Contact Frequency Necessary to Recreate Protein Structure

Andrew Latham*
Massachusetts Institute of Technology
Department of Chemistry
6-222
77 Massachusetts Ave.
Cambridge, MA 02139
(Dated: May 18, 2018)

Protein folding presents a challenge to experimental and theoretical studies alike. The combinatoric complexity of long polymers makes simple ensemble searches impractical. This leads to the conclusion that in naturally selected proteins, energetics must guide the protein to a low-energy native state capable of function. Many studies have attempted to use combinations of physical insight and bioinformatics to model protein structure. In this vein, our study utilizes the biasing of protein structures in molecular dynamics simulations towards their native state, using either harmonic or maximum entropy approaches. We then compare these two conditions, and show that protein structure can be recreated with a small fraction of overall contacts. While we apply this to lysine-free ubiquitin (K0-Ub), our broadly applicable methods should enable study of more compelling proteins in the future.

I. INTRODUCTION

Protein structure and function are inexorably linked: this has lead to extensive theoretical and experimental investigation into protein folding.[1] Proteins must fold both quickly and accurately to serve biological functions, yet a simple conformation search would need to explore approximately 10^{143} states.[2] A more complex interplay between energy and entropy seems necessary to reliably guide a protein to its native, active fold.

A plethora of theoretical and experimental investigations into the dynamics and energetics of protein folding have produced many models of this vital process. Models such as the random energy model[3] and designed random energy model[4] attempt to explain folding through looking at energy differences between metastable glass transition states and the true native configuration. Meanwhile, experimental techniques often probe how changes in residue affect protein folding rates.[5] These studies ultimately investigate how stabilizing interactions such as hydrogen bonding, hydrophobic interactions, and solvent interactions influence protein structure.

Computational methods have attempted to use knowledge of these interactions to describe protein folding. Models such as AMBER or CHARMM use quantum mechanics and experimental data to derive a force field that describes the vast diversity of inter-residue interactions.[6] Similarly, the AWSEM force field is a coarse-grained description of protein interactions that combines physically motivated terms with bioinformatics data. [7] These bioinformatics terms rely on data from homologous sequences to predict novel structures. Additionally, the AWSEM force field utilizes crystal structure measurements to bias the energy landscape. This biasing is done through a Go model[8] that biases the structure towards having the same tertiary contact frequencies seen in the crystal structure. These canonical models work well for solid proteins if the crystal structure is similar to the active structure. However, many proteins' native conformation either differs from their crystal structure or is hard to determine.

Nuclear magnetic resonance (NMR) experiments allow the study of proteins in their ensemble of native conformations. Nuclear Overhauser effect spectroscopy (NOESY) uses cross relaxation to detect through-space correlations between nuclei.[9] The magnitude of these cross correlation peaks is related to distances in the protein structure. 15N-NOESY-HSQC experiments correlate an individual amide nitrogen and its proton to all other protons that are nearby, from which inter-residue contacts can be inferred. [10] Computational techniques then piece together the native structure by combining sequence information with these contacts. Our work shall examine a complete structure of lysine-free ubiquitin (K0-Ub) determined by 15N-NOESY-HSQC, shown in Figure 1.[11]

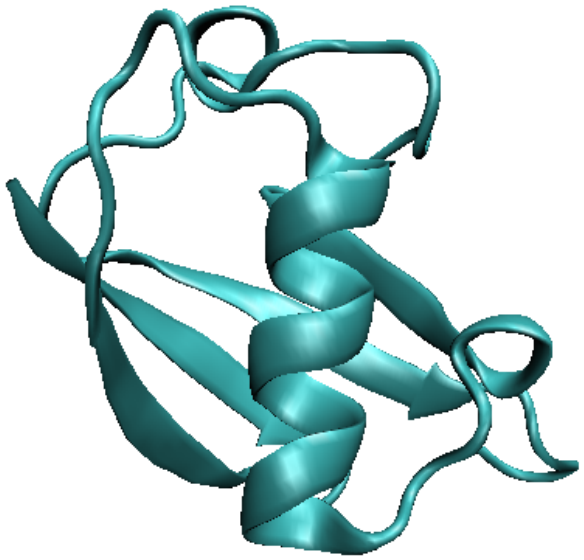

FIG. 1. Above is the determined structure of lysine-free ubiquitin (K0-Ub), used in this study.[11]

Using a solved structure of ubiquitin, our study will exam-

* aplatham@mit.edu

ine the amount of known contacts necessary to adequately maintain native structure. We compare results from harmonic and maximum entropy biases to study the fraction of known distances necessary to adequately characterize a protein. Hopefully, this method could be developed to study more complicated proteins, where only a fraction of the overall contacts are known.

