COMBINING PHYSICS AND SUPERCOMPUTERS TO PREDICT PROTEIN STRUCTURES FROM SEQUENCES

Allocation: NSF PRAC/5,750 Knh
PI: Ken Dill[1]
Co-PI: Alberto Perez[1]

[1]State University of New York at Stony Brook

EXECUTIVE SUMMARY

The team uses physics and supercomputers to reduce the time and cost of determining what proteins look like, atom by atom. This has been a grand challenge in computational biology for the last 50 years. We have developed the MELD (Modeling Employing Limited Data) method to run on GPUs and to combine physics with additional information using Bayesian inference. This reduces the computer time needed to fold proteins and increases the accuracy of the results.

Our MELD approach is the only atomistic physics-based approach to predict a protein's structure from its sequence that is fast enough to compete in a worldwide protein structure prediction competition (CASP). This competition has very strict deadlines, and so most methods used by participants are bioinformatics-based. We have shown that our method can be competitive and that it can bring new insights to the field. And, because it is physics-based, it gives us the option to know alternate states and pathways.

RESEARCH CHALLENGE

Proteins consist of a sequence of amino acids that folds into a 3D shape. This shape allows the protein to do its work inside our bodies. Proteins are drug targets, so developing new drugs requires knowing what these proteins look like. However, determining the structures experimentally is time-consuming, expensive, and not always possible. Bioinformatics tools are good at detecting proteins whose sequence is similar to known structures. Thus, these methods use the known structure for drug design—but there are many proteins that cannot be tackled with this approach.

Physics-based computer simulations mimicking the folding process are an alternative methodology. They used to require years of simulation time and returned incorrect structures. With new force fields [1] and our new MELD [2,3] computer method that leverages external information, we are in a great position to tackle this problem. It is very timely, as well, to collaborate with experimental labs that can produce proteins with complicated sequence information.

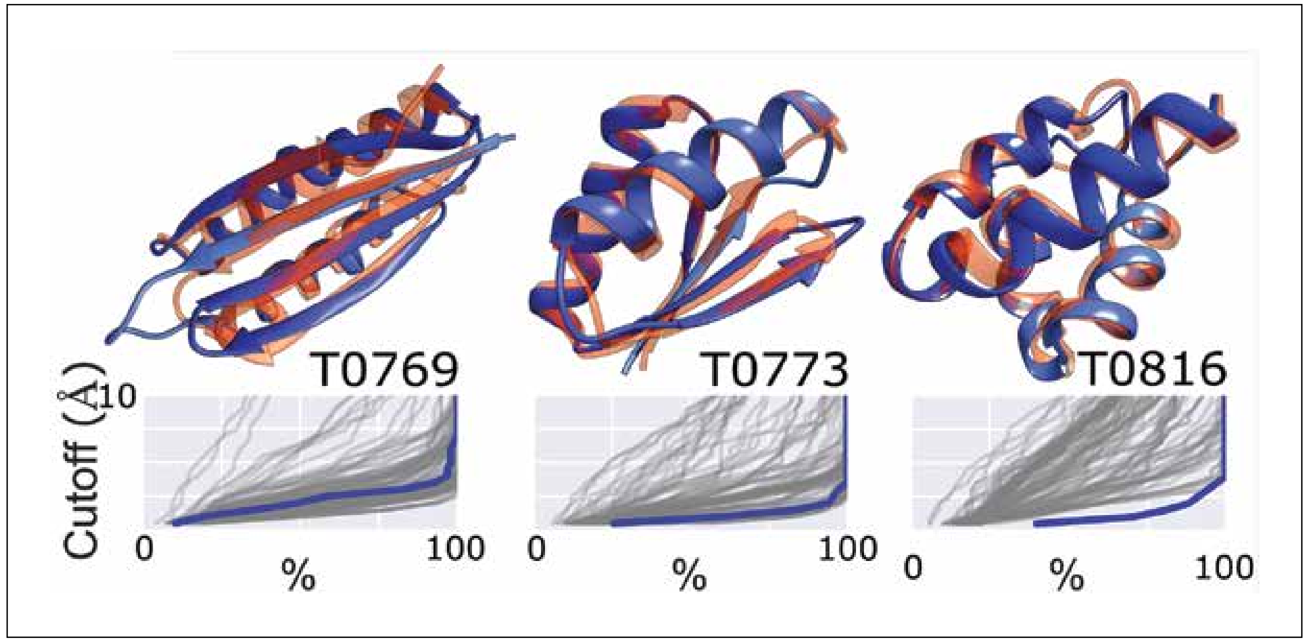

Figure 1: Past CASP results produced with MELD and Blue Waters. On top are three protein structures predicted blindly (blue ribbons), along with the comparison with the experimental structure (red ribbons). Below are the results of all groups in CASP (grey lines) and our prediction (blue line). All the results were calculated by the CASP prediction center and posted online (http://www.predictioncenter.org). For T0816, MELD was the only one to provide high-accuracy structures. (Figure adapted with permission from [5].)

