JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

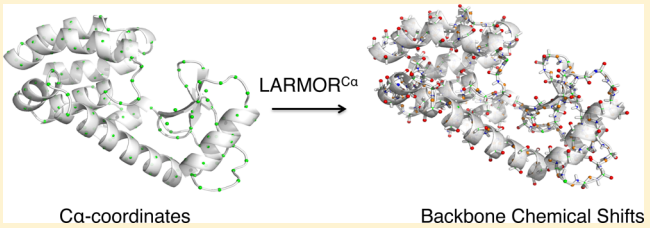
Predicting Protein Backbone Chemical Shifts From Cα Coordinates: Extracting High Resolution Experimental Observables from Low Resolution Models

Aaron T. Frank,*[,†,‡] Sean M. Law,[†,‡] Logan S. Ahlstrom,[†,‡] and Charles L. Brooks, III*[,†,‡]

[†]Department of Chemistry and [‡]Biophysics Program, University of Michigan, Ann Arbor, Michigan 48109, United States

Supporting Information

ABSTRACT: Given the demonstrated utility of coarse-grained modeling and simulations approaches in studying protein structure and dynamics, developing methods that allow experimental observables to be *directly* recovered from coarse-grained models is of great importance. In this work, we develop one such method that enables protein backbone chemical shifts (^{1}HN, ^{1}Hα, ^{13}Cα, ^{13}C, ^{13}Cβ, and ^{15}N) to be predicted from Cα coordinates. We show that our Cα-based method, LARMOR$^{C\alpha}$, predicts backbone chemical shifts with comparable accuracy to some all-atom approaches. More importantly, we demonstrate that LARMOR$^{C\alpha}$ predicted chemical shifts are able to resolve native structure from decoy pools that contain both native and non-native models, and so it is sensitive to protein structure. As an application, we use LARMOR$^{C\alpha}$ to characterize the transient state of the fast-folding protein gpW using recently published NMR relaxation dispersion derived backbone chemical shifts. The model we obtain is consistent with the previously proposed model based on independent analysis of the chemical shift dispersion pattern of the transient state. We anticipate that LARMOR$^{C\alpha}$ will find utility as a tool that enables important protein conformational substates to be identified by "parsing" trajectories and ensembles generated using coarse-grained modeling and simulations.



INTRODUCTION

Characterizing the folding/unfolding pathway of proteins remains an outstanding and significant challenge in structural biology. Though the emphasis has been on characterizing the native state structure of proteins, new experimental techniques are now being developed that enable transiently populated intermediates along the protein folding pathway to be characterized at the atomic scale.[1−6] Identifying such folding intermediates has always been viewed as an important task, but now as these and other non-native states have been implicated in several diseases,[7] developing approaches that enable the "complete" folding pathway to be characterized is of even greater importance.

In principle, classical molecular dynamics (MD) simulations, which can generate full atomic trajectories of a protein by propagating Newton's equations of motions, can be used to characterize its folding pathway(s). However, rigorous MD simulations are computationally expensive, making it difficult to simulate protein folding; typical simulations are on the order of nanosecond to microseconds, whereas proteins (with the exception of fast-folders) fold on a time scale of milliseconds and beyond. Though recent advances in computer hardware, software, and methodology[8−11] now allow the long time scale dynamics of some proteins to be studied,[12] these approaches still require significant computational resources. Thus, there remains a keen need for approaches that allow the folding pathway(s) of proteins to be characterized using readily available computational resources.

Coarse-grained molecular simulations, in which the full atomic system is reduced to a smaller less complex system of interacting "coarse-grained" particles, have been used to overcome the "mismatch" between simulation and biological time scales by sacrificing resolution for enhanced sampling efficiency. Remarkably, despite their simplicity, coarse-grained modeling and simulation approaches have been used to provide significant insights into protein functioning.[13−16]

Of considerable interest is the use of coarse-graining within a "multiscale" approach, in which coarse-grained simulations are used to rapidly and exhaustively sample the conformational space of a target protein, and then "selected" conformers from the coarse-grained simulations are used to "seed" more rigorous all-atom simulations. One approach to identifying relevant "seed" conformations is to use advanced clustering[17−19] and other data-reduction techniques.[20] Alternatively, experimental data can be used to select relevant "seed" structures by constructing ensembles that are consistent with the ensemble averaged data.[21−25] A prerequisite for such an "experimentally-augmented" identification of relevant conformational substates is the ability to calculate experimental observables from structural models, and, in the context of coarse-grained modeling, this typically requires mapping reduced models back to their all-atom representations.[26−31] Unfortunately, in addition to suffering from issues of nonuniqueness, this

Received: October 14, 2014
Published: December 8, 2014

mapping incurs an additional computational cost; typically coarse-grained approaches generate on the order 10^6 conformers, so this additional cost can be significant. Techniques are therefore needed that maximize the structural information that can be directly extracted from coarse-grained models and thus obviate the need for all-atom reconstruction of an entire trajectory or ensemble generated using coarse-grained simulations.

NMR relaxation dispersion (NMR-RD) experiments have recently garnered significant attention because they allow NMR observables of low-populated intermediates to be detected. These observables can be then used to "unveil" the structure of these previously "invisible" states. Using such an approach, NMR-RD derived chemical shifts, which are exquisitely sensitive to protein structure, have now been used to structurally characterize the folding intermediates of several proteins.[32−34] Incorporating NMR-RD derived chemical shifts into the analysis of coarse-grained simulations would allow relevant intermediate states that are sampled along the folding/unfolding pathways to be identified. As a first step toward being able to use NMR chemical shifts to "parse" trajectories or ensembles generated using coarse-grained modeling and simulations, we introduce LARMOR$^{C\alpha}$, a prediction method that allows protein backbone (^{1}HN, ^{1}Hα, ^{13}Cα, ^{13}C, ^{13}Cβ, and ^{15}N) chemical shifts to be predicted based *only* on Cα-based atomic coordinates. In what follows we (*i*) describe the model used to generate LARMOR$^{C\alpha}$; (*ii*) assess the accuracy of LARMOR$^{C\alpha}$; (*iii*) assess the sensitivity of LARMOR$^{C\alpha}$ predicted chemical shifts to protein structure; and (*iv*) use coarse-grained simulations and LARMOR$^{C\alpha}$ to characterize the transient state of the gpW, a small fast-folding protein.

