Title

Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference

Permalink
https://escholarship.org/uc/item/9533k9p3

Journal
Proceedings of the National Academy of Sciences of the United States of America, 112(22)

ISSN
0027-8424

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

Publication Date
2015-06-02

DOI
10.1073/pnas.1506788112

Peer reviewed

Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference

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

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

Contributed by Ken A. Dill, April 13, 2015 (sent for review February 3, 2015; reviewed by Richard Lavery and Andrej Sali)

More than 100,000 protein structures are now known at atomic detail. However, far more are not yet known, particularly among large or complex proteins. Often, experimental information is only semireliable because it is uncertain, limited, or confusing in important ways. Some experiments give sparse information, some give ambiguous or nonspecific information, and others give uncertain information—where some is right, some is wrong, but we don't know which. We describe a method called Modeling Employing Limited Data (MELD) that can harness such problematic information in a physics-based, Bayesian framework for improved structure determination. We apply MELD to eight proteins of known structure for which such problematic structural data are available, including a sparse NMR dataset, two ambiguous EPR datasets, and four uncertain datasets taken from sequence evolution data. MELD gives excellent structures, indicating its promise for experimental biomolecule structure determination where only semireliable data are available.

protein structure | molecular modeling | integrative structural biology | Bayesian inference

Increasingly, structures are determined using integrative structural biology approaches, where direct experimental data are combined with computer-based models (1). Important successes in integrative structural biology have come from pioneering methods such as Modeler (2, 3), methods based on Rosetta (4–7), and others (8). Atomistic molecular dynamics (MD) simulations can be a powerful tool in integrative structural biology, because they capture physical principles and thermodynamic forces—information that is otherwise orthogonal to purely structural observations.

However, there remain many situations in which it is not yet possible to properly integrate external knowledge with atomistic MD to infer biomolecular structures. Often, the external knowledge is challenging in one or more of the following ways. (*i*) Sparse data provide too little information to fully constrain the structure. (*ii*) Ambiguous data are not very specific, allowing alternative structural interpretations. (*iii*) Uncertain data cannot be interpreted at face value, because they contain false-positive signals that can be misdirective. Determining new challenging protein structures requires ways to handle semireliable data.

Here, we describe a physics-based, Bayesian computational method called MELD (Modeling Employing Limited Data). It is a procedure for making rigorous inferences from limited or uncertain data. We build upon previous Bayesian approaches (9–14), which share the key feature of combining prior belief with the available data to produce statistically consistent samples from a posterior distribution, rather than searching for a single well-scoring model. The key properties of MELD are the rigorous treatment of statistical mechanics, a novel likelihood function that can handle uncertain data, and a graphics processing unit (GPU)-accelerated sampling strategy that makes the calculations tractable.

MELD uses free energy as the principle for choosing between different possible interpretations of the data and provides a rigorous, robust, statistical-mechanical approach for resolving sparsity, ambiguity, and uncertainty. Most importantly, MELD provides proper Boltzmann populations of states. Knowing the relative populations of different states is essential for inferring free energies, stabilities, rates, and biological mechanisms and is a key reason for using MD simulations over other types of conformational sampling.

Overview of MELD

Integrative Structural Biology Can Be Formulated As an Inference Problem. The task of integrative structural biology is to infer the ensemble of likely structures, $p(\mathbf{x}|\mathbf{D})$, given some data, $\mathbf{D}$. Here, $\mathbf{x}$ is the $3N$-dimensional vector of atomic coordinates. To solve this problem, we invoke Bayes' theorem:

$$\overbrace{p(\mathbf{x}|\mathbf{D})}^{\text{posterior}} = \frac{p(\mathbf{D}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{D})} \sim \overbrace{p(\mathbf{D}|\mathbf{x})}^{\text{likelihood}}\,\overbrace{p(\mathbf{x})}^{\text{prior}}. \qquad [1]$$

We must specify two things: (*i*) a prior probability distribution, $p(\mathbf{x})$, over the space of structures and (*ii*) a likelihood function, $p(\mathbf{D}|\mathbf{x})$, that predicts the likelihood of the data given a structure. The denominator, $p(\mathbf{D})$, is called the data likelihood and can be regarded as a normalization factor and ignored for our purposes.

The prior, $p(\mathbf{x})$, expresses our belief in the probability of observing each protein structure in the absence of any additional data. For MELD, we choose as our prior the Boltzmann distribution produced by a recent version of the Amber

Significance

Modeling Employing Limited Data (MELD) is a method of integrative structural biology. It serves to determine protein structures by a Bayesian approach combining physical models with experimental data that are only semireliable, by virtue of being either too sparse, too ambiguous, or too uncertain. For eight proteins for which both the correct native structure and semireliable data were available, MELD gives excellent structures.

Author contributions: J.L.M., A.P., and K.A.D. designed research; J.L.M. and A.P. performed research; J.L.M. and A.P. contributed new reagents/analytic tools; J.L.M. and A.P. analyzed data; and J.L.M., A.P., and K.A.D. wrote the paper.