II. METHODOLOGY

Our goal is to use harmonic potentials or maximum entropy to bias simulations with varying amounts of known contacts. Contacts were implemented as distance restraints and only considered between alpha carbons. For harmonic potentials, the percentage of known contacts varied from 0% to 10% by 1% and from 10% to 100% by 10%. For sufficient statistics, 20 samples were taken, with different contacts randomly sampled each time. The maximum entropy samples were taken from 0% to 100% by 10%. Due to computational cost, only one trial was attempted in the maximum entropy case. Each simulation was performed in the NVT ensemble using LAMMPS.[12] They ran for 2 ns, with timesteps of 2 fs using the AWSEM potential.[7] AWSEM terms that require any knowledge of structure or make bioinformatics predictions were excluded from the potential.

A. Harmonic Potential

For each system, contacts were randomly selected at a the contact percentages listed above. Then, each distance was biased according to

$$\hat{H} = \hat{H}_{AWSEM} + \sum_i k_i (d_{ab,exp} - d_{ab})^2 \tag{1}$$

where k_i is the spring constant, i is the particular contact, $d_{ab,exp}$ is the contact distances determined by NMR, and d_{ab} is the simulation distance at each timestep. For this simulation, $k = 1 \frac{kcal}{\mathring{A}^2 * mol}$, which is only one quarter the strength of a bond.

B. Maximum Entropy

Our work utilizes maximum entropy principles to bias toward the distances determined experimentally. This approach, similar to the model of chromatin investigated by Bin Zhang, is minimally biased and simple to simulate.[13] According to the maximum entropy principle,[14] the overall Hamiltonian is

$$\hat{H}_{ME} = \hat{H}_{AWSEM} + \sum_i \alpha_i f_i(r_{ab}) \tag{2}$$

where $\hat{H}_{AWSEM}$ is the energy according to the AWSEM simulation package[7] and $f_i(r_{ab})$ is ensemble averages of collective variables which also are measured experimentally. In this case, each $f_i(r_{ab})$ is the distance between atom a and atom b.

The Lagrangian multipliers α_i from Equation 2 are determined through minimization of the objective function,

$$\Gamma(\alpha) = \frac{\beta^2}{2} \alpha^T * B * \alpha - \beta [< f_i > - f_{i,exp}]^T \tag{3}$$

which results from a Taylor expansion of a ratio of the partion function with and without the perturbation,[13] where

$$B_{ij} = < f_i f_j > - < f_i >< f_j > \tag{4}$$

is a Hermitian matrix related to the covariance of the ensemble functions. Ultimately, this leads to an iterative procedure where

$$\alpha_{j+1} = \alpha_j + \frac{1}{\beta} B^{-1} [< f_i > - f_{i,exp}]^T \tag{5}$$

updates α after each iteration j. This procedure is continued until the overall error,

$$\frac{\sum_i | < f_i > - f_{i,exp}|}{\sum_i f_{i,exp}} \tag{6}$$

is less than a minimum tolerance. At this point, realistic simulation parameters have been established and results are used.

Unfortunately, our error, shown in Equation 6, never converged for our studied system. Therefore, our results are based on the first twenty studied iterations.

III. RESULTS

Two different metrics measured differences between our original structure and the simulated structure. The RMSD

$$RMSD(\vec{u}, \vec{v}) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} ||\vec{u}_i - \vec{v}_i||^2} \tag{7}$$

measures the distance between each atom in its current structure, $\vec{u}_i$, and its position in a reference structure, $\vec{v}_i$, in this case the NMR structure determined experimentally. Lower values of RMSD confirm a better match to experimental data.

Another similarity parameter, Q, is defined

$$Q = \frac{2}{(N-2)(N-3)} \sum_{i<j-2} exp\left(- \frac{(r_{ij} - r_{ij,exp})^2}{2\sigma_{ij}^2} \right) \tag{8}$$

where N is the number of atoms, r_{ij} is the distance in the simulation, and $r_{ij,exp}$ is the distance in the NMR structure. σ_{ij} measures the distance along the sequence for the two amino acids and is given by $\sigma_{ij} = (1 + |i-j|^{0.15})$. Q can take values between 0 and 1, where 1 is a perfect match to experimental structures. Using RMSD, Q, and sum error, we can sufficiently characterize when we are able to reproduce NMR structures.

A. Harmonic Biasing

As discussed previously, each contact was harmonically biased toward the experimental structure with $k = 1\frac{kcal}{\mathring{A}^2 * mol}$ for each contact. After twenty attempts with each fraction of restraints, we had sufficient statistics to analyze the effect of percentage of known contacts on structure stability.

