GA

BW

BI

PETASCALE INTEGRATIVE APPROACHES TO PROTEIN STRUCTURE PREDICTION

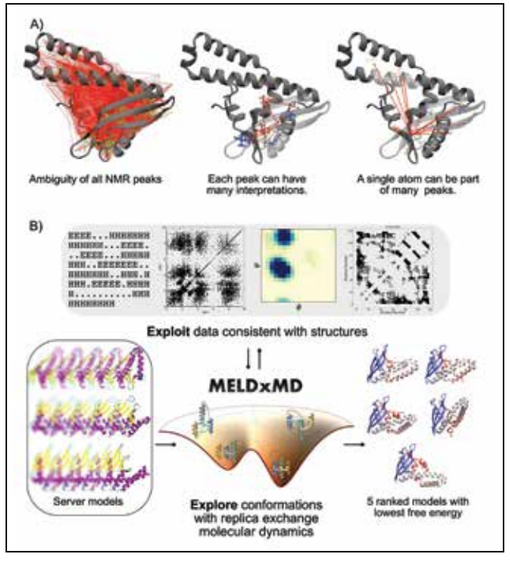
Allocation: NSF PRAC/5,600 Knh
PI: Ken Dill[1]
Co-PI: Alberto Perez[2]

[1]SUNY at Stony Brook
[2]University of Florida

EXECUTIVE SUMMARY

The research team applies physics-based molecular dynamics computations to protein modeling by leveraging external information through the team's Modeling Employing Limited Data (MELD) accelerator method. MELD runs on GPUs and is able to harness sparse, noisy, and ambiguous information using "sub-hay-stacking" Bayesian inference. It can lead to orders-of-magnitude speedups in protein folding and protein–protein docking over traditional molecular dynamics (MD) methods.

MELD x MD was ranked first in last summer's Critical Assessment of Structure Prediction (CASP) protein structure prediction competition in the experimental NMR data-assisted predictions category. This event tests how well computations can utilize real-world NMR data to determine high-resolution protein structures. This is an important blind test that shows that molecular dynamics forcefields can meet the challenge and that MELD accelerates these computations sufficiently so that it can address large enough protein sizes to become an important new method for structural biology.



RESEARCH CHALLENGE

A central step in understanding how proteins perform their biological actions and how to design drugs to inhibit or activate them is to know a protein's 3D atomistic structure. Experimental methods, such as X-ray crystallography, nuclear magnetic resonance spectroscopy (NMR), and cryo-electron microscopy (cryo-EM) provide the underlying data, but such methods require a computational means of converting that data into a meaningful structure. Different experiments have different limitations: data can be sparse, or ambiguous, or noisy and combinatoric. Researchers need computational approaches that can handle these limitations. The best would be physics-based simulations, properly sampled, to provide proper Boltzmann populations and free energies.

This is the challenge the research team is addressing with its MELD-accelerated molecular dynamics (MELD x MD) [1,2]. MELD x MD is a substantial enhancement of traditional MD in biomolecular simulations because it allows for the exploration of more limited data in determining protein structures or larger proteins [3] and larger motions in all the applications of physical molecular simulations to matters of protein structure and mechanism [4].

METHODS & CODES

The team developed a plugin (MELD) to the MD package OpenMM [5]. 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, only a fraction is enforced. The part to be enforced changes at every timestep and is chosen in a deterministic way.

RESULTS & IMPACT

High-performance computing (HPC) on Blue Waters has been essential to this work in two ways: (1) the invention and development of the MELD method, and its application to proteins, has been very costly computationally and could not have been done without these national HPC resources; and (2) the principal way that researchers measure success of their methods, and compare them to other methods in the field, is through blind competitions, such as the biennial CASP event. Physics-based methods have been too computationally slow to enter these many-protein time-limited events in the past. CASP13 in 2018 involved 100 protein challenges, each with a three-week computational deadline. MELD x MD now brings physical methods into the realm of these real-world comparative tests.

Figure 1: Noisy and ambiguous NMR data pose unique challenges during the CASP13 competition.