Reviewers: R.L., Institut de Biologie et Chimie des Protéines; and A.S., University of California, San Francisco.

The authors declare no conflict of interest.

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

[2] To whom correspondence may be addressed. Email: justin.maccallum@ucalgary.ca or dill@laufercenter.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1506788112/-/DCSupplemental.

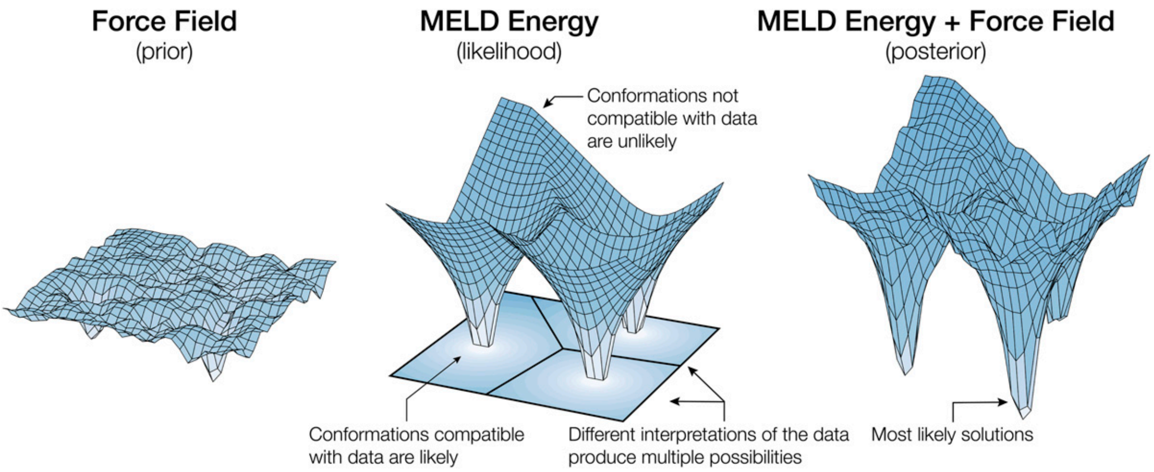
Force Field
(prior)
MELD Energy
(likelihood)
MELD Energy + Force Field
(posterior)
Conformations not compatible with data are unlikely
Conformations compatible with data are likely
Different interpretations of the data produce multiple possibilities
Most likely solutions

Fig. 1. MELD uses Bayesian inferences to predict structures from sparse, ambiguous, and uncertain data. Each panel shows the log-likelihood of a probability distribution in schematic form. The prior represents our belief about the probability of each structure and is given by the Amber force field. The likelihood models the probability that the observed data were given by each structure. Because our data are ambiguous and uncertain, there are multiple possible interpretations that can give rise to different basins of possible structures. We sample from the posterior, which includes information from both the prior and the likelihood.

forcefield (15) combined with a generalized-Born implicit solvation model (16):

$$p(\mathbf{x}) \sim \exp[-\beta E_{\text{amber}}(\mathbf{x})]. \quad \textbf{[2]}$$

In the absence of other information, MELD will produce samples from Eq. **2**, which may alone be enough to correctly predict the structures of small proteins (17–19). The likelihood function, $p(\mathbf{D}|\mathbf{x})$, expresses our belief about how probable it is that the observed data were produced by a particular structure (discussed below).

This Bayesian framework allows us to incorporate both our knowledge of the physics of protein structure and the external information in a rigorous and consistent way (Fig. 1). In MELD, we sample from the posterior using standard MD algorithms. We use the negative logarithm of Eq. **1** as our energy function, which amounts to sampling from

$$E_{\text{total}} = E_{\text{amber}}(\mathbf{x}) - \beta^{-1} \ln p(\mathbf{D}|\mathbf{x}). \quad \textbf{[3]}$$

The available external information is turned into nonnegative restraints on geometric degrees of freedom, such as distances or torsion angles, as in standard restrained MD (see *SI Appendix* for details). These restraints quantify how well a structure $\mathbf{x}$ agrees with a particular piece of data, D_i, with $E_i^{\text{rest}}(\mathbf{x}) = 0$ indicating that $\mathbf{x}$ is consistent with D_i. Rather than trying to infer the exact geometry from the data, we use restraints with broad, flat, zero-energy regions that encompass all structures consistent with D_i. Higher values of $E_i^{\text{restraint}}(\mathbf{x})$ indicate increasing disagreement between the structure and the external information, and thus indicate that $\mathbf{x}$ is exponentially unlikely to have produced D_i. Our likelihood function for D_i is

$$p(D_i|\mathbf{x}) \sim \exp\left[-\beta E_i^{\text{restraint}}(\mathbf{x})\right]. \quad \textbf{[4]}$$

So far, this approach is identical to standard restrained MD, but expressed in Bayesian terms. The key difference in MELD is the form of the likelihood function. Consider a set of 100 residue–residue contacts predicted from evolutionary considerations, where only about 65 of the predictions are expected to be correct. In standard restrained MD, one simply adds all of the energies, or alternatively multiplies all of the likelihoods:

$$p(\mathbf{D}|\mathbf{x}) \sim \prod_{i=1}^{100} \exp\left[-\beta E_i^{\text{restraint}}(\mathbf{x})\right]. \quad \textbf{[5]}$$

However, this causes a problem because the 35 incorrect restraints will bias the structure toward wrong regions of conformational space. Standard restrained-MD approaches assume that all of the data are correct, so those approaches may perform poorly when faced with ambiguous or uncertain datasets.

