Title
Accelerating molecular simulations of proteins using Bayesian inference on weak information

Journal
Proceedings of the National Academy of Sciences of the United States of America, 112(38)

ISSN
0027-8424

Authors
Perez, Alberto
MacCallum, Justin L
Dill, Ken A

Publication Date
2015-09-22

DOI
10.1073/pnas.1515561112

Peer reviewed

Accelerating molecular simulations of proteins using Bayesian inference on weak information

Alberto Perez[a,1,2], Justin L. MacCallum[b,1], and Ken A. Dill[a,c,d,2]

[a]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, NY 11794; [b]Department of Chemistry, University of Calgary, Calgary, AB T2N 1N4, Canada; [c]Department of Chemistry, Stony Brook University, Stony Brook, NY 11794; and [d]Department of Physics & Astronomy, Stony Brook University, Stony Brook, NY 11794

Contributed by Ken A. Dill, August 7, 2015 (sent for review June 27, 2015; reviewed by Jie Liang)

Atomistic molecular dynamics (MD) simulations of protein molecules are too computationally expensive to predict most native structures from amino acid sequences. Here, we integrate "weak" external knowledge into folding simulations to predict protein structures, given their sequence. For example, we instruct the computer "to form a hydrophobic core," "to form good secondary structures," or "to seek a compact state." This kind of information has been too combinatoric, nonspecific, and vague to help guide MD simulations before. Within atomistic replica-exchange molecular dynamics (REMD), we develop a statistical mechanical framework, modeling using limited data with coarse physical insight(s) (MELD + CPI), for harnessing weak information. As a test, we apply MELD + CPI to predict the native structures of 20 small proteins. MELD + CPI samples to within less than 3.2 Å from native for all 20 and correctly chooses the native structures (<4 Å) for 15 of them, including ubiquitin, a millisecond folder. MELD + CPI is up to five orders of magnitude faster than brute-force MD, satisfies detailed balance, and should scale well to larger proteins. MELD + CPI may be useful where physics-based simulations are needed to study protein mechanisms and populations and where we have some heuristic or coarse physical knowledge about states of interest.

protein folding | molecular dynamics | integrative structural biology | Bayesian inference

Computer modeling is an important source of insights into the properties of protein molecules. There are two main approaches, each with different main areas of applicability: comparative modeling and atomistic molecular dynamics (MD) simulations. Comparative modeling draws inferences from a database of the more than 100,000 known native structures of proteins (1); it is an information-centric approach. A key area of applicability is in predicting the native structures of previously unknown proteins. These methods are often tested in the community-wide blind event for predicting native protein structures, called community assessment of structure prediction (2, 3). In contrast, physics-based atomistic simulations are aimed at computing proper relative populations of the many different states of a system; this type of modeling is an energy-centric approach. Computing proper populations (or, correspondingly, free energies) is essential for elucidating stabilities, motions, and mechanistic actions of protein molecules.

Physical simulations offer important advantages in the long run, providing a principled and transferrable basis for understanding properties; the capability to go beyond just native structures alone to dynamics, binding, folding, and mechanisms; applicability where databases are limited, including membrane proteins or other foldable polymers, such as peptoids (4); and extensibility to other temperatures, solvents, and binding conditions, for example. A proper physical model requires a plausible physical energy function that can accurately predict native structures (validation); that applies across many different proteins (transferrable); that satisfies Boltzmann's law (physical); that scales up to sufficiently large proteins (practical); and, when predicting folding, that begins from the fully unfolded state (to avoid inadvertent biases). These objectives are largely not met by bioinformatics algorithms, which do not satisfy Boltzmann's law, or by current atomistic simulations, which are too computationally expensive to tackle sizable proteins starting from fully unfolded states.

Major Challenge in MD Is Conformational Sampling

MD simulations are computationally expensive for the levels of conformational sampling needed to fold proteins from unfolded states. Integrating Newton's equations of motion a few femtoseconds at a time (required for satisfactory approximation of differential equations by difference equations), finding the native state can take millions (microseconds) or billions (milliseconds) of integrations, which translates into weeks, months, and even years of computer time depending on system size and machine architecture. However, in many situations, we care mostly about particular "states of interest." For example, for protein folding, one key state of interest is the protein's native structure. For mechanistic actions, we may know something about the structures of the beginning and ending states. The present work focuses on problems involving particular states of interest, even when we do not know their exact structures.

There is a long history of integrating information-centric with energy-centric methods in seeking states of interest. Integrative structural biology combines them, for example, in pioneering methods, such as Modeler (5, 6); methods based on Rosetta (7–9); and others (10). However, in such marriages, the energetic modeling is secondary; it does not satisfy Boltzmann's law or give proper populations or free energies. Here, because our end goal is fundamentally to get proper populations, we seek a method

Significance

An important challenge has been to develop computer methods that can predict protein native structures from their sequences and satisfy the thermodynamic principle of Boltzmann's Law, which requires that the sampling method obey detailed balance. The latter is needed to study mechanisms and dynamics, which require an understanding of relative populations of states. These dual goals are met by atomistic model simulations, but they have been too expensive computationally. Here, we join together molecular dynamics (MD) with Bayesian inferences derived from loose insights (proteins have "hydrophobic cores" and "secondary structures"). We show that this method can speed up MD simulations by up to five orders of magnitude, allowing for the accurate predictions of small native protein structures with only atomistic potentials.