METHODS

Training and Testing Set. LARMOR$^{C\alpha}$ predictors were trained and tested using the RefDB database.[35] Briefly, the data set contains proteins for which both high-resolution X-ray structures and NMR chemical shifts are available in the Protein Data Bank (PDB: http://www.pdb.org) and Biological Magnetic Resonance Bank (BMRB: http://www.bmrb.wisc.edu/), respectively. The training and testing set used here consisted of 196 and 61 proteins, respectively (Table S1).

Cα-Cα Distance-Based Structure Features. The distance-based structural features used to predict backbone chemical shifts from Cα coordinates were identical to those recently used by the program PCASSO to assign protein secondary structure from Cα coordinates.[36] Briefly, for a given residue, *i*, a set of 43 features are calculated from the Cα coordinates and the pseudocenter coordinates, respectively (see Table S2 and also ref 36 for a list of all features). The pseudocenter for the i^{th} residue is defined as the center-of-geometry between C$\alpha(i)$ and C$\alpha(i+1)$. The feature vector, $V(i)$, for the i^{th} residue is made up by features from the i^{th}, $i-1$th, and $i+1$th residues which results in a total of 258 feature elements (Table S2).

Generating LARMOR$^{C\alpha}$ Using Randomized Decision Trees. For all proteins in the training set, the Cα-Cα distance-based structural features described above were extracted from the X-ray structures and combined with their corresponding chemical shift data. The resulting data set was used as input to build a set of models to separately predict ^{1}HN, ^{1}Hα, ^{13}Cα, ^{13}C, ^{13}Cβ, or ^{15}N backbone chemical shifts. Specifically, for each backbone nucleus type, the random forest machine learning technique, implemented in the RandomForest module in the Scikit-learn Python package,[37] was used to build a predictive model. Each random forest predictor consisted of 50 randomized decision trees, and the maximum depth was set to 25. Each node in a given tree was split using the best splitting variable from a subset of 16 randomly chosen feature variables. The minimum number of samples required for splitting an internal node and the minimum number of samples required in a leaf node were both set to 5. Default values were used for all other parameters.

Molecular Dynamics (MD) Simulations. Coarse-grained decoy pools were generated for four arbitrarily chosen proteins in the testing set (PDB ID 1LM4: chain B,[38] 2C5L: chain C,[39] 1DYT: chain A[40] and 1H4A: chain X[41]) using G$\bar{o}$ model MD simulations. The native contacts used to define the G$\bar{o}$ models were derived from the coordinates in the corresponding PDBs using the MMTSB G$\bar{o}$ Model Server (https://mmtsb.org/webservices/gomodel.html).[42,43] All simulations were carried out using CHARMM MD simulation package.[44] Trajectories of 7.5 μs each were propagated using Langevin dynamics at 300 K with a friction coefficient of 0.10 ps^{-1}. All bonds were constrained using SHAKE,[45] and nonbond interactions were truncated at 25 Å with a smooth switching function between 21 and 23 Å. G$\bar{o}$ model MD simulations of the gpW protein were carried out using the same procedure described above. In this case, the native contacts used to define the G$\bar{o}$ model were derived from the gpW NMR structure (PDBID: 2L6Q; model 1).[46,47] For all the proteins, simulations at 300 K enabled both native and non-native conformations to be sampled.

Chemical Shift Analysis. For each of the four test systems, backbone chemical shifts were predicted from the resulting trajectory using LARMOR$^{C\alpha}$, and then the weighted-root-mean-squared-error (*w*RMSE) between predicted and measured chemical shifts along each trajectory were calculated. The *w*RMSE is given as

$$w\mathrm{RMSE} = \sqrt{\frac{1}{N_{CS}} \sum_{N_{CS}}^{i=1} w_i^2 (\delta_i^{\mathrm{pred}} - \delta_i^{\mathrm{meas}})^2} \quad (1)$$

where δ_i^{pred}, δ_i^{meas}, and w_i are the predicted chemical shift, the measured chemical shift, and weighting factor, respectively, for a given nucleus, *i*. The summation runs over the set of N_{CS} chemical shifts. The weighting factors (w_i) were used to account for the differential accuracy of the predictors. Specifically,

$$w_i = \frac{R_i}{\mathrm{MAE}_i} \quad (2)$$

where R_i and MAE_i are the estimated Pearson correlation coefficient and estimated mean-absolute-error, respectively, between the measured and LARMOR$^{C\alpha}$ predicted chemical shifts for the nucleus type associated with *i*. The weighting factor also scales the contribution to the overall error such that nuclei with different dispersion ranges can contribute equally to the *w*RMSE.

In addition to extracting the model with the lowest *w*RMSE and then comparing to the native structure, for each of the four systems receiver-operator-characteristic (ROC) analysis was carried out. First, the fraction of native contacts, *Q*, was calculated for each conformer along each trajectory. Conformers along the trajectories were then designated as native if $Q > 0.90$ and non-native otherwise. ROC curves were plotted

dx.doi.org/10.1021/ct5009125 | J. Chem. Theory Comput. 2015, 11, 325−331

for each test case, and the area-under-curve (AUC) was determined. Here the AUC, which ranges between 0 and 1, was used as a measure of the resolving power of the LARMOR$^{C\alpha}$ predicted chemical shifts. An AUC approaching 1 indicated that the models with the lowest error *w*RMSE corresponded to the native conformer and thus the *w*RMSE was effective at resolving native and non-native conformers, whereas an AUC of 0.5 indicated that the use of *w*RMSE to distinguish native from non-native conformers was no better than a random designation. In addition to carrying out ROC analysis using the total *w*RMSE, parallel analyses were carried out using only ^{1}HN, ^{1}Hα, ^{13}Cα, ^{13}C, ^{13}Cβ, or ^{15}N chemical shifts.