Instead, in MELD, we recognize that some of the restraints are wrong. We specify that 65 of the predicted contacts are correct and 35 should be treated as wrong.* Our likelihood function will bias this system toward conformations consistent with the correct information, while simultaneously ignoring the incorrect information. However, how do we decide which information is correct and which is not? For each structure $\mathbf{x}$, we make a local assumption: that the 35 restraints that are least well-satisfied (i.e., the 35 highest energy restraints) are likely to be incorrect and that the rest are correct. Our likelihood function is

$$p(\mathbf{D}|\mathbf{x}) \sim \prod_{i=1}^{65} \exp\left[-\beta E_i^{\text{restraint}}(\mathbf{x})\right], \quad \textbf{[6]}$$

where the restraints are sorted by energy,† so that

$$E_1^{\text{restraint}} \le E_2^{\text{restraint}} \le \ldots \le E_{100}^{\text{restraint}}. \quad \textbf{[7]}$$

By correctly sampling from the posterior distribution, $p(\mathbf{x}|\mathbf{D})$, MELD simultaneously infers the globally most likely configurations and the corresponding restraints that are globally most likely to be correct. MELD generates the minimum-free-energy ensemble from Eq. **3**, which provides the principle for selecting the correct interpretation from the sea of possibilities.

This per-timestep sorting and partitioning of restraints into active and ignored results in a multifunneled energy landscape (Fig. 1). The different funnels that MELD constructs on the landscape are because $p(\mathbf{D}|\mathbf{x})$ in Eq. **6** entails different restraints for different conformations. In practice, MELD uses a somewhat

*The number of restraints to be treated as correct must be specified as an input. We set this value empirically based on past experience with a particular source of data. Setting the number of correct restraints too high will force MELD to account for incorrect data, whereas setting it too low will throw away useful information. One could develop a Bayesian formulation that can simultaneously estimate the amount of correct information and the structural ensemble.

†Eqs. **6** and **7** lead to a continuous energy landscape, but with discontinuities in the force (cusps in Fig. 1), which result in occasional integration errors. In *SI Appendix*, we demonstrate that these occasional errors have no detectable effect on the conformational distributions produced by MELD.

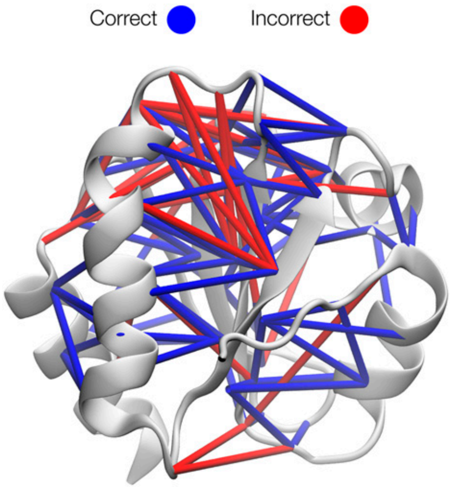

Fig. 2. Summary of input data for thioredoxin based on EvFold (31, 32) predicted contacts. The upper panel shows the EvFold predicted contacts overlaid on the crystal structure. Both correct (blue) and incorrect (red) contacts are shown. Secondary structure restraints are also enforced (see *SI Appendix*, Fig. S9).

more complex scheme with two layers of grouping, selection, and sorting, which allows for more sophisticated sculpting of the energy landscape (see *SI Appendix* for full details).

MELD is implemented using the GPU-accelerated OpenMM library (20). To avoid kinetic traps, MELD uses Hamiltonian and temperature replica exchange MD (H-REMD, see *Materials and Methods*) (21, 22). The top replicas have a high temperature and weak restraints, which allow the system to sample broadly. The bottom replicas are at room temperature with strong restraints, focusing sampling on low-free-energy regions. We identify the low-free-energy regions by clustering; we do not expect the energies in Eq. **3** to be directly useful for scoring, because they ignore entropy.

Input to a Typical MELD Calculation. The typical inputs to MELD are (*i*) the protein sequence, (*ii*) predicted secondary structures from PSIPRED (23), and (*iii*) externally supplied information about residue–residue contacts or distances; see Fig. 2 and *SI Appendix*, Figs. S5–S12.

We use three different classes of residue–residue distance information that we call sparse, ambiguous, or uncertain. Each class uses a different restraint strategy. Sparse restraints are few but are otherwise unambiguous and reliable specifications of short residue–residue distances. Ambiguous restraints provide residue–residue distance information over a wide range, being ambiguous about the exact distances, but are free from false positives. Uncertain restraints provide information about residue–residue contacts, but many of these restraints are false positives, so by our prior knowledge of their reliability we set them to be only 80% active, as outlined above. For all systems, secondary structures are set to be 75% active, based on the approximate accuracy of current secondary structure predictions.