Author contributions: A.P., J.L.M., and K.A.D. designed research; A.P. performed research; A.P. and J.L.M. analyzed data; and A.P., J.L.M., and K.A.D. wrote the paper.

Reviewers included: J.L., University of Illinois at Chicago.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]A.P. and J.L.M. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: alberto@laufercenter.org or dill@laufercenter.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1515561112/-/DCSupplemental.

that satisfies detailed balance. We take the energy-centric approach as primary.

How might we guide MD simulations to states of interest when we do not know what those structures are? We describe an approach based on coarse physical insight(s) (CPI), that is, heuristic knowledge about the states of interest. For example, we know the generic features of single-domain, water-soluble, globular proteins. They have hydrophobic cores. They have substantial secondary structures and are compact. They have β-strands that are usually paired. Such information alone is much too vague, nondirective, and combinatoric for a computer algorithm to find the correct native structure, given only an amino acid sequence. However, we show here how that level of "weak information" can be used to create multiple funnels on MD energy landscapes, accelerating conformational search while preserving the relative populations of the states of interest.

Method of Modeling Using Limited Data + CPI

Our approach has two components. First, modeling using limited data (MELD) is a Bayesian inference approach (11). It combines, on the one hand, prior information (Eq. **1**) based on MD simulations of an atomistic model with the underlying distribution coming from a force field. On the other hand, sparse, ambiguous, and uncertain information for the determination of protein structures is used and evaluated as the likelihood that each structure is compatible with the information (11) (Eq. **1**). Sparse refers to data that are accurate but insufficient on their own to specify a structure. Ambiguous refers to data that are not very precise or where there are different possible interpretations. Uncertain refers to data that are only partially correct, where a subset of information is wrong and would lead to incorrect structures. MELD integrates data that is limited in these ways with Hamiltonian and temperature replica-exchange molecular dynamics (H,T-REMD) simulations to refine protein structures:

$$p(x|D) = \frac{p(D|x)p(x)}{p(D)} \sim p(D|x)p(x), \qquad [1]$$

where x represents structures, D represents experimental data, $p(x|D)$ is the probability of the structure given the data, $p(D|x)$ is the likelihood of the data given the structure, $p(x)$ is the Boltzmann probability distribution of structures from the atomistic force field model, and $p(D)$ is an irrelevant normalization factor. Restraints are used to incorporate the data into simulations.

The second component of our method is the use of CPI to guide REMD simulations toward states of interest. In particular, we illustrate the principles on a problem of finding protein native structures from extended chain states using REMD. The CPIs that we use here are (*i*) that proteins have secondary structures, (*ii*) that proteins have hydrophobic cores, (*iii*) that β-strands pair up, and (*iv*) that proteins have compact structures. The challenge is in how to formulate these well-known rules into a formulation that is more directive than misdirective in an MD simulation.

We do not know which particular interactions will be satisfied in a given protein. Instead, from collecting statistics in the Protein Data Bank (PDB) before simulations, we know the fraction, f_{CPI}, of the possible interactions that will typically be satisfied. For example, a globular protein of up to 100 residues typically makes 8% of its possible hydrophobic contacts ($f_{hyd} = 0.08$), and 70–80% of secondary structure predictions from Psi-blast–based secondary structure prediction (PSIPRED) (12, 13) or PORTER (14, 15) are typically correct ($f_{ss} = 0.8$). The combinatorics of CPIs have a small directive signal toward folding: Only a few of the exponentially many possible combinations are consistent with the native structure. MELD + CPI simultaneously infers both which restraints are correct and the corresponding structural ensemble. Full details of MELD + CPI are given in *Materials and Methods* and *SI Appendix, Methods*, and details of MELD are provided elsewhere (10).

Each type of CPI is turned into a set of possible restraints with a flat-bottom harmonic functional form (*SI Appendix, Methods*). Then, at each time step, given the current configuration, and for each type of CPI, MELD + CPI will sort all of the restraints by energy and will activate the fraction f restraints with lowest energy, the "least-stretched heuristic restraints," to guide the simulation until the next time step. Choosing these least-stretched springs is very fast and reduces the combinatoric problem to deterministic choice. MELD + CPI uses Hamiltonian and temperature replica exchange, where the restraints are weak at the highest temperature, whereas the restraints are strong at the lowest temperature. This pipeline is illustrated schematically in Fig. 1 in an HP lattice model. The Hamiltonian and temperature change in the replica exchange. At the highest replica, the restraint force constants are zero; hence, configurations are sampled all over the potential energy surface (PES). Moving down in the replica ladder, the spring constants increase, funneling the PES toward regions compatible with different combinations of springs. Because the springs have a flat bottom, the spring energy (and force) is zero inside the funneled region. Hence, the sampling inside those regions is just driven by the force field. The relative populations inside such different regions are the same as in the original force field. Because the restraint energy is always greater than or equal to zero, regions that were not preferred by the force field before will not become stabilized.