For the gpW protein, the *w*RMSE between LARMOR$^{C\alpha}$ predicted chemical shifts and measured chemical shifts corresponding to (*i*) the native states and (*ii*) the transient state were determined.[46] The conformation closest to the average structure of the 10 models exhibiting the lowest *w*RMSE was then extracted and considered representative of the state.

RESULTS AND DISCUSSION

The prospect of predicting backbone chemical shifts directly from Cα atomic coordinates opens up the possibility of utilizing chemical shifts to parse trajectories of Cα-based coarse-grained simulations and so identify intermediate states along the folding pathway of proteins. However, relying *only* on Cα atomic coordinates reduces the information content in the models and thus places an inherent limit on how accurately chemical shifts can be predicted. In what follows, we first examine the accuracy with which LARMOR$^{C\alpha}$ predicts backbone chemical shifts and then compare it to all-atom prediction methods. We then assess whether, given its current accuracy, LARMOR$^{C\alpha}$ predicted backbone chemical shifts are likely to be of utility in resolving protein structure. The latter is essential because it is sensitivity to protein structure rather than absolute prediction accuracy that will be most important when utilizing chemical shifts to study the folding pathway of proteins.

LARMOR$^{C\alpha}$ Backbone Chemical Shift Prediction Accuracy Is Comparable to Some All-Atom-Based Approaches. We began our analysis by determining the accuracy with which LARMOR$^{C\alpha}$ predicts protein backbone chemical shifts for the proteins in the testing set (Table S1). The accuracy of the predictions for ^{1}HN, ^{1}Hα, ^{13}Cα, ^{13}C, ^{13}Cβ, and ^{15}N nuclei were quantified by computing the root-mean-square-error (RMSE), mean-absolute-error (MAE), and the Pearson correlation coefficient (*R*) between LARMOR$^{C\alpha}$ predicted chemical shifts and measured chemical shifts. The results are summarized in Table 1. For ^{1}HN, ^{1}Hα, ^{13}Cα, ^{13}C, ^{13}Cβ, and ^{15}N the RMSE, MAE and *R*, calculated over all corresponding chemical shifts in the testing set, were 0.54, 0.35, 1.21, 1.79, 4.18, and 3.32 ppm, 0.39, 0.25, 0.90, 1.03, 1.55, and 2.45 ppm, and 0.67, 0.77, 0.82, 0.93, 0.94, and 0.81, respectively.

The large discrepancy between the RMSE and MAE is indicative of the presence of a small set of large prediction outliers. To confirm this, outlier analysis was carried out for each backbone nucleus. Specifically, we identified possible prediction outliers using the 3-sigma rule, i.e. a prediction outlier was identified as one that had an error that exceeded the median error by more than three standard deviations. When excluding the prediction outliers—on average ~4.0% of the total testing set for each nucleus—the RMSE and MAE decreased to 0.44, 0.28, 1.06, 0.99, 1.06, and 2.88 ppm and 0.35, 0.22, 0.83, 0.76, 0.81, and 2.25 ppm and the *R* increased to 0.75, 0.83, 0.86, 0.98, 0.99, and 0.85, for ^{1}HN, ^{1}Hα, ^{13}Cα, ^{13}C, ^{13}Cβ, and ^{15}N nucleus, respectively.

Table 1. Backbone Chemical Shifts Prediction Accuracy[a]

nucleus	RMSE (ppm)	MAE (ppm)	*R*	no. of shifts	% prediction outliers
Hα	0.54 (0.44)	0.39 (0.35)	0.67 (0.75)	5776	2.98
HN	0.35 (0.28)	0.25 (0.22)	0.77 (0.83)	8346	3.34
Cα	1.21 (1.06)	0.90 (0.83)	0.82 (0.86)	8856	2.31
C	1.79 (0.99)	1.03 (0.76)	0.93 (0.98)	7218	5.96
Cβ	4.18 (1.06)	1.55 (0.81)	0.94 (1.00)	7322	7.39
N	3.32 (2.88)	2.45 (2.25)	0.81 (0.85)	8125	2.43

[a]The root-mean-square-error (RMSE), mean-absolute-error (MAE), and Pearson correlation coefficient (*R*) between LARMOR$^{C\alpha}$-predicted and experimental chemical shifts. For each nucleus, the RMSE, MAE, and *R* statistics were calculated using all chemical shifts in the testing set. In parentheses are the statistics obtained when prediction outliers are excluded. Also listed for each nucleus is the number of chemical shifts over which the statistics were computed and the percentage of prediction outliers identified. A prediction outlier is identified as having an error that exceeds the median error by more than three standard deviations (i.e., the 3-sigma rule).

Consistent with our expectation, backbone chemical shifts predicted using LARMOR$^{C\alpha}$ were generally less accurate than those calculated using all-atom methods. For example, SHIFX2[48] and SPARTA+,[49] which are currently the "gold-standard" for empirical structure-based protein chemical shift prediction, exhibited significantly lower RMSE over the testing set (Table S3) and had mean *R*s of 0.98 and 0.92, respectively, compared with 0.82 for LARMOR$^{C\alpha}$ (Table S4). A similar picture emerges when comparing LARMOR$^{C\alpha}$ to CamShift;[50] the mean *R* for CamShift was 0.89 (Tables S3 and S4). However, LARMOR$^{C\alpha}$ predicts backbone chemical shifts with an accuracy comparable to PROSHIFT[51] and SHIFTS;[52] the mean *R* for PROSHIFT and SHIFTS were 0.86 and 0.81, respectively, compared to LARMOR$^{C\alpha}$'s 0.82. When prediction outliers were accounted for, the overall accuracy of LARMOR$^{C\alpha}$ prediction accuracy was on par with Camshift (Tables S3 and S4). Together these results show that although LARMOR$^{C\alpha}$ generally predicts backbone chemical shifts less accurately than all-atom methods, with the exception of SHIFTX2 and SPARTA+, the drop off in accuracy is not too severe, this despite predicting backbone chemical shifts based *only* on Cα coordinates.