Results

We applied MELD to eight proteins for which (*i*) the native structure is already known and (*ii*) sparse, ambiguous, or uncertain datasets were available (Fig. 3). Our goal was to learn whether MELD could produce correct structures from the semi-reliable data. For each system, we also compare the MELD result to a baseline calculation using the same input information, but performed with X-PLOR-NIH (24) (Fig. 4; see *SI Appendix* for details). X-PLOR-NIH is representative of methods typically used in structural biology that are designed to work with plentiful, largely unambiguous, and correct data.

How Efficiently Does MELD Target and Sample Near-Native Structures? Fig. 4 shows that in all eight cases MELD samples the important native-like states much better than X-PLOR does. In six cases, MELD produces structures that are within 2.5 Å of native, and the best structures for five of them are within 1.8 Å (Fig. 4). Of course, the computational demands for MELD are much greater, but GPU acceleration makes MELD tractable.

How Well Does MELD Choose the Correct Structures? To pick out native structures accurately requires not only good conformational-space coverage but also a good discrimination function that distinguishes more native-like from less native-like structures, such as having the lowest free energies. We clustered our conformations and chose the medoids of the three most populous clusters as representatives (Fig. 3). For six targets, MELD identifies a structure within 3.5 Å backbone rmsd from native and within 2.8 Å of native for five targets within our simulation times. *SI Appendix*, Table S1 reports the rmsds and populations of the top three clusters for each target.

MELD typically finds that one of the three most populous clusters is within ~1.0–1.5 Å rmsd of the best structure in the entire ensemble. This indicates that MELD conformational sampling is tightly focused with many conformations close to the native structure. In contrast, X-PLOR generates broad ensembles, having an average rmsd typically more than 8 Å from native.

Finding Structures Using Sparse Information from Solid-State NMR. An example of sparse information is the solid-state NMR data of Huber et al. (25) on ubiquitin. The experimental data provide information about short proton–proton interactions between terminal methyl groups on Ile, Leu, and Val (ILV) residues. The experiments measure 49 interactions, which we turn into 33 restraints between carbon atoms (Fig. 5*A*) by ignoring stereospecific proton assignments. This leaves ~0.4 restraints per residue, much less than the 10–20 restraints per residue that are typical in solution-NMR determination. The measured interactions help to approximately define the core of the protein, but there are several stretches longer than 10 aa each that are devoid of any experimental restraints (Fig. 5*A*).

Previously, Huber et al. (25) made two tests. They produced an ensemble of 200 structures using CYANA (26) coupled with assumed backbone ϕ/ψ angles [predicted from NMR chemical shifts using Talos+ (27)] and the measured ILV–methyl interactions. Even after selecting the 20 lowest-energy structures, the resulting ensemble is broad (*SI Appendix*, Fig. S1). Their second test included 139 more restraints from an additional experiment with a backbone-amide-labeled sample (25), which better defines the structure of β-sheets and improves the backbone rmsd to 1.6 Å.

Our aim here was to see whether MELD could produce correct structures even without the additional 139 restraints. We applied MELD to this structure-determination problem using the ILV-labeled (but not backbone-amide-labeled) dataset and replaced the detailed structural NMR chemical shift information with less accurate secondary structure predictions from PSIPRED (23). We found (*i*) that the centroid of our most populous cluster is 1.0 Å backbone rmsd from native, (*ii*) that structures as close as 0.6 Å were present in the ensemble, and (*iii*) that the core side-chain packing is also accurate, even in regions devoid of restraints (Fig. 5*B*).

The most populous cluster produced by MELD (with only the ILV–methyl data) was more accurate than even the best structure produced by CYANA (even with the additional amide-labeled data). The best structure produced by X-PLOR-NIH was poor (5.9 Å, Fig. 4).[‡] Without restraints, even state-of-the-art MD simulations on the special-purpose Anton supercomputer are

[‡] The results with CYANA are better than with X-PLOR. This may be due to differences in torsion angles; Huber et al. (25) used TALOS+ predictions based on NMR data, whereas our X-PLOR calculations used sequence-based predictions.

