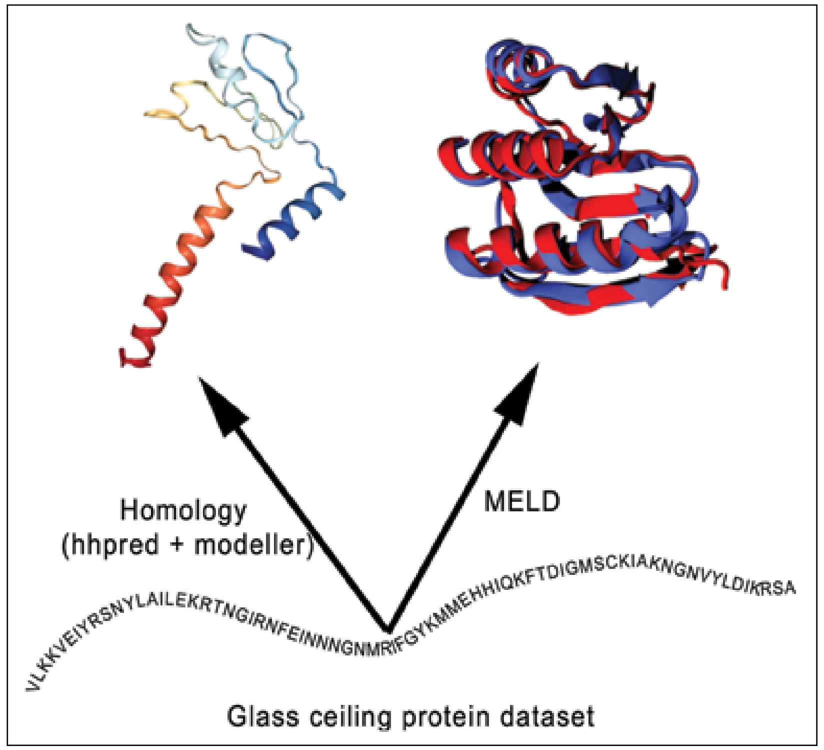
Homology
(hhpred + modeller)
MELD
VLKKVEIYRSNYLAILEKRTNGIRNFEINNNGNMRIFGYKMMEHHIQKFTDIGMSCKIAKNGNVYLDIKRSA
Glass ceiling protein dataset

Figure 1: There are 465 nonthreadable proteins in the Protein Data Bank whose structure is known. When predicting the structure based on sequence using state-of-the-art homology methods (left) the predicted structure is wrong. Using MELD and physics, we get the right answer (right: experiment in red, our prediction in blue).

METHODS & CODES

Physics models are accurate but too slow to tackle problems such as folding. MELD [1,2] is our solution. In this approach, we accelerate molecular dynamics (MD) simulations while integrating ambiguous, noisy, and sparse data through a Bayesian inference approach. MELD allows us to find the best agreement between physics and a subset of the data. On one hand, this reduces the search space of the system, accelerating the convergence time of physics-based methods. On the other hand, MELD allows us to deal with imperfect sets of information, such as data derived from a generic knowledge of the system from experiments. Our MD engine is based on the OpenMM program [3] and AMBER [4] for force fields and setup.

RESULTS & IMPACT

With Blue Waters, we have taken MELD and physics to the limit. We have participated in CASP, the blind protein structure prediction competition. This worldwide event in which more than 200 groups participate has taken place every other year for the last 24 years. During the three months of competition, hundreds of targets are released and predictions have to be submitted in a timely manner. Before MELD, atomistic physics-based simulations had not been possible this quickly. Thanks to MELD and Blue Waters, not only were we able to make predictions, some of them were the best of the entire competition. This is an unprecedented result in the last 24 years of CASP and in using physics-based approaches for structure prediction.

We put MELD to work on problems where current technologies fail. MELD is particularly suited for predicting the structure of small proteins with little homology. These are at the intersection of cases where the faster database methods do not work and where the computational cost of MELD is still acceptable. We are working on microproteins (<70 amino acids) with no known experimental structure as well as a glass-ceiling set of proteins for which the experimental structures are known [5]. We are obtaining encouraging results in both cases, and we have active collaborations with experimentalists to prove the quality of our prediction (using circular dichroism and nuclear magnetic resonance spectroscopy). In the second case, we have so far had a 25% success rate (see Fig. 1).

Sometimes protein folds upon binding to other proteins. This means that a disordered protein in solution obtains a specific 3D shape when it interacts with a specific receptor. Common computational tools fail to predict the binding pose due to the rearrangement of the structure during the process. With MELD, we have been able to predict binding poses and relative binding free energies for the complex P53–MDM2, which is involved in the development of cancer. The protocols we designed for this project are unique and have significantly advanced the field. Without Blue Waters, this project would have taken us about 16 times longer to simulate.

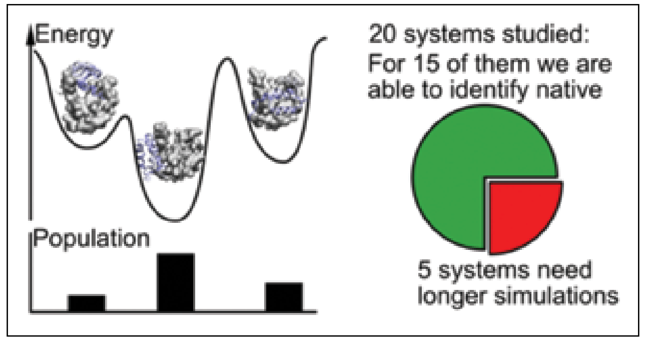

Figure 2: Protein-protein conformations predicted with MELD (top left). Populations (lower left) are related to free energies; hence, we can pick up the right answer by clustering results. Our method so far succeeds in 15 of 20 cases (right).

Proteins bind to each other and to other molecules. With MELD, we are able to investigate the way complexes of two proteins are arranged in space (see Fig. 2) and how small ligands (i.e., drugs) bind to a protein. Ultimately, physics governs both processes, and MELD allows for a tightly focused search process. This is an exciting new line of research made possible by the computational resources of Blue Waters.

WHY BLUE WATERS

Blue Waters has been the perfect resource for us. It is the only cluster in the United States where we can get enough throughput GPU (graphics processing unit) usage, especially for the time-sensitive CASP competition. The staff is also helpful and quick to respond to solve issues and help us to maximize our Blue Waters usage for efficiency. We have a 100-GPU cluster in our lab (slow 2050s and 2070s models). Every MELD calculation requires at least 30 GPUs, so at most three calculations can run in the lab, whereas on Blue Waters we can sometimes run up to 30 calculations.

PUBLICATIONS AND DATA SETS

Perez, A., et al., Blind protein structure prediction using accelerated free-energy simulations. *Science Advances*, 2 (2016), pp. e1601274–e1601274.

Morrone, J. A., et al., Molecular Simulations Identify Binding Poses and Approximate Affinities of Stapled α-Helical Peptides to MDM2 and MDMX. *J. Chem. Theory Comput.*, 13 (2017), pp. 863–869.

Morrone, J., A. Perez, J. MacCallum, and K. Dill, Computed Binding of Peptides to Proteins with MELD-Accelerated Molecular Dynamics. *J. Chem. Theory Comput.*, 13 (2017), pp. 870–876.