First, we computed the sum error, shown in Equation 6. For the sum error, only atoms being biased contribute. Therefore, it is not defined when no contacts are made. Figure 2 plots the sum error as a function of the percentage of known contacts. Note that the sum error monotonically decreases, within statistical error. These results show that as more contacts are known, the structure deviates less from the experimental structure. This similarity causes known distances to vary less as more restraints are created.

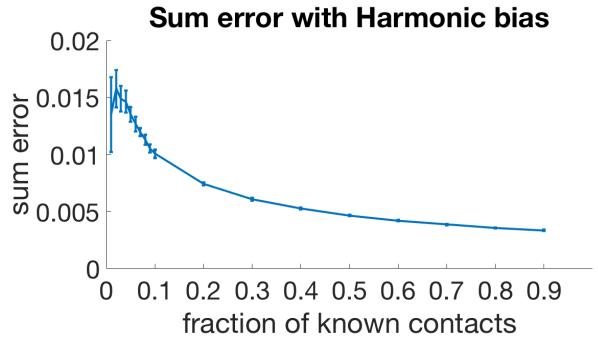

FIG. 2. The sum error is plotted against the fraction of known contacts for harmonic biasing. Note that the error decreases as more contacts are known. The realization of other contacts in the system stabilizes already known contacts, causing the structure to more closely resemble the experimental structure.

While analysis of the sum error is useful, it only analyzes biased contacts. Therefore, RMSD and Q are more useful structural comparisons. These variables capture how only a small fraction of contacts can result in similarity throughout the structure. RMSD is plotted as a function of known contacts in Figure 3, while Q is plotted as a function of known contacts in Figure 4.

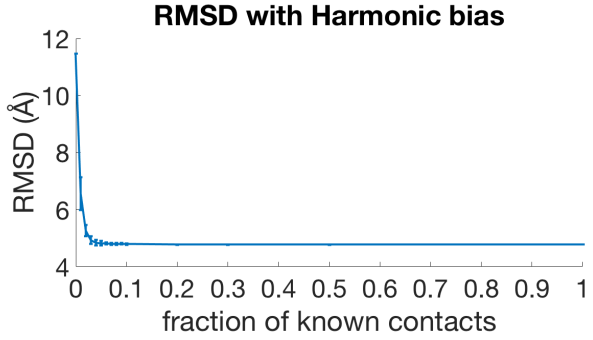

FIG. 3. The RMSD is plotted against the fraction of known contacts for harmonic biasing. Note how the system converges to an RMSD of 4.8 after 3% of contacts are known.

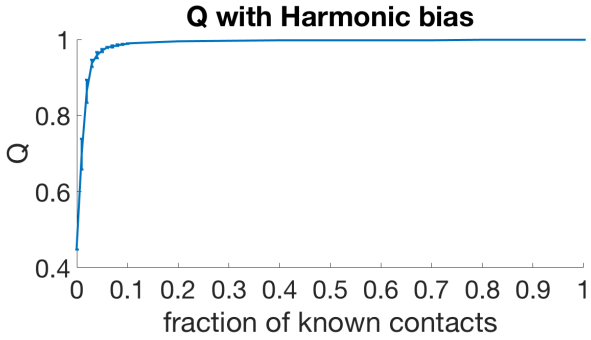

FIG. 4. The RMSD is plotted against the fraction of known contacts for harmonic biasing. Note how the system converges to an Q of 1 when many contacts are known. While the Q converges slower than the RMSD, good agreement is still seen after 3% of distance restraints.

The RMSD, Figure 3, decreases as a function of known contacts. In fact, going from zero contacts to 1% of known contacts decreases the RMSD by more than half. This continues to decreases until converging to a known structure. At approximately 3% of contacts, the structure accurately reproduces the experimental structure. Similarly, Figure 4 shows Q increasing as a function of time. Again, even biasing 1% of contacts improves Q from 0.45 to 0.69. Complete convergence to a known structure, seen by a Q value of 1, is seen at about 20% of known contacts. However, good agreement between structures is seen above 3% of contacts, similar to RMSD.

B. Maximum Entropy Biasing

As an iterative algorithm, maximum entropy suggests the the sum error should monotonically decrease with the number of iterations. Otherwise, solutions to the Lagrangian multipliers, α_i will diverge. Unfortunately, we were unable to effectively implement maximum entropy biasing. The sum error is displayed across iterations in Figure 5.