Fig. 2 shows in a qualitative way how this procedure makes the landscape more funneled and frustrated. Under the influence of the springs, it is not possible to exchange from one minimum to another. To escape those valleys, excursions to higher replicas are needed. The temperature increases and spring force constants decrease as a "walker" moves to higher replicas. Thus, the

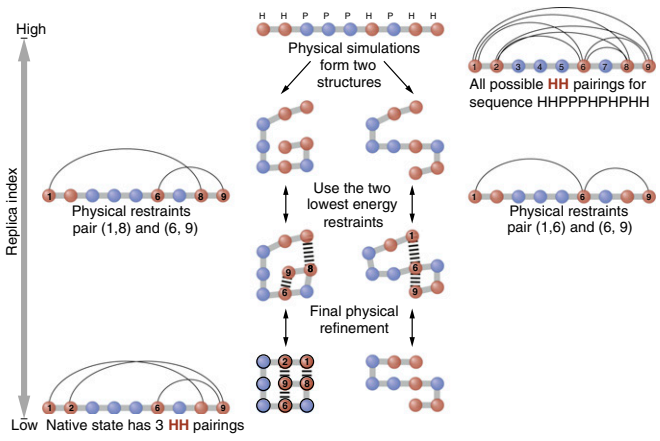

Fig. 1. Illustrating how MELD + CPI works, in terms of molecular structures. The principle of the method is simplest to convey by using a toy HP lattice model of a short chain in two dimensions. We are given the sequence HHPPPHPHH. For simplicity, the heuristic information (CPI) that we use in this case is just the pairing of hydrophobics in HH contacts. (*Top Right*) All seven possible HH pairings are shown. Our starting knowledge is that the native structure will have about 2 HH contacts, but we do not know which ones. The second row of the figure shows two possible conformations that are achieved after partial conformational sampling. The third row shows that for a certain conformation, only the lowest restraint energy HH contact springs will be guiding the system (i.e., those contact springs that are most compatible with each given conformation). The fourth row shows the conformations that those springs lead to. Based on the populations (or number of HH contacts in this simple model), we can differentiate which of those two conformations will be the native state. Note that there are many other pathways leading to other conformations. These conformations were found by a combination of the physical simulation plus the two heuristic springs that were imposed by using the knowledge that the protein should have a hydrophobic core.

PES becomes less frustrated, and there is more kinetic energy to allow greater configuration sampling, driving each walker in the REMD to new regions compatible with different springs.

In summary, we sample from a multifunneled energy landscape using H,T-REMD. Increasing the temperature weakens both the physical and heuristic-restraint interactions. The H,T-REMD serves to move from different regions of conformational space. At the highest replica, the temperature is high and the restraints are inactive. Hence, the sampling is broad, covering the entire conformational landscape. As we go down the replica-exchange ladder, the temperature decreases and the restraints become stronger, efficiently funneling down toward the region of conformational space where the set of f_{CPI} contacts are satisfied (restraint energy = 0). We use flat-bottom harmonic potentials so that multiple microstates are compatible with this f_{CPI}. At the end of the simulations, the last half of the bottom five replicas is combined and clustered using average-linkage clustering (13). The centroids of the top five clusters by population are used as representative of the folded state.

Results

MELD + CPI Samples Near-Native Structures Very Efficiently. We applied MELD + CPI to 20 small proteins (*SI Appendix*, Table S1) drawn from two datasets (14, 16), ranging from 20 to 92 residues in length. We assessed our predicted folded structures using three different measures. The first, *Best1Pop*, reports the backbone rmsd of the centroid of the single most populated cluster. The second, *Best5Pop*, reports the lowest backbone rmsd from the centroids of the five most populated clusters. These two measures test the combined success of the force field and the completeness of the conformational sampling. The third measure, *BestStruc*, reports the lowest backbone rmsd of any single structure sampled in the simulations. This test is more specific of just MELD + CPI itself, which helps us to distinguish any flaws of MELD + CPI from flaws of the force field, per se. We define success as a backbone rmsd difference of the predicted and experimental structures below 4 Å.

By the *BestStruc* criterion, MELD + CPI is successful at sampling native and near-native structures for all 20 of the proteins (*SI Appendix*, Table S2). By the *Best5Pop* criterion, we find that MELD + CPI successfully identifies the native topology for 15 of the 20 targets (Fig. 3 and *SI Appendix*, Table S2). By the *Best1Pop* criterion, we find that the single lowest free-energy cluster had a backbone rmsd below 4 Å in 11 of 20 cases (*SI Appendix*, Table S2). These results show that the principal limitations are the force field (17), the solvent model (18, 19), or the sampling time, rather than the coverage of conformational space, indicating that these factors are not limitations of MELD + CPI, which is a sampling method.