BIOPHYSICS AND COMPUTATIONAL BIOLOGY

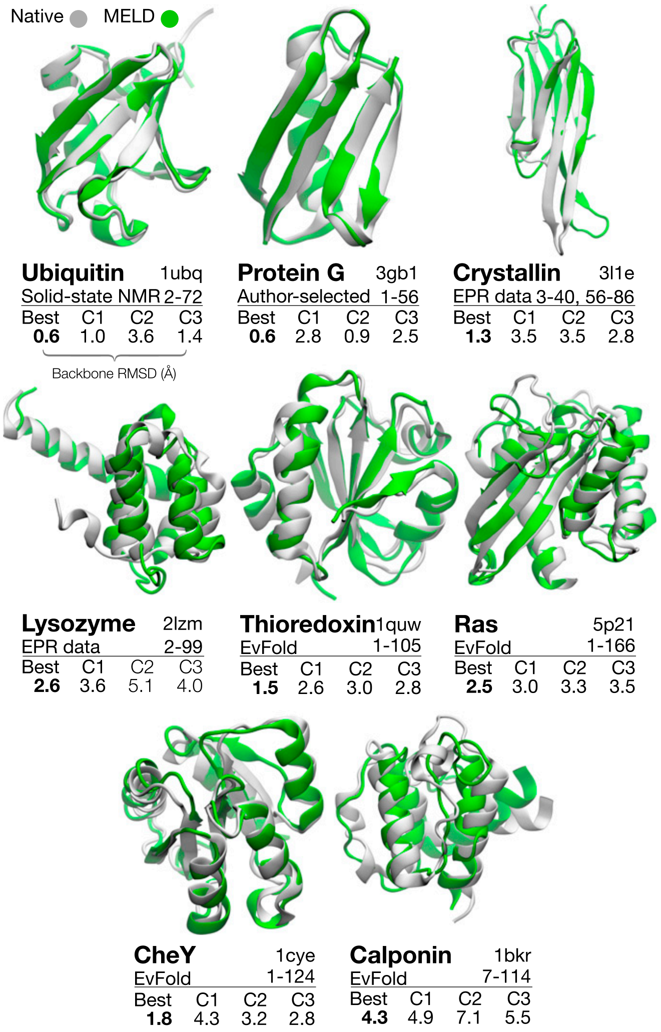

Fig. 3. Summary of MELD results for the eight case studies in this paper. Each panel shows the best (lowest backbone rmsd) structure (green) superposed on the native structure (gray) and reports the rmsd of the best structure and three most populous clusters (C1–C3).

not able to find the structure of ubiquitin starting from an extended chain, because ubiquitin folds only on the millisecond timescale (17). In MELD, however, even though the restraints are sparse, this set is sufficient to provide a highly funneled free-energy landscape (*SI Appendix*, Fig. S2) that leads to rapid folding.

The Experimental Data Alone Do Not Uniquely Define the Structure. Fig. 6 shows how the MELD restraint energy varies for the ubiquitin dataset (other systems are qualitatively similar; *SI Appendix*, Fig. S3). There is a broad cloud of structures within 5 Å that have MELD energies indistinguishable from the most native-like structures. More remarkably, there are structures that are 6–11 Å from native that also have very low MELD energies, with a particularly dense cluster around 9 Å. These are structures that are in good agreement with the experimental data but have nonnative folds. In the absence of a physical model, the experimental data cannot distinguish between these alternative structures, but MELD can by rigorously sampling from a well-defined distribution.

Even a Few Restraints Can Be Highly Informative. To illuminate the limits of maximum sparsity in MELD, we considered Protein G, where we combined secondary structure predictions with four restraints selected to help define the main β-sheet (*SI Appendix*, Fig. S6). The calculation was started from a completely extended conformation. Even this limited amount of sparse data enables MELD to find accurate structures (Fig. 3). Although these four restraints were hand-picked as a test of what is minimal, this result indicates that a few well-placed, highly informative restraints may be sufficient, in general, for structure determination with MELD.

Finding Structures Using Ambiguous Information from EPR. Inferring protein structure from site-directed spin-label EPR experiments is challenging because (*i*) there is inherent "fuzziness" in mapping the experimental signal into a distance between two spin-label probes, (*ii*) the probes are attached to the protein by flexible linkers, which introduces further uncertainty, and (*iii*) the distances measured are often large (20–40 Å) and contain less information than short (<10 Å) distances (28).

To obtain EPR restraints, we followed the two strategies previously used in ROSETTA-EPR (5, 29). We augment that distance information with secondary structure predictions from PSIPRED. We simulated two different systems (αA-Crystallin and Lysozyme) starting from completely extended conformations using available EPR data (29, 30).

The results for αA-Crystallin are good (Fig. 3), with a best overall structure of 1.3 Å and best cluster of 2.8 Å backbone rmsd from native. Deviations occur in the long hairpin (residues 40–56 of the modeled sequence), which extends into space as a monomer but forms a strand pair in the crystal lattice.

The results for T4-Lysozyme are not as good (Fig. 3). The best rmsd of all models is 2.6 Å and the best cluster is 3.6 Å from native. This is still enough to define the overall fold of the protein correctly, but many details are wrong. To assess the source of this error, we performed two additional calculations starting from the native structure: one without restraints and one with the full set of EPR and secondary structure restraints. In the simulations without restraints, in 200 ns the rmsd rises rapidly to >7 Å while the protein unfolds. In the simulations with restraints, the rmsd stabilizes at ∼4 Å, similar to our result starting from an extended chain. There are three possibilities: (*i*) We simulated a truncated protein that is missing the N-terminal β-domain and the loss of this domain could make the protein unstable, (*ii*) there are systematic errors in the force field, or (*iii*) our treatment of the experimental data has errors.