Sensitivity to Structure Allows LARMOR$^{C\alpha}$ To Distinguish Native and Non-Native States. Next, we examined whether chemical shifts predicted by LARMOR$^{C\alpha}$ were sensitive to protein structure by assessing their ability to resolve native structure from decoy conformational pools that contained both native and non-native conformers. If sensitive to protein structure, the native-like models in the decoy pool should exhibit the lowest error between LARMOR$^{C\alpha}$ predicted chemical shifts and measured chemical shifts and *vice versa*.

To test this, decoy pools for 4 arbitrarily chosen proteins in the testing were generated using Gō model MD simulations. The final pools contained a total of 100,000 conformers. As shown in Figure 1, the decoy pools generally contained a mixture of native and non-native conformers. For each protein, LARMOR$^{C\alpha}$ was used to predict backbone chemical shifts for every conformer in the decoy pool, and then the corresponding

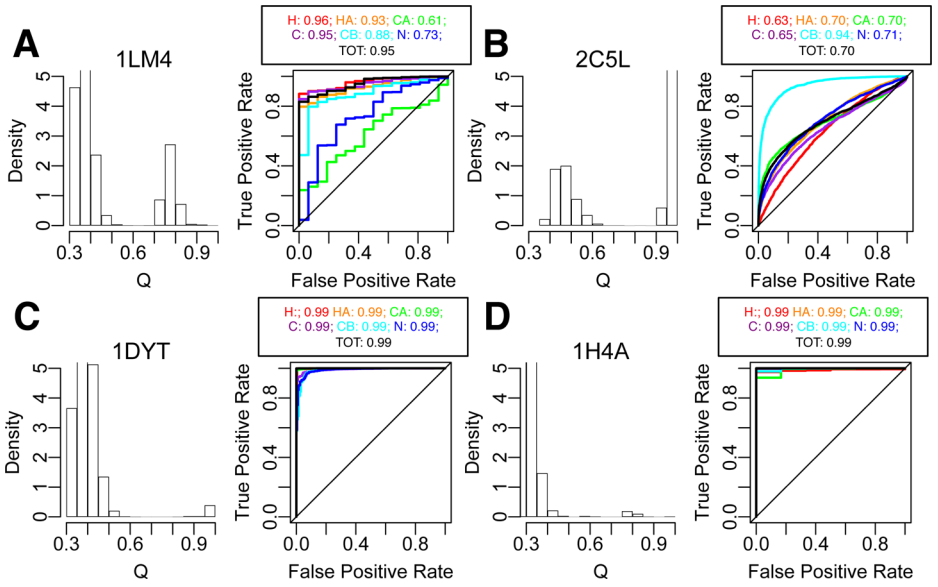


Figure 1. Sensitivity of LARMOR$^{C\alpha}$ chemical shifts to protein structure (I). Shown are the results of using LARMOR$^{C\alpha}$ predicted chemical shifts to resolve native conformers from decoy pools generated using Cα-G$\bar{o}$ model MD simulations. Results are shown for four arbitrarily chosen proteins in the testing set: PDB IDs (A) 1LM4, (B) 2C5L, (C) 1DYT, and (D) 1H4A, respectively. Shown for each protein are plots of the distribution of the fraction of native contacts (Q) in the decoy pool and the ROC curves (right). The plots characterize the degree to which the *w*RMSE between measured and LARMOR$^{C\alpha}$ predicted chemical shifts can distinguish native from non-native conformers in the decoy pools. In addition to ROC curves obtained using the total chemical shift error (black), separate ROC curves are shown when using only ^{1}HN (red), ^{1}Hα (orange), ^{13}Cα (green), ^{13}C (purple), ^{13}Cβ (cyan), or ^{15}N (blue) chemical shifts. The AUC values associated with each ROC curve are shown in boxes.

*w*RMSE was computed. The fraction of native contacts (Q) was also determined for every conformer in the decoy pool. Receiver-operator-characteristic (ROC) analysis was then carried out to assess the extent to which native-like conformers ($Q > 0.90$) could be resolved from non-native conformers ($Q \leq 0.90$).

With the exception of 2C5L, the AUC determined from ROC curves (when using all available backbone chemical shift data) were all ≥0.95; the AUC for 2C5L was ∼0.70 (Figure 1). Similar results were obtained if ^{1}HN, ^{1}Hα, ^{13}Cα, ^{13}C, ^{13}Cβ, or ^{15}N chemical shifts were used separately; with the exception of Cα and N nuclei for 1LM4 and ^{1}HN, ^{1}Hα, ^{13}Cα, ^{13}C, and ^{15}N nuclei for 2C5L, the AUC were all ≥0.88 (Figure 1). Encouragingly, for all four proteins, when using all available backbone chemical shifts, the models with the lowest *w*RMSE had $Q \geq 0.97$ (Figure 2).

As a further test of its sensitivity to structure, we examined whether LARMOR$^{C\alpha}$ predicted chemical shifts could be used to resolve the difference between conformational substates of the phage T4 lysozyme (T4L). The free-energy landscape of a mutant T4L, L99A, has been recently studied using NMR-RD experiments, allowing chemical shifts to be obtained of a transient low-populated (∼3%) conformational substate.[33] Using a mutate-to-trap approach, chemical shifts were also obtained for a triple mutant (L99A-G113A-R119P T4L) that was purported to "resemble" the transient state. The structures of the transient L99A and the triple T4L mutants were determined using CS-Rosetta and confirmed that the structure of the transient state closely resembles that of the triple T4L mutant. The RMSD between the transient state of the L99A mutant and the triple T4L mutant was ∼0.8 Å, whereas the RMSDs of the transient single and triple mutants compared to the highly populated state of L99 T4L were ∼2.5 and 2.3 Å, respectively (Figure 3).