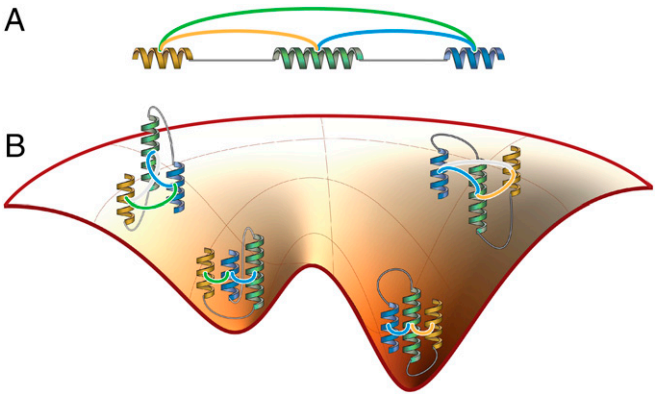

Fig. 2. Illustrating how MELD + CPI works, in terms of energy landscapes. An illustration of the same principle, except using an energy landscape picture instead of a lattice model and using helix pairings instead of HH contacts as a heuristic principle, is shown. (*A*) Suppose a heuristic identifies three possible pairings of helices. Suppose that we know that, on average, only two pairings occur in the native structure, but we do not know which ones. (*B*) MELD + CPI creates funnels on the energy landscape in specific regions that satisfy different CPI. The molecular simulation finds the deepest of these wells.

Fig. 4 compares the performance of MELD + CPI against unconstrained MD simulations (20). Although this comparison is not an apples-to-apples comparison, because those simulations were not all initiated from fully unfolded states, used a different solvation model, and were targeted at questions of kinetics (21), there are few simulations more relevant for comparison. MELD + CPI is up to five orders of magnitude faster than the estimated folding times from those unrestrained simulations (Fig. 4*A*).

Unrestrained MD simulations were performed by Nguyen et al. (22), whose work is similar to ours in the force-field, implicit-solvent model (23); REMD sampling; and initiation from completely extended chain states (22). Fig. 4*B* shows "receiver operating characteristic-like" (ROC-like) plots indicating that MELD + CPI finds conformations near the native structure much more effectively than the 500 ns of unrestrained REMD (22) (Fig. 3*B*) or the longer time simulations (22) from that work. MELD + CPI samples native states in all 20 proteins, whereas only a fraction of them were sampled by using MD (22) or T-REMD (22). Especially relevant is protein G (PDB ID code 1mi0): Neither an ∼60-μs MD run nor an ∼30-μs REMD run can reach native-like structures (*Best5Pop* = 7.5 Å in the unrestrained REMD) (22). Fig. 4 *C* and *D* shows a more complete measure of performance than either simulation time alone or sampling efficiency alone. A simple measure, which reflects both sampling effectiveness and search speed, is $P = \frac{f_{\text{folded}}}{t}$, where P is the performance (higher P means more efficient simulations), f_{folded} is the fraction of structures in the full ensemble that are less than 4 Å rmsd from the native structure, and t is the total simulation time (including all replicas). By this definition, P is also useful for rating the success of single trajectories on the same footing as sampling from replica exchange. Fig. 4*C* shows that MELD + CPI has better sampling performance than in the corresponding standard REMD simulations. Fig. 4 *C* and *D* shows that for very small proteins, there is not much advantage to using this strategy, because residues are close enough that they will often come in contact due to thermal fluctuations. However, as the system gets larger, MELD + CPI provides an improvement in efficiency.

The advantage of physics-based strategies is having a proper thermodynamic way to identify the native state based on populations. MELD + CPI is based in REMD so it obeys detailed balance (11) and hence populations are meaningful. For populations to be significant, the REMD should be converged. *SI Appendix*, Fig. S1 shows the convergence of the REMD ladder by plotting the rmsd distributions to the same random structure of every walker as it visits different replicas. The greater the overlap between the different distributions, the closer they are to convergence. For proteins like 1fme, 1prm, 2f4k, or 2jof, the distributions overlap significantly, increasing the likelihood of success on clustering. On the other hand, proteins like 1lmb, 1ubq, or 2hba would require more sampling to converge the REMD. When taking the same measure based on a native-centric view (*SI Appendix*, Fig. S2), we can count how many independent replicas have found native-like conformations. The higher the number, the more likely is identification of the native state. Longer simulations would increase convergence and the amount of cases in which just the first cluster is enough to identify the native state. Convergence of the simulations is out of the scope of this paper and is not considered in the extrapolations of Fig. 4*D*. Clustering is performed based on structure similarity: Unfolded structures are structurally diverse from each other, leading to small clusters, whereas native-like structures are