The results using X-PLOR are poor (Fig. 4), and none of the generated models is within 6 Å of the native structure for either protein. Using ROSETTA-EPR, Meiler and coworkers (5) folded T4-Lysozyme to 1.8Å and αA-Crystallin to 4.0Å (29) Cα rmsd from native. Overall, MELD significantly outperforms the simple X-PLOR approach and has performance comparable to that of ROSETTA-EPR, but with the important advantage of

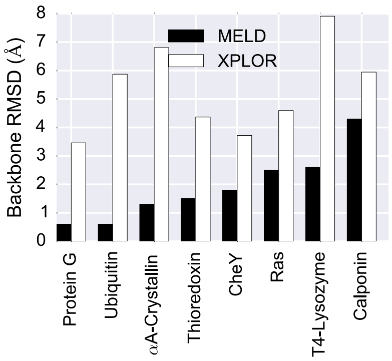

Fig. 4. MELD samples more accurate structures than X-PLOR-NIH for all test cases in this study. Each bar represents the single best structure produced for that target by each method.

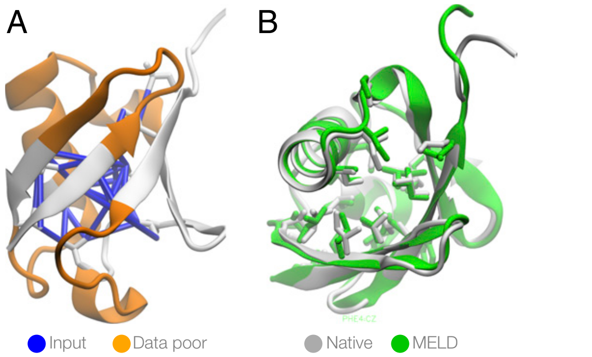

Fig. 5. Structure determination of ubiquitin using MELD with sparse solid-state NMR data and Talos+ secondary structure predictions. (*A*) The input restraints overlaid on the crystal structure. Data-poor regions longer than 10 residues are shown in orange. (*B*) Overlay of native and MELD prediction showing the remarkable agreement in the prediction of side-chain conformations.

coming from a fully physical model capable of giving Boltzmann populations.

Finding Structures Using Uncertain Information from Evolutionarily Inferred Contacts. EvFold (31, 32) belongs to a new family of methods (33, 34) that predict residue–residue contacts from coevolution in multiple sequence alignments.

Here, we wanted to see whether such probabilistic restraints, when applied using MELD, could accurately predict protein structures. We took restraints from EvFold (31) for four targets: thioredoxin, Ras, CheY, and calponin. For each target, we combined the top N_{res} contacts predicted by EvFold with secondary structure predictions from PSIPRED. For these systems, rather than starting from an extended conformation, we seeded each replica with one of the structures produced by the EvFold pipeline, because that information is readily available.[§]

For all systems, the best structures sampled by MELD are more accurate than those from X-PLOR (Fig. 4). Except for calponin, MELD samples best structures that are accurate (<2 Å) and gives accurate (<2.8 Å) models in the three most populous clusters (Fig. 3). *SI Appendix*, Table S2 shows that for all four systems MELD produces better structures than EvFold's structure generation procedure, which is based on CNS [Crystallography and NMR System (35)]. The most populous cluster MELD identifies is more accurate than even the best conformation (lowest rmsd) that EvFold samples, which might not be identifiable by score or energy. The average improvement of the most populous cluster from MELD over the lowest-energy structure from EvFold is 2.5 Å.

Enforcing Incorrect Restraints Reduces Accuracy. Incorporating EvFold data into MELD requires specifying the fraction of active contacts and the cutoff distance defining a contact (see *SI Appendix* for details). In the present study, we chose 0.8 and 6 Å, based on previous EvFold results (31). With these parameters, the MELD energies of the native structures are high (*SI Appendix*, Fig. S3). Excluding calponin, these results are indicative of a small number of modest (1–2 Å) restraint violations (*SI Appendix*, Fig. S4). The energy of the native structure for these systems could be reduced to near zero by a small reduction in the active fraction to 0.65–0.70, or by a small increase in cutoff distance to 7–8 Å. The structural quality for these systems is good, suggesting that these parameters do not have substantial negative impact.

The results for calponin are not as good, where even the best sampled structure is more than 4 Å from native. For this system, the EvFold predictions are less accurate, resulting in higher energies (*SI Appendix*, Fig. S4). To achieve a MELD energy near zero, the active fraction would need to be reduced to 0.4–0.5, or the cutoff distance increased to 10–11 Å. In this case, it seems that the suboptimal parameters may have led to poor structures. However, unrestrained simulations starting from native indicate that there may also be systematic force-field errors. Although it is possible that results could be improved by parameters tuning, this is not our aim here.

One obvious current limitation of MELD is the need to specify parameters such as the active fraction and the cutoff distance. We are developing a more general approach that places priors on these parameters and then treats them as parameters to be inferred jointly with the structural distribution. Similar approaches have been successfully used with reliable datasets, for example to infer correct parameters for converting NMR crosspeak intensities into distances (9, 12).