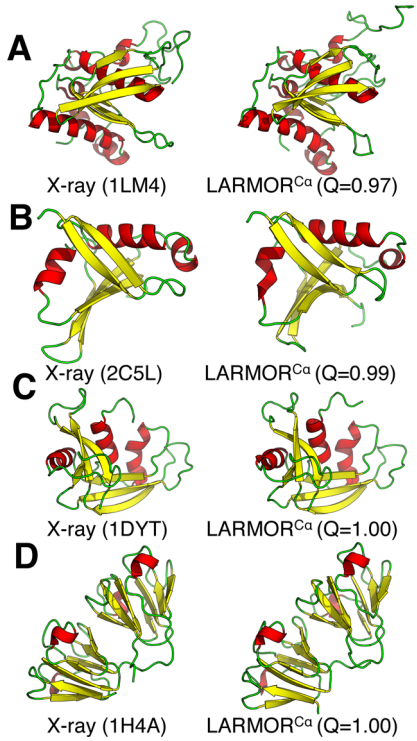


Figure 2. Comparison between X-ray structures and models with the lowest chemical shift error. Side-by-side comparison of the X-ray structure (*left*) and the models with the lowest total error (*w*RMSE) between experimental and LARMOR$^{C\alpha}$-predicted chemical shifts (*right*) for proteins corresponding to PDB IDs (A) 1LM4, (B) 2C5L, (C) 1DYT, and (D) 1H4A. In each panel, the fraction of native contacts (Q) is indicated below the model with the lowest chemical shift error.

dx.doi.org/10.1021/ct5009125 | *J. Chem. Theory Comput.* 2015, 11, 325−331

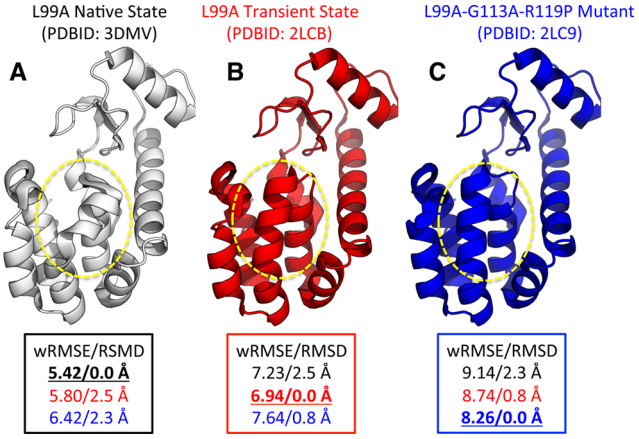


Figure 3. Sensitivity of LARMOR$^{C\alpha}$ chemical shifts to protein structure (II): Structures of three conformational substates of T4L: (A) native L99A T4L, (B) the transiently populated intermediate of L99A T4L, and (C) the L99A-G113A-R119P T4L triple mutant, respectively. The region in the transient intermediate of L99A and the triple mutant that differs significantly from native L99A T4L is circled (*yellow dotted*). LARMOR$^{C\alpha}$ backbone chemical shifts were predicted from the Cα coordinates taken from the solved structure of each of these three substates and then compared to NMR-RD-derived chemical shifts of the native L99A T4L, the transient intermediate state of L99A T4L, and the triple T4L mutant, respectively. For each structure, the *w*RMSE relative to the native L99A T4L, the transient intermediate state of L99A T4L, and the triple T4L mutant are shown in black, red, and blue, respectively (*boxes*) and the lowest is highlighted (*bold* and *underlined*). Also, for each structure, the structural RMSD relative to the native L99A T4L, the transient intermediate state of L99A T4L, and the L99A-G113A-R119P T4L mutant structure are shown in black, red, and blue, respectively (*boxes*). Here, the *w*RMSE and RMSDs were calculated for residues 100−120 and 132−146.

To test whether LARMOR$^{C\alpha}$ could resolve the small structural differences between these three states (namely, the highly and transiently populated states of L99A T4L and the conformation of the triple mutant), LARMOR$^{C\alpha}$ was used to predict backbone chemical shifts from the solved structures of each species. For each species, we computed the *w*RMSE between the predicted and experimental chemical shifts; the *w*RMSEs were computed using data for residues 100−120 and 132−146 as these were the only residues that exhibited significant changes in chemical shifts between the different the states of T4L. We expect that the structures with the lowest *w*RMSE should match the system associated with reference (experimental) chemical shifts. As shown in Figure 3, this was indeed the case. The L99A T4L structure exhibited the lowest *w*RMSE relative to the chemical shifts computed for the highly populated state, the transient state L99A T4L structure showed the lowest *w*RMSE relative to the transient-state chemical shifts, and the triple mutant structure displayed the lowest *w*RMSE relative to the mutant chemical shifts. These results indicate that LARMOR$^{C\alpha}$ was able to resolve the small structural difference between conformational substates of T4L.

Although LARMOR$^{C\alpha}$ was able to resolve the "correct" structure based upon the chemical shifts, the errors for the L99A transient state and the triple mutant were higher than the error for the L99A T4L. The higher errors for the transient states suggest that models for these states can be refined even further. Indeed, during the CS-Rosetta protocol used to generate these models, it was assumed, based upon chemical shifts dispersion patterns, that only residues 100−120 and 132−146 were significantly different between the L99A transient state and the triple mutant. Thus, during refinement only atoms in these residues were allowed to deviate from the native L99A T4L structure.

Together these results indicate that backbone chemical shifts predicted by LARMOR$^{C\alpha}$ are sufficiently sensitive to protein structure to allow chemical shifts to be used in resolving native from non-native structure. Even small structural differences between similar conformational substates can be detected. As such, NMR chemical shifts should be useful in "parsing" trajectories and ensembles generated using coarse-grained simulations to identify physically relevant conformational substates along the folding pathway of proteins.

Analysis Using LARMOR$^{C\alpha}$ Indicates That the Transient State of gpW Is Locally Unfolded. Recently, NMR relaxation dispersion (RD) experiments were used to study the free-energy landscape of gpW, a 62-residue α+β fast-folding protein (see Figure 4). NMR RD experiments allowed chemical shifts to be obtained for both the native-state and a low-populated transient state.[46] Analysis of the chemical shift dispersion pattern of the transient state revealed that the helices remained intact, whereas the beta-strand region was unfolded.[46] In principle combining LARMOR$^{C\alpha}$ with coarse-grained simulations should allow for structures consistent with the chemical shifts of the transient state to be identified. Thus, we used LARMOR$^{C\alpha}$ to probe the folding pathway of gpW during Gō model MD simulations.