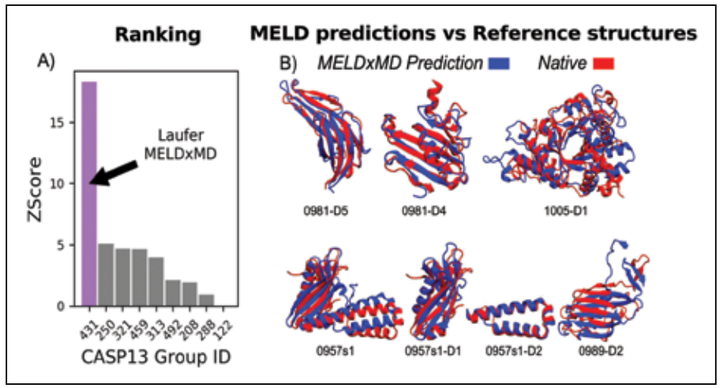


Figure 2: MELD x MD was the highest-ranked group in NMR data-assisted CASP13. a) The MELD x MD method (group code 431) had the highest Z-score of all groups that competed in NMR data-assisted CASP13. b) MELD x MD predictions are aligned to reference structures.

We believe that MELD x MD is potentially field-changing [1–4,6]. It can bring orders-of-magnitude greater speedups to MD simulations of a broad range of protein properties. From the NMR event and other such CASP tests, we are learning that the forcefields [7] and solvent models are quite good [8] and are often now better than many bioinformatics and database-driven approaches [9], with the added important power that they provide proper populations and free energies [10,11]. We are learning that MELD x MD can succeed in many venues—assisting in structure determination [3,6], *ab initio* protein folding [1,2,9], ligand binding, protein–protein docking [13], and others. Going forward, we believe the method will have additional niches important to scientists working to understand structures, dynamics, mechanisms, and to improve drug discovery. Two such applications currently in progress are assisting with cryo-EM to determine large protein structures and complexes, and determining the structures of amyloid aggregates, such as those that are problematic across the spectrum of neurodegenerative diseases.

WHY BLUE WATERS

Blue Waters is the only system in the United States that has enough GPUs for the research team to compete in CASP and allows many jobs using a relatively low number of GPUs (30 each) to run for up to 48 hours. Compilation of both Amber and OpenMM/MELD have not been trivial to optimize, and support from project staff has been invaluable, especially during the deployment of the new Python standard libraries. Furthermore, conversations with the staff have been invaluable to set up ways to run jobs efficiently during the CASP competition.

PUBLICATIONS & DATA SETS

A. Perez, J. L. MacCallum, and K. A. Dill, "Accelerating molecular simulations of proteins using Bayesian inference on weak information," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 112, no. 38, pp. 11846–11851, 2015.

A. Perez, J. A. Morrone, E. Brini, J. L. MacCallum, and K. A. Dill, "Blind protein structure prediction using accelerated free-energy simulations," *Sci. Adv.*, vol. 2, no.11, p. e1601274, 2016.

J. A. Morrone, A. Perez, J. L MacCallum, and K. A. Dill, "Computed binding of peptides to proteins with MELD-accelerated molecular dynamics," *J. Chem. Theory Comput.*, vol. 13, no. 2, pp. 870–876, 2017.

J. A. Morrone *et al.*, "Molecular simulations identify binding poses and approximate affinities of stapled α-helical peptides to MDM2 and MDMX," *J. Chem. Theory Comput.*, vol. 13, no. 2, pp. 863–869, 2017.

J. C. Robertson, A. Perez, and K. A. Dill, "MELD x MD folds nonthreadables, giving native structures and populations," *J. Chem. Theory Comput.*, vol. 14, no.12, pp. 6734–6740, 2018.

A. Perez, F. Sittel, G. Stock, and K. Dill, "MELD-path efficiently computes conformational transitions, including multiple and diverse paths," *J. Chem. Theory Comput.*, vol.14, no. 4, pp. 2109–2116, 2018.

E. Brini, D. Kozakov, and K. A. Dill, "Predicting protein dimer structures using MELD x MD," *J. Chem. Theory Comput.*, 2019.

J. C. Robertson *et al.*, "NMR-assisted protein structure prediction with MELD x MD," *Proteins: Structure, Funct., and Bioinform.*, vol. 87, no. 12, pp. 1333–1340, 2019. doi: 10.1002/prot.25788