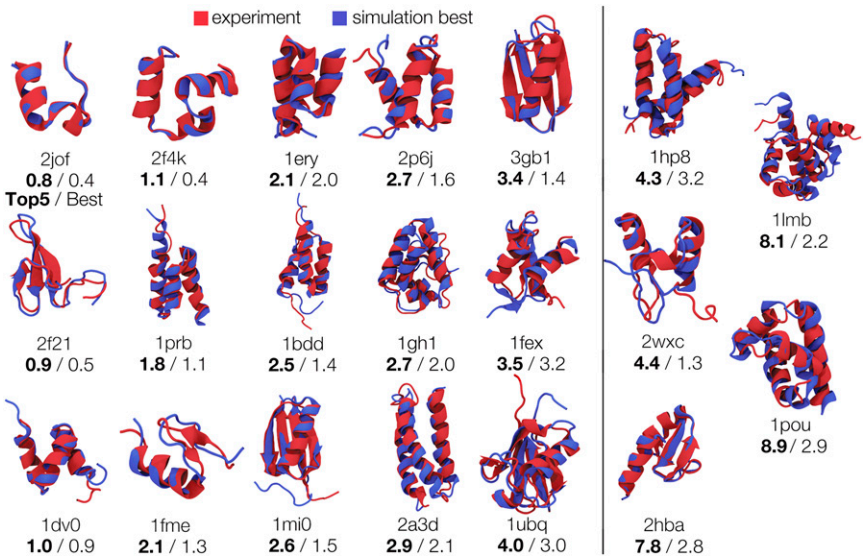

Fig. 3. Predicted native vs. experiments. The *Best5Pop* rmsd is indicated in bold; the *BestStruct* is not indicated in bold and is shown superposed on the native structure. Predictions that are closer than 4 Å rmsd to native in *Best5Pop* are shown above the line.

clustered together, leading to higher populations, allowing us to identify native states even in some cases where the replica-exchange ladder is not converged.

Of special interest were the five proteins that sampled the native structures well but did not identify them. Here, we can distinguish between force field errors and convergence problems. We reran these simulations starting from the native state (*SI Appendix*, Fig. S3). We find that the native state is only stable for one of the five proteins (2hba). So, for the other four proteins, the problem is the force field rather than the convergence. For 2hba, expanding the folding trajectory of 2hba from 500 to 800 ns starting from the unfolded state (*SI Appendix*, Fig. S4) shows an increase in the native-like population, demonstrating that our convergence was the problem in this case.

Different CPI-Restraint Types Play Different Roles in Reaching Native Structures. We use different temperature dependencies for our restraints in the REMD temperature ladders. Our restraints on

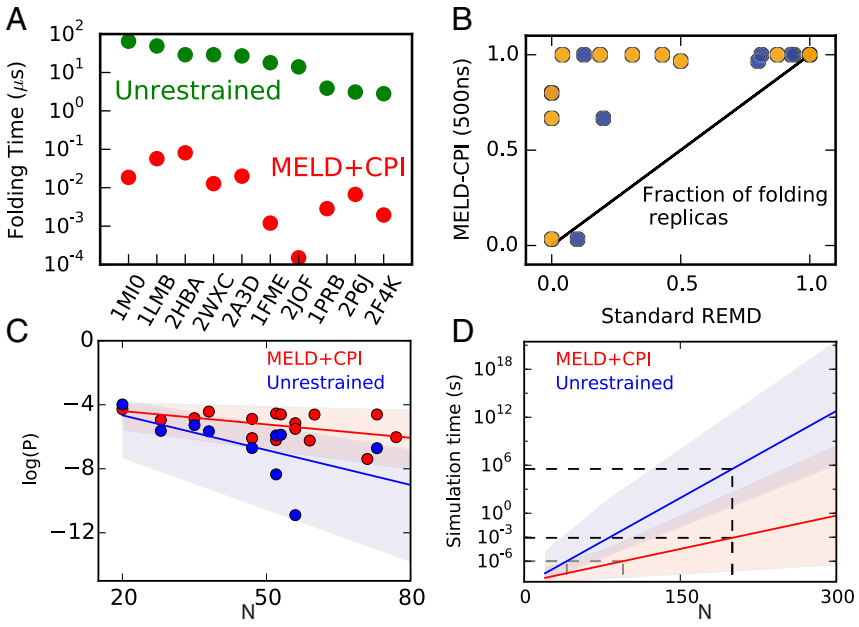

Fig. 4. Performance of MELD + CPI vs. unrestrained MD simulations. (*A*) Comparison of time to fold for MELD + CPI vs. average folding time predicted from explicit solvent MD (14). (*B*) Receiver operating characteristic plot: The *y* axis is the fractional native sampling by MELD + CPI in 500 ns of simulation, whereas the *x* axis is the corresponding fraction of native sampling by unrestrained REMD (22). Orange dots indicate 500 ns of sampling in the unrestrained REMD, and blue dots indicate the whole unrestrained REMD trajectory (22). (*C*) Performance *P* (main text) of MELD + CPI vs. unrestrained implicit solvent simulations (22). *N* is the number of residues in the protein. (*D*) Predicted simulated time from extrapolation of data in *C* to longer chain lengths. The simulation times represent the time needed to achieve a population of 0.01 native in the ensemble. Dashed gray lines indicate the expected protein size that can be sampled in 1 μs. Dashed black lines indicate the expected simulation time for a 200-residue protein with both methods. In *D*, the scalings are projected to longer protein chains. These extrapolations are just based on the sampling, and are not intended to address the scaling of force-field inaccuracies that will also increase with system size. For 200-mer proteins, the figure shows that the MELD + CPI recruitment of external heuristic knowledge should reduce the computational costs by about nine orders of magnitude relative to pure brute-force MD.