Conclusions

The challenge in structural biology is to determine ever larger and more complex structures. It requires making ever better use of diverse, ambiguous, and confusing experimental data. At the same time, the power of molecular simulations in this enterprise is in filling in the fine-grained detail in space and time, and in going beyond structures, to inform us also about populations, stabilities, kinetics, motions, and mechanisms. Molecular simulations alone, however, are challenged by simulation errors (in sampling and force fields) that increase with the number of degrees of freedom. In MELD, we combine the advantages of simulations and imperfect data. MELD draws Bayesian inferences from semireliable data in the context of atomistic REMD computer simulations, to give accurate protein structures.

In a way, MELD follows from an old line of thought that if we knew the physical mechanisms for how proteins fold so fast, we could invent fast ways to search their conformational spaces to

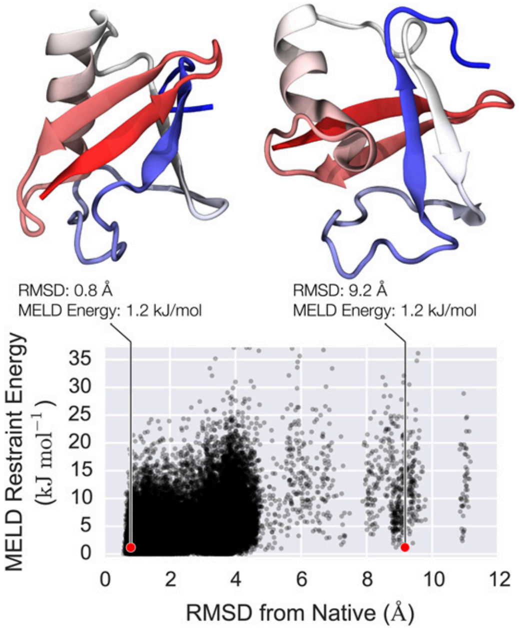

Fig. 6. The available solid-state NMR data do not uniquely define the structure of ubiquitin. Structures as far as 11 Å rmsd from native have MELD energies that are comparable to native. These alternative structures may even have different folds (upper panels) but are nevertheless in good agreement with the available sparse experimental data.

[§] We would expect similar results if we started from an extended conformation, because the EvFold pipeline relies on restrained MD using CNS (35), although we have not performed these calculations.

find native like states. From the early days, there has been interest in parlaying knowledge of kinetic routes of protein folding to an understanding of principles of protein structures (36, 37). Since then, it has become clear that a unifying principle is that folding energy landscapes are funnel-shaped (30, 38, 39). Here, MELD establishes funnel-shaped potentials, driven by uncertain—but nevertheless powerful—information from experiments or bioinformatics. We find that these funnels greatly accelerate the identification of the native structure.

Materials and Methods

This section provides an overview. Full details can be found in *SI Appendix*. The MELD source code is freely available at https://github.com/maccallumlab/meld.

Simulation Software and Parameters. All simulations were performed using a modified version of OpenMM (20). Proteins were modeled using a version of the ff12sb force field (15) that included a CMAP-like (40) correction to better reproduce the balance between α and β secondary structures (parameters are included in the github repository). Solvation was modeled using the OBC model (16). X-PLOR was used as a baseline for comparison; a sample script can be found in *SI Appendix*.

Turning Information into Restraints. Secondary structure predictions from PSIPRED (23) were turned into restraints acting on overlapping 5-mer fragments (*SI Appendix*). These restraints include both backbone torsion angle and intrachain distance restraints, which we found critical for reproducing secondary structure. We specified that 75% of these compound 5-mer restraints be active. The residue–residue distance restraints were handled differently depending on the source of the input data (see *SI Appendix* for details). For the sparse and ambiguous cases, all of the distance restraints were active. For the uncertain cases, 80% of restraints were active.

Hamiltonian and Temperature Replica Exchange. The number of replicas was adjusted, based on the size of the system and available computational resources, to between 24 and 48 replicas. Exchanges were performed every 20 ps or every 50 ps. For the non-EvFold systems, the temperature was varied geometrically from 300 K at the bottom replica to 450 K at the middle replica and held constant over the top half of the REMD ladder. Conversely, the strength of imposed distance restraints was varied from zero at the top of the ladder, to full strength at the middle replica, and held at full strength over the bottom half of the REMD ladder. We found that this arrangement improved the number of folding trajectories. The secondary structure restraints were held at full strength throughout. For the EvFold systems, we varied both the temperature and strength of restraints across the full REMD ladder.

Clustering. Trajectories were clustered based on Cα coordinates, using average-linkage clustering with $\epsilon = 4$ Å (41). The centroids of the three most populous clusters were chosen as representative structures.