Possible reasons for this error are several. Two discussed possibilities relate to the nature of studying molecular dynamics. Modifying atomic potentials should not agitate the original system dynamics, but rather gently guide the dynamics toward the desired state. One issue originated from the magnitude of the forces created by the algorithm. The forces generated by maximum entropy were too large, causing issues where neighboring atoms came in contact before neighbor lists could be updated. Normalizing the bias to our simulation allowed the simulation to run, but may have lost necessary information in the process. Additionally, the forces may have still been too large to accurately model dynamics. The linear nature of the biasing potential creates a discontinuos force, which may be another issue in our simulation. In molecular dynamics, discontinuos forces may cause particles to jump because of a quick switch in velocity. These jumps endanger neighbor lists, inaccurately represent dynamics, and may cause the poor results seen here. In the future, implementing

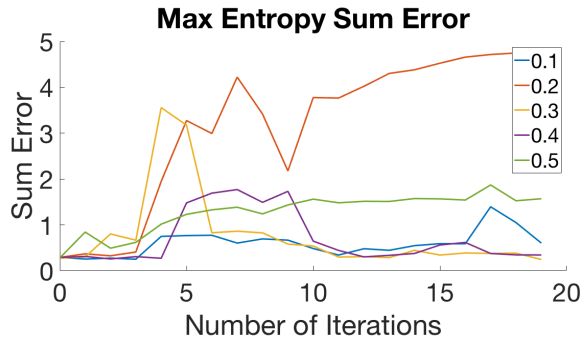

FIG. 5. The sum error is plotted for the maximum entropy case. Five of the ten fractions of known contacts are shown for simplicity, as indicated in the Legend. It is believed that the overall biasing term was too large relative to the rest of the simulation. Therefore, the sum error is not a monotonically decreasing function of the number of iterations.

a switching or step function may fix this issue and allow for better results.

IV. CONCLUSIONS

In this study, maximum entropy biasing and harmonic biasing were introduced to examine how to couple protein structure with simulation data. More work determining realistic energy sizes is necessary to implement maximum entropy analysis. However, the harmonic results were promising. From analysis of RMSD and Q, we conclude that only 3% of contacts are necessary to recreate accurate structures. In fact, even 1% of random contacts creates drastic improvements to structural predictions. These results are deduced from data in Figure 3 and Figure 4.

In conclusion, even small amounts of structural information help with structure prediction. While this study only considered one protein, further work should analyze fractions of contacts necessary to recreate structure across a wider variety of proteins. Furthermore, steps such as determining the correct energy bias magnitude and implementing a differentiable functional form may allow a more successful maximum entropy biasing, which creates a minimally biased restraint.[14] Finally, our results have shown that knowing even 1% of contacts drastically improves structure prediction. Therefore, we hope to apply similar techniques to study more interesting biological problems, such as intrinsically disordered proteins, in the future.

[1] Anfinsen, C. B. *Science* **1973**, *181*, 223–230.
[2] Levinthal, C. How to fold graciously. Mossbauer spectroscopy in biological systems. Urbana, IL, 1969.
[3] Bryngelson, J. D.; Wolynes, P. G. *The Journal of Physical Chemistry* **1989**, *93*, 6902–6915.
[4] Abkevich, V. I.; Gutin, A. M.; Shakhnovich, E. I. *Folding and Design* **1996**, *1*, 221–230.
[5] Baldwin, R. L. *Protein Science* **2008**, *9*, 207–207.
[6] Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. *PLOS ONE* **2012**, *7*.
[7] Davtyan, A.; Schafer, N. P.; Zheng, W.; Clementi, C.; Wolynes, P. G.; Papoian, G. A. *The Journal of Physical Chemistry B* **2012**, *116*, 8494–8503.
[8] Hiroshi, T.; Yuzo, U.; Gō, N. *International Journal of Peptide and Protein Research* **1975**, *7*, 445–459.
[9] Anet, F. A. L.; Bourn, A. J. R. *Journal of the American Chemical Society* **1965**, *87*, 5250–5251.
[10] Aue, W. P.; Bartholdi, E.; Ernst, R. R. *The Journal of Chemical Physics* **1976**, *64*, 2229–2246.
[11] Huang, T.; Li, J.; Byrd, R. A. *Protein Science 23*, 662–667.
[12] Plimpton, S. *Journal of Computational Physics* **1995**, *117*, 1–19.
[13] Zhang, B.; Wolynes, P. G. *Proc Natl Acad Sci U S A* **2015**, *112*, 6062–6067.
[14] Pitera, J. W.; Chodera, J. D. *Journal of Chemical Theory and Computation* **2012**, *8*, 3445–3451, PMID: 26592995.