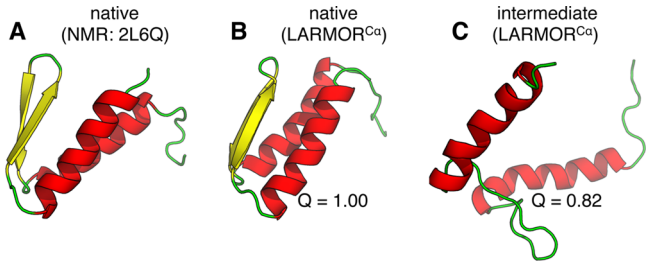


Figure 4. Resolving native and transient states along the folding pathway of the fast-folding protein gpW using LARMOR$^{C\alpha}$. The folding pathway of gpW was studied using Cα-based Gō-model MD simulations. Shown are cartoon representations comparing (A) the solved native-state structure of the gpW and the representative models of (B) the native and (C) the transient states selected from the Cα-trajectory using LARMOR$^{C\alpha}$. Representative models were selected by comparing LARMOR$^{C\alpha}$-predicted chemical shifts to recently reported NMR-RD-derived backbone chemical shifts for the native and the transient intermediate states.[46] The models in (B) and (C) correspond to the two models that were closest to the average structure of the 10 models that exhibited the lowest error (*w*RMSE) between LARMOR$^{C\alpha}$-predicted and the measured chemical shifts of the native state and the transient state, respectively.

The representative model based on the native state chemical shifts was found to contain α+β topology (Figure 4B), indicating that the LARMOR$^{C\alpha}$ was able to resolve the native structure from the ensemble of structures generated during the Gō model simulations. In contrast to the representative model of the native states, the representative model of the transient-state exhibited an unfolded beta-region (Figure 4C). These results agree well with the analysis of Kay and co-workers,[46] and they serve to further confirm that LARMOR$^{C\alpha}$ can be used to efficiently parse coarse-grained trajectories and ensembles to

dx.doi.org/10.1021/ct5009125 | J. Chem. Theory Comput. 2015, 11, 325−331

identify important conformational substates (i.e., both native and intermediary states).

Though in the current study we focused on using LARMOR$^{C\alpha}$ to *seamlessly* incorporate backbone chemical shifts into the analysis of coarse-grained MD simulations, LARMOR$^{C\alpha}$ is also well suited for incorporation into most protein structure prediction methods where it can be used to enable backbone chemical shifts to actively guide conformation sampling. Additionally, LARMOR$^{C\alpha}$ can also be used to parse large all-atom trajectories and ensembles to identify a smaller subset of relevant conformational states. In the spirit of "multi-scale analysis",[36] more accurate and complete chemical shifts prediction (i.e., prediction of both backbone and side-chain chemical shifts) can then be carried out for the smaller subset using all-atom prediction approaches.

CONCLUSION

In summary, we have developed LARMOR$^{C\alpha}$, a Cα-based approach that enables the prediction of backbone chemical shifts from coarse-grained models of proteins. We show that in addition to predicting chemical shifts with accuracy comparable to some all-atom approaches, LARMOR$^{C\alpha}$ was capable of resolving protein structure. This sensitivity to protein structure enables LARMOR$^{C\alpha}$ to identify conformational substates from coarse-grained simulations that are consistent with available NMR chemical shifts. An exciting application of the method is to identify "invisible" intermediate substates using chemical shifts obtained from NMR relaxation dispersion experiments, as was demonstrated here for the gpW fast-folding protein. Structural information on transiently populated intermediates afforded by the combination of coarse-grained simulation and LARMOR$^{C\alpha}$ has the potential to offer functional insights into the mechanism of protein folding, misfolding, and aggregation, and their role in folding-related diseases.[7] Beyond coarse-grained simulations, LARMOR$^{C\alpha}$ could be used to quickly parse all-atom MD trajectories and also be incorporated into existing structure prediction methods. To facilitate its use, the source code for LARMOR$^{C\alpha}$ is made freely available at https://github.com/atfrank/LARMORCA.

ASSOCIATED CONTENT

Supporting Information

Tables S1−S4: (**S1**) training and testing sets; (**S2**) LARMOR$^{C\alpha}$ structure features used to predict backbone chemical shifts; (**S3**, **S4**) Comparison between LARMOR$^{C\alpha}$ and all-atom chemical shift prediction methods. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Authors

*Phone: 734 615-0609. Fax: 734 647-1604. E-mail: afrankz@umich.edu (A.T.F.).

*Phone: 734 647-6682. Fax: 734 647-1604. E-mail: brookscl@umich.edu (C.L.B).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the NSF through the Center for Theoretical Biological Physics (PHY0822283). A.T.F. was supported through the University of Michigan President's Postdoctoral Fellowship, and L.S.A. received funding from the NIH Ruth L. Kirschstein NRSA Postdoctoral Fellowship (GM108298).

REFERENCES

(1) Bai, Y.; Sosnick, T.; Mayne, L.; Englander, S. Protein-Folding Intermediates: Native-State Hydrogen-Exchange. *Science* **1995**, *269*, 192−197.

(2) Juneja, J.; Udgaonkar, J. B. Characterization of the Unfolding of Ribonuclease a by a Pulsed Hydrogen Exchange Study: Evidence for Competing Pathways for Unfolding. *Biochemistry* **2002**, *41*, 2641−2654.

(3) Sekhar, A.; Kay, L. E. NMR Paves the Way for Atomic Level Descriptions of Sparsely Populated, Transiently Formed Biomolecular Conformers. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 12867−12874.

(4) Long, D.; Bouvignies, G.; Kay, L. E. Measuring Hydrogen Exchange Rates in Invisible Protein Excited States. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 8820−8825.

(5) Hansen, A. L.; Kay, L. E. Measurement of Histidine pKa Values and Tautomer Populations in Invisible Protein States. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, E1705−E1712.