ACKNOWLEDGMENTS. We thank B. Meier for providing the raw ensembles produced by CYANA in ref. 25, S. M. Islam and B. Roux for providing the EPR data in ref. 30, and D. S. Marks and C. Sander for providing raw data from EVfold. J.L.M. thanks P. K. Eastman, V. S. Pande, and the SimTK organization for supporting the development of MELD in OpenMM. The authors acknowledge the generous support of the Laufer family. This work was supported by Extreme Science and Engineering Discovery Environment MCB120052 and NIH Grants 5R01GM090205 and 1R01GM10710401.

1. Ward AB, Sali A, Wilson IA (2013) Biochemistry. Integrative structural biology. *Science* 339(6122):913–915.
2. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234(3):779–815.
3. Eswar N, et al. (2006) *Comparative Protein Structure Modeling Using Modeller* (Wiley, Hoboken, NJ).
4. 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.
5. 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.
6. Shen Y, et al. (2008) Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci USA* 105(12):4685–4690.
7. Dominguez C, Boelens R, Bonvin AMJJ (2003) HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* 125(7):1731–1737.
8. Li W, Zhang Y, Skolnick J (2004) Application of sparse NMR restraints to large-scale protein structure prediction. *Biophys J* 87(2):1241–1248.
9. Rieping W, Habeck M, Nilges M (2005) Inferential structure determination. *Science* 309(5732):303–306.
10. Habeck M, Nilges M, Rieping W (2005) Replica-exchange Monte Carlo scheme for bayesian data analysis. *Phys Rev Lett* 94(1):018105.
11. Habeck M, Nilges M, Rieping W (2005) Bayesian inference applied to macromolecular structure determination. *Phys Rev E Stat Nonlin Soft Matter Phys* 72(3 Pt 1):031912–031919.
12. Rieping W, Nilges M, Habeck M (2008) ISD: A software package for Bayesian NMR structure calculation. *Bioinformatics* 24(8):1104–1105.
13. Fisher CK, Huang A, Stultz CM (2010) Modeling intrinsically disordered proteins with bayesian statistics. *J Am Chem Soc* 132(42):14919–14927.
14. Voelz VA, Zhou G (2014) Bayesian inference of conformational state populations from computational models and sparse experimental observables. *J Comput Chem* 35(30):2215–2224.
15. Case DA, et al. (2015) Amber 2015 (University of California, San Francisco).
16. 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.
17. Piana S, Lindorff-Larsen K, Shaw DE (2013) Atomic-level description of ubiquitin folding. *Proc Natl Acad Sci USA* 110(15):5915–5920.
18. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE (2011) How fast-folding proteins fold. *Science* 334(6055):517–520.
19. 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.
20. 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.
21. Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett* 314(1-2):141–151.
22. Fukunishi H, Watanabe O, Takada S (2002) On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J Chem Phys* 116(20):9058.
23. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 292(2):195–202.
24. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. *J Magn Reson* 160(1):65–73.
25. Huber M, et al. (2011) A proton-detected 4D solid-state NMR experiment for protein structure determination. *ChemPhysChem* 12(5):915–918.
26. Güntert P, Mumenthaler C, Wüthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J Mol Biol* 273(1):283–298.
27. Shen Y, Delaglio F, Cornilescu G, Bax A (2009) TALOS+: A hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J Biomol NMR* 44(4):213–223.
28. Havel TF, Crippen GM, Kuntz ID (1979) Effects of distance constraints on macromolecular conformation. II. Simulation of experimental results and theoretical predictions. *Biopolymers* 18(1):73–81.
29. Alexander N, Bortolus M, Al-Mestarihi A, Mchaourab H, Meiler J (2008) De novo high-resolution protein structure determination from sparse spin-labeling EPR data. *Structure* 16(2):181–195.
30. Islam SM, Stein RA, McHaourab HS, Roux B (2013) Structural refinement from restrained-ensemble simulations based on EPR/DEER data: Application to T4 lysozyme. *J Phys Chem B* 117(17):4740–4754.
31. Marks DS, et al. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* 6(12):e28766.
32. Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. *Nat Biotechnol* 30(11):1072–1080.
33. Jones DT, Buchan DWA, Cozzetto D, Pontil M (2012) PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28(2):184–190.
34. Kamisetty H, Ovchinnikov S, Baker D (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci USA* 110(39):15674–15679.
35. Brunger AT (2007) Version 1.2 of the Crystallography and NMR system. *Nat Protoc* 2(11):2728–2733.
36. Kim PS, Baldwin RL (1982) Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annu Rev Biochem* 51(1):459–489.
37. Kim PS, Baldwin RL (1990) Intermediates in the folding reactions of small proteins. *Annu Rev Biochem* 59(1):631–660.
38. Dill KA, Chan HS (1997) From Levinthal to pathways to funnels. *Nat Struct Biol* 4(1):10–19.
39. Dill KA, MacCallum JL (2012) The protein-folding problem, 50 years on. *Science* 338(6110):1042–1046.
40. Best RB, et al. (2012) Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain $\chi(1)$ and $\chi(2)$ dihedral angles. *J Chem Theory Comput* 8(9):3257–3273.
41. Shao J, Tanner SW, Thompson N, Cheatham TEI (2007) Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms. *J Chem Theory Comput* 3(6):2312–2334.