secondary structures and compactness are formed over a wide range of temperatures, whereas our hydrophobic and strand-pairing restraints are scaled to weaken to a force constant of zero at high temperatures. This procedure loosely mimics the folding-kinetics idea of zipping and assembly (24), namely, that local structures (secondary structures) form early in folding and nonlocal interactions form later. One general observation is that the more diverse the information, the faster is the computational first passage time. Hence, we expect that introducing other types of heuristic information, from experiments or evolution (17), might speed up simulations further.

We studied our ubiquitin simulation. Ubiquitin is a challenge for brute-force MD due to its slow folding time, but it is folded well by MELD + CPI. We studied the role of the different types of CPIs (*SI Appendix*, Table S3) in accelerating the folding of ubiquitin. In the native state, 18% of the possible hydrophobic contacts are satisfied in ubiquitin, compared with only the 8% that we imposed, which is representative of the PDB. We asked whether adding more hydrophobic restraints would have improved the results. We found improvement of our best rmsd structure (*BestStruct*) by 0.6 Å, but we were not able to detect the native state in the top five clusters. This failure could indicate a longer convergence time when the accuracy is close to the real native accuracy (there are many possible sets of hydrophobic pairs enforcing 8% of the restraints but only one that enforces the correct 18%), or it could be an effect of backtracking (25).

To test this balance between sampling correct structures and identifying them further, we tried to fold ubiquitin only using secondary structure predictions. Surprisingly, our *BestStruct* is close to the case where we use hydrophobic contacts and strand pairing. However, the clustering results are significantly worse (4 vs. 8 Å). The heuristic on the secondary structure is a local one: It limits the conformational sampling based on the local environment (helix or strand) but provides no information about long-range interactions. At the other extreme, hydrophobic contacts and strand pairings give us long-range information but do not impose restrictions on the local environment. This set-up leads to many correct, but not stable, contacts. Without secondary structure restraints, our simulations did not sample the native state. Hence, there needs to be a balance in the restraints: Long-ranged contacts overcome diffusive barriers, whereas short-ranged ones predispose the local environment to stable long-range interactions. Without the correct local environment, successful long-range interactions are less likely to happen.

How Can We Measure the Performance of Constrained Conformational Search Methods? How can we measure the performance of computer methods that aim for both speed and accuracy in predicting native protein structures? Computational speed is simple to determine. Here, we want to know how well a conformational search method, such as the present one, is able to explore a localized targeted space, such as around the native structure. We focus on how much the method restricts conformational searching. The Flory–Huggins (FH) theory of polymer chain conformations (11) gives us a physical basis for computing the reduction in conformational searching due to different numbers of constraints, in a mean-field approximation. In FH theory, ρ is the number of contacts made in the chain divided by the maximum possible number of contacts, so this value corresponds to the fraction of the maximum possible number of springs that could possibly be enforced. So ρ goes from 0 (no spring restraints) to 1 (maximally compact structure defined by springs). Hence, $\Delta S(\rho) = R \ln W(\rho)$ is the FH conformational chain entropy as a function of the relative number of such spring constraints, and W is the size of the conformational space. The conformational entropy of the remaining degrees of freedom can also be described as $\Delta S(\rho) = [(1-\rho)/\rho] * \log(1-\rho)$, which is a mean-field estimate of the reduction of conformational searching as a function of informational springs. *SI Appendix*, Fig. S5 exemplifies this point: in *A*, three proteins are simulated with different types of heuristics restraints (2HBA, protein G, and ubiquitin), showing that as the number of springs increases, so does accuracy. *SI Appendix*, Fig. S5*B* shows the increase in performance compared with simulations without springs with an increased fraction of restraints (ρ). The plots showcase the ability to identify native states better by clustering with shorter simulations as the amount of restraints increases relative to a given protein chain length.

What Are the "Computational Pathways" to the Native State? We have studied the restraint pathways that MELD + CPI finds as it seeks the native structure. These restrain pathways are not physical pathways because the intermediate states include restraint potentials; these restrain pathways are just sequences of events that are observed well in the REMD simulations from one restraint to the next on the way to the native structure. However, at the end points of our computational folding, there are no restraints still operative, because they are flat-bottom potentials. Just as in physical protein folding, MELD + CPI produces different microscopic routes to the native structures (*SI Appendix*, Fig. S2). We have used the MSMBuilder tool (20) to cluster and process the information from the 30 replicas for each protein. Our interest is in understanding how MELD + CPI and REMD help to guide and accelerate folding, rather than trying to understand the physical folding kinetics, which do not make sense in our REMD scheme. We track p-fold values, replica indexes, and rmsd for each of the states identified by MSMBuilder and then use MSMExplorer (21) to visualize the resulting pathways.