(6) Malhotra, P.; Udgaonkar, J. B. High-Energy Intermediates in Protein Unfolding Characterized by Thiol Labeling Under Nativelike Conditions. *Biochemistry* **2014**, *53*, 3608−3620.

(7) Dobson, C. M. Protein Folding and Misfolding. *Nature* **2003**, *426*, 884−890.

(8) Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; Shaw, D. E. Anton, a Special-Purpose Machine for Molecular Dynamics Simulation. *Commun. ACM* **2008**, *51*, 91−97.

(9) Beberg, A. L.; Ensign, D. L.; Jayachandran, G.; Khaliq, S.; Pande, V. S. *Folding@ Home: Lessons From Eight Years of Volunteer Distributed Computing* **2009**, 1−8.

(10) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. MSMBuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale. *J. Chem. Theory Comput.* **2011**, *7*, 3412−3419.

(11) Eastman, P.; Friedrichs, M. S.; Chodera, J. D.; Radmer, R. J.; Bruns, C. M.; Ku, J. P.; Beauchamp, K. A.; Lane, T. J.; Wang, L.-P.; Shukla, D.; Tye, T.; Houston, M.; Stich, T.; Klein, C.; Shirts, M. R.; Pande, V. S. OpenMM 4: a Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J. Chem. Theory Comput.* **2013**, *9*, 461−469.

(12) Lane, T. J.; Shukla, D.; Beauchamp, K. A.; Pande, V. S. To Milliseconds and Beyond: Challenges in the Simulation of Protein Folding. *Curr. Opin. Struct. Biol.* **2013**, *23*, 58−65.

(13) Clementi, C. Coarse-Grained Models of Protein Folding: Toy Models or Predictive Tools? *Curr. Opin. Struct. Biol.* **2008**, *18*, 10−15.

(14) Hills, R. D.; Brooks, C. L. Insights From Coarse-Grained Gō Models for Protein Folding and Dynamics. *Int. J. Mol. Sci.* **2009**, *10*, 889−905.

(15) Takada, S. Coarse-Grained Molecular Simulations of Large Biomolecules. *Curr. Opin. Struct. Biol.* **2012**, *22*, 130−137.

(16) Chan, H. S.; Zhang, Z.; Wallin, S.; Liu, Z. Cooperativity, Local-Nonlocal Coupling, and Nonnative Interactions: Principles of Protein Folding From Coarse-Grained Models. *Annu. Rev. Phys. Chem.* **2011**, *62*, 301−326.

(17) Karpen, M.; Tobias, D.; Brooks, C. Statistical Clustering-Techniques for the Analysis of Long Molecular-Dynamics Trajectories - Analysis of 2.2-Ns Trajectories of YPGDV. *Biochemistry* **1993**, *32*, 412−420.

(18) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. Peptide Folding: When Simulation Meets Experiment. *Angew. Chem., Int. Ed.* **1999**, *38*, 236−240.

(19) Hubner, I. A.; Deeds, E. J.; Shakhnovich, E. I. Understanding Ensemble Protein Folding at Atomic Detail. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 17747−17752.

(20) Rajan, A.; Freddolino, P. L.; Schulten, K. Going Beyond Clustering in MD Trajectory Analysis: an Application to Villin Headpiece Folding. *PLoS One* **2010**, *5*, e9890.