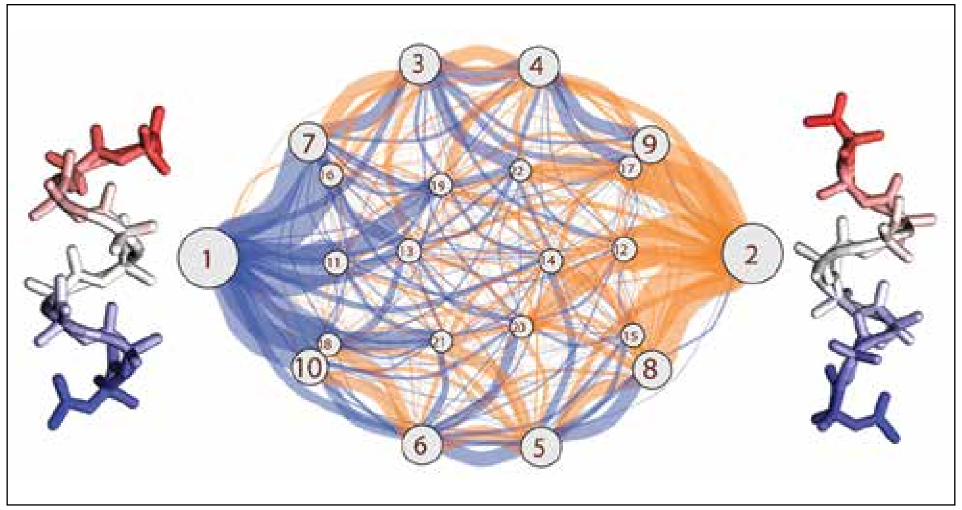

Figure 2: MELD allows us to identify states to seed multiple unbiased simulations and recover pathways. In this example, we show all states and pathways visited in exploring a helix-to-helix transition in which the system transitions from a left-handed helix conformation to a right-handed one. (Figure adapted with permission from [6].)

METHODS & CODES

We developed a plugin (MELD) to the Molecular Dynamics (MD) package OpenMM. MELD consists of a Hamiltonian and Temperature replica exchange MD protocol in which the Hamiltonian varies according to external information coming from experiment, general knowledge, or bioinformatics. What is unique about MELD is that the information is expected to be unreliable. Hence, rather than enforcing all of it, we impose only a fraction. The part to be enforced changes at every timestep and is chosen in a deterministic way.

RESULTS & IMPACT

With Blue Waters and MELD we are able to compete in a time-sensitive protein structure prediction competition (CASP [4]) that would otherwise be impossible. The competition occurs from May to August; each weekday several sequences are released, along with a three-week deadline to submit the structure predictions. With standard computer resources we might be able to target a few proteins during this time period but not the 100 or more independent protein systems that we target with the help of Blue Waters.

With Blue Waters, we have shown the ability to predict structures that fail with all other state-of-the art methods (see Fig.1) [5]. These proteins, known as being "unthreadable," come from orphan genes and have no homology to other known proteins. The abundance of unthreadable proteins encouraged us to specifically target these systems. We have already successfully folded a handful of them and look forward to folding more.

In addition, we are using the states predicted by MELD to predict the pathways that proteins fold to or use to transition between different states. We have named this set of methodologies "MELD-path" and recently published the first work produced with this methodology (see Fig. 2) [6]. We derived a set of nearly 300 states from MELD simulations and then started ~13,800 unbiased simulations using Amber on GPUs. For this system, each trajectory had to run for less than an hour. With infinite GPU resources it would have taken just one hour to run all the simulations. Using the backfilling queue on Blue Waters, we managed to collect all the data in under 17 days. This impressive feat would have taken us around 255 days to accomplish on a single GPU. Using Markov State Modeling theory we were able to recover the most relevant pathways for a helix-to-helix transition. The pathways clearly showed a preference for unfolding or refolding starting from either end of the helix and proceeding sequentially, rather than starting in random places on the helix.

WHY BLUE WATERS

Blue Waters is the only system in the United States that has enough GPUs for us to compete in CASP, and the only one that allows many jobs requiring a relatively low number of GPUs (30 each) to run for up to 48 hours. Members of the Blue Waters staff provided invaluable support in the compilation of both the Amber and OpenMM/MELD packages, which required nontrivial effort, especially during the deployment of the new Python site libraries. Furthermore, our conversations with the staff have been instrumental in improving the efficiency of running jobs during the CASP competition.

PUBLICATIONS & DATA SETS

Perez, A., et al., MELD-Path Efficiently Computes Conformational Transitions, Including Multiple and Diverse Paths. *J. Chem. Theory Comput.*, 14 (2018), pp. 2109–2116.