We make two observations: (*i*) the MELD + CPI procedure explores multiple topologies in parallel through independent walkers, and (*ii*) there are many possible computational pathways that satisfy the folding process in the presence of the heuristics (*SI Appendix*, Fig. S6). In general, at high replica indices, the procedure explores a very broad range of extended states, whereas at lower replica-exchange indices, the structures become compact, resembling molten globule states. At the lowest replica indices, the protein is often native-like. A common theme in most pathways we have observed (except for some of the simpler proteins, such as TRP-cage) is that they will fold into intermediates that have certain characteristics of the native state but can have some secondary structure elements in incorrect orientations. These structures have to unfold, going back to higher replica indices, and then refold into native-like topologies (*SI Appendix*, Fig. S6).

Limitations of the Method. MELD + CPI is a sampling method. It cannot fix deficiencies in the force field. Although much faster sampling is accomplished, convergence can be an issue. The sequence and secondary structure predictions define the restraints; hence, for some proteins, they will be more directive (converge faster) than others.

The basic engine is classical MD; hence, there is no reactivity. If disulfide bonds are present in the native state but not specified in the simulations, they can never be formed. The lack of reactivity can limit the success of the method in some cases (*SI Appendix*, Table S4) due to steric clashes between reduced Cys that would not be present in the oxidized state. Disulfide bond information can be determined experimentally (26), greatly improving the results of the simulations.

Finally, not surprisingly, our "globular-protein" heuristics fail on proteins that are not globular. We tested the present heuristics on three nonglobular proteins (16). These proteins make fewer hydrophobic contacts than expected by our heuristics, forcing MELD to enforce incorrect restraints and ultimately leading to incorrect structures (*SI Appendix*, Table S5). These three proteins are helix bundles, so the only nonlocal heuristics in effect are the hydrophobic contacts. Our accuracy parameter for this heuristic is set at 8%, but looking at the native structures, we find that only 4%, 5%, and 6%, respectively, of the hydrophobic

contacts are satisfied in the native state. Not surprisingly, we were not able to identify native-like structures. For the native state, there is no combination of 8% of springs that have zero restraint energy. Hence, we no longer fall in the regime where comparing populations for the native state within MELD is comparable to comparing them with the original force field. Ultimately, there is no basis why MELD + CPI should work (*SI Appendix*, Table S5) in this case. Looking at this table, for one of the structures, we sampled native-like conformations. In this case, we do not expect that longer, more converged simulations will help, because there is a problem of matching the wrong heuristics to the wrong problem. Different definitions of heuristics to deal with nonglobular proteins would be needed in those cases.

Conclusions

In summary, MELD + CPI harnesses the desirable features of two approaches to protein structure prediction. Because it entails REMD simulations with atomistic force fields that satisfy detailed balance, it does not require specific template protein structures, samples the protein degrees of freedom extensively, uses transferrable physical potentials, computes populations rather than just structures, and will be useful where knowledge bases are limited. However, because it also uses external structural insights, it is much faster than MD. The power of MELD + CPI is that the information it uses is not exact and correct and specific but, rather, is vague, unreliable, and combinatoric, such as "having a hydrophobic core" or "having good secondary structures." In MELD + CPI, the CPI speeds up the MD and the MD "picks out" the native-like constraints. MELD + CPI is a practical application of the fact that protein folding is sped up by funnel-shaped landscapes. This method samples the native structures of 20 of 20 small proteins well, predicts the native structures for 15 of them well, does so much faster than unrestrained MD simulations (14), can be performed on laboratory-sized computing clusters, and appears promising for scaling to larger proteins.

Materials and Methods

This section provides an overview. Full details can be found in *SI Appendix*.

MD. We model the proteins in full atomistic detail, combined with the implicit-solvation model of Onufriev et al. (27). For the protein interactions, we used an in-house modified version of the AMBER12SB force field (28) that adds a correction map (29) term to reproduce the balance between α- and β-regions of the Ramachandran plot (the correction map is available at https://github.com/maccallumlab/meld). All our simulations are 500 ns long (per replica) unless otherwise noted. Initial conformations are fully extended as generated by the tleap (28) sequence command. We use the OpenMM suite of programs (30) with the MELD plug-in (11) with a 2-fs time step and Langevin dynamics.

REMD. For efficient conformational sampling, we use an H,T-REMD sampling approach with 30 replicas. The temperature ranges from 300 K in the lowest replica to 450 K in the highest, increasing geometrically. The heuristic restraints weaken at higher temperatures. At a low replica index, force constants are strong (250 $kJ \cdot mol \cdot nm^{-2}$) and at a high replica index, they are zero, changing exponentially from the lowest to highest replica. It is also important to point out that in MELD + CPI, we have used MD as a sampling method, but other methods that obey detailed balance [e.g., Monte Carlo (MC) or a hybrid MD/MC approach] could also be used for sampling.