(21) Marsh, J. A.; Forman-Kay, J. D. Ensemble Modeling of Protein Disordered States: Experimental Restraint Contributions and Validation. *Proteins: Struct., Funct., Bioinf.* **2011**, *80*, 556−572.
(22) Chen, Y.; Campbell, S. L.; Dokholyan, N. Deciphering Protein Dynamics From NMR Data Using Explicit Structure Sampling and Selection. *Biophys. J.* **2007**, *93*, 2300−2306.
(23) Berlin, K.; Castañeda, C. A.; Schneidman-Duhovny, D.; Sali, A.; Nava-Tudela, A.; Fushman, D. Recovering a Representative Conformational Ensemble From Underdetermined Macromolecular Structural Data. *J. Am. Chem. Soc.* **2013**, *135*, 16595−16609.
(24) Fisher, C. K.; Huang, A.; Stultz, C. M. Modeling Intrinsically Disordered Proteins with Bayesian Statistics. *J. Am. Chem. Soc.* **2010**, *132*, 14919−14927.
(25) Xiao, X.; Kallenbach, N.; Zhang, Y. Peptide Conformation Analysis Using an Integrated Bayesian Approach. *J. Chem. Theory Comput.* **2014**, *10*, 4152−4159.
(26) Holm, L.; Sander, C. Database Algorithm for Generating Protein Backbone and Side-Chain Coordinates From a C-Alpha Trace Application to Model-Building and Detection of Coordinate Errors. *J. Mol. Biol.* **1991**, *218*, 183−194.
(27) Sali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **1993**, *234*, 779−815.
(28) Feig, M.; Karanicolas, J.; Brooks, C. L. I. MMTSB Tool Set: Enhanced Sampling and Multiscale Modeling Methods for Applications in Structural Biology. *J. Mol. Graphics Modell.* **2004**, *22*, 377−395.
(29) Petrey, D.; Xiang, Z. X.; Tang, C. L.; Xie, L.; Gimpelev, M.; Mitros, T.; Soto, C. S.; Goldsmith-Fischman, S.; Kernytsky, A.; Schlessinger, A.; Koh, I. Y. Y.; Alexov, E.; Honig, B. Using Multiple Structure Alignments, Fast Model Building, and Energetic Analysis in Fold Recognition and Homology Modeling. *Proteins: Struct., Funct., Bioinf.* **2003**, *53*, 430−435.
(30) Rotkiewicz, P.; Skolnick, J. Fast Procedure for Reconstruction of Full-Atom Protein Models From Reduced Representations. *J. Comput. Chem.* **2008**, *29*, 1460−1465.
(31) Li, Y. Q.; Zhang, Y. REMO: a New Protocol to Refine Full Atomic Protein Models From C-Alpha Traces by Optimizing Hydrogen-Bonding Networks. *Proteins: Struct., Funct., Bioinf.* **2009**, *76*, 665−674.
(32) Korzhnev, D. M.; Religa, T. L.; Banachewicz, W.; Fersht, A. R.; Kay, L. E. A Transient and Low-Populated Protein-Folding Intermediate at Atomic Resolution. *Science* **2010**, *329*, 1312−1316.
(33) Bouvignies, G.; Vallurupalli, P.; Hansen, D. F.; Correia, B. E.; Lange, O.; Bah, A.; Vernon, R. M.; Dahlquist, F. W.; Baker, D.; Kay, L. E. Solution Structure of a Minor and Transiently Formed State of a T4 Lysozyme Mutant. *Nature* **2012**, *477*, 111−114.
(34) Neudecker, P.; Robustelli, P.; Cavalli, A.; Walsh, P.; Lundstrom, P.; Zarrine-Afsar, A.; Sharpe, S.; Vendruscolo, M.; Kay, L. E. Structure of an Intermediate State in Protein Folding and Aggregation. *Science* **2012**, *336*, 362−366.
(35) Zhang, H.; Neal, S.; Wishart, D. S. RefDB: a Database of Uniformly Referenced Protein Chemical Shifts. *J. Biomol. NMR* **2003**, *25*, 173−195.
(36) Law, S. M.; Frank, A. T.; Brooks, C. L. PCASSO: a Fast and Efficient Cα-Based Method for Accurately Assigning Protein Secondary Structure Elements. *J. Comput. Chem.* **2014**, *35*, 1757−1761.
(37) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(38) Kreusch, A.; Spraggon, G.; Lee, C. C.; Klock, H.; McMullan, D.; Ng, K.; Shin, T.; Vincent, J.; Warner, I.; Ericson, C.; Lesley, S. A. Structure Analysis of Peptide Deformylases fromStreptococcus Pneumoniae, Staphylococcus Aureus, Thermotoga maritima and Pseudomonas Aeruginosa: Snapshots of the Oxygen Sensitivity of Peptide Deformylase. *J. Mol. Biol.* **2003**, *330*, 309−321.
(39) Bunney, T. D.; Harris, R.; Gandarillas, N. L.; Josephs, M. B.; Roe, S. M.; Sorli, S. C.; Paterson, H. F.; Rodrigues-Lima, F.; Esposito, D.; Ponting, C. P.; Gierschik, P.; Pearl, L. H.; Driscoll, P. C.; Katan, M. Structural and Mechanistic Insights Into Ras Association Domains of Phospholipase C Epsilon. *Mol. Cell* **2006**, *21*, 495−507.
(40) Mallorqui-Fernandez, G.; Pous, J.; Peracaula, R.; Aymami, J.; Maeda, T.; Tada, H.; Yamada, H.; Seno, M.; de Llorens, R.; Gomis-Ruth, F. X.; Coll, M. Three-Dimensional Crystal Structure of Human Eosinophil Cationic Protein (RNase 3) at 1.75 Angstrom Resolution. *J. Mol. Biol.* **2000**, *300*, 1297−1307.
(41) Basak, A.; Bateman, O.; Slingsby, C.; Pande, A.; Asherie, N.; Ogun, O.; Benedek, G. B.; Pande, J. High-Resolution X-Ray Crystal Structures of Human gammaD Crystallin (1.25 a) and the R58H Mutant (1.15 a) Associated with Aculeiform Cataract. *J. Mol. Biol.* **2003**, *328*, 1137−1147.
(42) Karanicolas, J.; Brooks, C. L. Improved Go-Like Models Demonstrate the Robustness of Protein Folding Mechanisms Towards Non-Native Interactions. *J. Mol. Biol.* **2003**, *334*, 309−325.
(43) Karanicolas, J.; Brooks, C. L., III. The Origins of Asymmetry in the Folding Transition States of Protein L and Protein G. *Protein Sci.* **2009**, *11*, 2351−2361.
(44) Brooks, B.; Bruccoleri, R.; Olafson, B.; et al. CHARMM: a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.
(45) Barth, E.; Kuczera, K.; Leimkuhler, B.; Skeel, R. D. Algorithms for Constrained Molecular-Dynamics. *J. Comput. Chem.* **1995**, *16*, 1192−1209.
(46) Sanchez-Medina, C.; Sekhar, A.; Vallurupalli, P.; Cerminara, M.; Muñoz, V.; Kay, L. E. Probing the Free Energy Landscape of the Fast-Folding gpW Protein by Relaxation Dispersion NMR. *J. Am. Chem. Soc.* **2014**, *136*, 7444−7451.
(47) Sborgi, L.; Verma, A.; Muñoz, V.; de Alba, E. Revisiting the NMR Structure of the Ultrafast Downhill Folding Protein gpW From Bacteriophage Λ. *PLoS One* **2011**, *6*, e26409.
(48) Han, B.; Liu, Y.; Ginzinger, S. W.; Wishart, D. S. SHIFTX2: Significantly Improved Protein Chemical Shift Prediction. *J. Biomol. NMR* **2011**, *50*, 43−57.
(49) Shen, Y.; Bax, A. SPARTA+: a Modest Improvement in Empirical NMR Chemical Shift Prediction by Means of an Artificial Neural Network. *J. Biomol. NMR* **2010**, *48*, 13−22.
(50) Kohlhoff, K. J.; Robustelli, P.; Cavalli, A.; Salvatella, X.; Vendruscolo, M. Fast and Accurate Predictions of Protein NMR Chemical Shifts From Interatomic Distances. *J. Am. Chem. Soc.* **2009**, *131*, 13894−13895.
(51) Meiler, J. J. PROSHIFT: Protein Chemical Shift Prediction Using Artificial Neural Networks. *J. Biomol. NMR* **2003**, *26*, 25−37.
(52) Xu, X.-P.; Case, D. A. Probing Multiple Effects on 15N, 13C Alpha, 13C Beta, and 13C′ Chemical Shifts in Peptides Using Density Functional Theory. *Biopolymers* **2002**, *65*, 408−423.