Clustering into Representative Structures. At the ends of each simulation, we collect together the most similar structures into clusters, as is commonly done in structure predictions. We have used average-linkage clustering (12, 13) with a ε value of 2, which is standard (14, 15). As input for the clustering, we took the five lowest temperature replicas. We test the accuracy of clustering by computing the rmsd of the centroid to the native state. To avoid situations of loops and termini disrupting the clusters, the clustering is done on the Cα carbons of residues having predicted secondary structures. For the comparison with the native state, we consider the Cα of all residues, excluding flexible termini, as is standard in the field. *SI Appendix*, Table S1 contains a description of the residues used for each protein. We arbitrarily define a threshold in which structures closer than 4 Å to native are regarded as being within the native basin.

ACKNOWLEDGMENTS. We thank D. E. Shaw Research for making their folding simulations available to us and the Simmerling group for sharing their simulation data. We thank Sarina Bromberg for artistic help with the figures and Ivet Bahar for helpful comments and discussion. We are grateful for support from the Laufer Center, as well as support from NIH Grants GM107104 and 5R01GM090205 and from the Natural Sciences and Engineering Research Council (Canada). We appreciate the computing resources provided by The Extreme Science and Engineering Discovery Environment (XSEDE) Grant MCB120052.

1. Berman HM, et al. (2000) The Protein Data Bank. *Nucleic Acids Res* 28(1):235–242.
2. Moult J (2005) A decade of CASP: Progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol* 15(3):285–289.
3. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A (2014) Critical assessment of methods of protein structure prediction (CASP)–round x. *Proteins* 82(Suppl 2): 1–6.
4. Butterfoss GL, et al. (2012) De novo structure prediction and experimental characterization of folded peptoid oligomers. *Proc Natl Acad Sci USA* 109(36):14320–14325.
5. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234(3):779–815.
6. Eswar N, et al. (2006) Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics Chapter 5:Unit 5.6.
7. Simons KT, Bonneau R, Ruczinski I, Baker D (1999) Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* (Suppl 3):171–176.
8. Hirst SJ, Alexander N, McHaourab HS, Meiler J (2011) RosettaEPR: An integrated tool for protein structure determination from sparse EPR data. *J Struct Biol* 173(3): 506–514.
9. Shen Y, et al. (2008) Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci USA* 105(12):4685–4690.
10. Li X, Jacobson MP, Friesner RA (2004) High-resolution prediction of protein helix positions and orientations. *Proteins* 55(2):368–382.
11. MacCallum JL, Perez A, Dill KA (2015) Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc Natl Acad Sci USA* 112(22):6985–6990.
12. Roe DR, Cheatham TE, III (2013) PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J Chem Theory Comput* 9(7): 3084–3095.
13. Shao J, Tanner SW, Thompson N, Cheatham TE (2007) Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms. *J Chem Theory Comput* 3(6):2312–2334.
14. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE (2011) How fast-folding proteins fold. *Science* 334(6055):517–520.
15. Daura X, Gademann K, Jaun B (1999) Peptide folding: When simulation meets experiment. *Angew Chem Int Ed* 38(1/2):236–240.
16. Wu GA, Coutsias EA, Dill KA (2008) Iterative assembly of helical proteins by optimal hydrophobic packing. *Structure* 16(8):1257–1266.
17. Marks DS, et al. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 6(12):e28766.
18. Flory PJ (1942) Thermodynamics of high polymer solutions. *J Chem Phys* 10(1):51–61.
19. Huggins ML (1943) Thermodynamic properties of solutions of high polymers: The empirical constant in the activity equation. *Ann N Y Acad Sci* 44(4):431–443.
20. Beauchamp KA, et al. (2011) MSMBuilder2: Modeling Conformational Dynamics at the Picosecond to Millisecond Scale. *J Chem Theory Comput* 7(10):3412–3419.
21. Cronkite-Ratcliff B, Pande V (2013) MSMExplorer: Visualizing Markov state models for biomolecule folding simulations. *Bioinformatics* 29(7):950–952.
22. Nguyen H, Maier J, Huang H, Perrone V, Simmerling C (2014) Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. *J Am Chem Soc* 136(40):13959–13962.
23. Nguyen H, Roe DR, Simmerling C (2013) Improved Generalized Born Solvent Model Parameters for Protein Simulations. *J Chem Theory Comput* 9(4):2020–2034.
24. Voelz VA, Dill KA (2007) Exploring zipping and assembly as a protein folding principle. *Proteins* 66(4):877–888.
25. Capraro DT, Roy M, Onuchic JN, Jennings PA (2008) Backtracking on the folding landscape of the beta-trefoil protein interleukin-1beta? *Proc Natl Acad Sci USA* 105(39):14844–14848.
26. Wu J, Watson JT (1997) A novel methodology for assignment of disulfide bond pairings in proteins. *Protein Sci* 6(2):391–398.
27. Onufriev A, Bashford D, Case DA (2004) Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* 55(2):383–394.
28. Case DA, et al. (2012) Amber12 (University of California, San Francisco).
29. Mackerell AD, Jr, Feig M, Brooks CL, 3rd (2004) Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem* 25(11):1400–1415.
30. Eastman P, et al. (2013) OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J Chem Theory Comput* 9(1): 461–469.