Published in final edited form as:

J Phys Chem B. 2012 March 1; 116(8): 2365–2375. doi:10.1021/jp209657n.

Structure Refinement of Protein Low Resolution Models Using the GNEIMO Constrained Dynamics Method

In-Hee Park¹, Vamshi Gangupomu¹, Jeffrey Wagner¹, Abhinandan Jain², and Nagara-jan Vaidehi¹

¹Division of Immunology, Beckman Research Institute of the City of Hope, Duarte, CA 91010, USA

²Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA

Abstract

The challenge in protein structure prediction using homology modeling is the lack of reliable methods to refine the low resolution homology models. Unconstrained all-atom molecular dynamics (MD) does not serve well for structure refinement due to its limited conformational search. We have developed and tested the constrained MD method, based on the Generalized Newton-Euler Inverse Mass Operator (GNEIMO) algorithm for protein structure refinement. In this method, the high-frequency degrees of freedom are replaced with hard holonomic constraints and a protein is modeled as a collection of rigid body clusters connected by flexible torsional hinges. This allows larger integration time steps and enhances the conformational search space. In this work, we have demonstrated the use of a constraint free GNEIMO method for protein structure refinement that starts from low-resolution decoy sets derived from homology methods. In the eight proteins with three decoys for each, we observed an improvement of ~ 2 Å in the RMSD to the known experimental structures of these proteins. The GNEIMO method also showed enrichment in the population density of native-like conformations. In addition, we demonstrated structural refinement using a “Freeze and Thaw” clustering scheme with the GNEIMO framework as a viable tool for enhancing localized conformational search. We have derived a robust protocol based on the GNEIMO replica exchange method for protein structure refinement that can be readily extended to other proteins and possibly applicable for high throughput protein structure refinement.

Keywords

Constrained MD; GNEIMO; Structure Refinement; Decoys

INTRODUCTION

The extent of conformational sampling afforded by molecular dynamics (MD) simulations for macromolecules such as proteins and polymers is a critical aspect of various challenging problems such as protein folding, protein structure prediction, protein-ligand docking and protein-protein docking. The main bottleneck in conformational sampling in unconstrained

Supporting Information Available: A summary of total simulation time for all-torsion, hierarchical GNEIMO MDs and unconstrained Cartesian MDs (Table S1); Dihedral angle G-factors of the initial decoys and that of the refined structures (Figure S1); Potential energy distributions of the refinement simulation via all-torsion GNEIMO-REXMD compared to that of the native ensemble via all-torsion GNEIMO MD at 310 K (Figure S2); Potential energy component distribution of decoy1-2cr7A case (Figure S3); The “Freeze and Thaw” (hierarchical) clustering scheme used in the GNEIMO method (Figure S4). This material is available free of charge via the Internet at <http://pubs.acs.org>.

all-atom MD simulations (henceforth referred to as Cartesian simulations) stems from the large number of degrees of freedom in the system of interest. There have been many attempts to enhance the sampling power of MD simulations (1) by using biased potentials such as Accelerated MD (2), Targeted MD (3), Steered MD (4), and Umbrella Sampling (5). Other methods based on the Generalized Boltzmann Ensemble algorithm, such as Replica Exchange MD (REXMD)(6), also enhance conformational sampling.

Constraining high-frequency degrees of freedom via holonomic constraints leads to substantial reduction in the number of degrees of freedom as well as increase in integration time step size. However the equations of motion are in internal coordinates and become computationally tedious to solve (7). We have adapted algorithms from spatial operator algebra to develop the computationally efficient Generalized Newton Euler Inverse Mass Operator (GNEIMO) method for solving the coupled equations of motion in internal coordinates (8)-(10). Other researchers have also adapted and tested this algorithm and incorporated it in software packages (11)-(12).

There is an increasing need for computational methods that refine low resolution homology models based upon sequence similarity to known protein structures (13). Given the potential usefulness of template-based homology models, subsequent structure refinement method is important to improve the model towards high resolution. However, the refinement step is generally skipped in the process of building a model using homology modeling since it generally worsens the resolution of the starting model (14). There are two important factors to be considered in a structure refinement algorithm: (i) an accurate energy function with the native structure as the lowest energy; and (ii) an efficient conformational search method to generate diverse set of conformations, possibly enriching the native like conformations. In this work, we will mainly focus on the second factor: namely the use of MD methods for performing extended conformational sampling for protein structure refinement.

Unconstrained Cartesian simulations are not effective in refining low resolution protein structural models (14). Poon et al. developed an elastic normal-mode-based protocol to improve the quality of low-frequency modes, which is applied to refinement of moderate resolution X-ray structures (16). Torsional angle dynamics methods which utilized the NEIMO algorithm have been used earlier in NMR (implemented in CYANA software) and X-ray (implemented in X-PLOR software) structure refinement applications (17)-(18). Chen et al. demonstrated the applications of torsional angle dynamics to augment conformational sampling of peptides and proteins with examples of folding and structure refinement. However they used restraints from NMR measurements in the refinement procedure (12). In this work, we show that performing all-torsion MD simulations using the GNEIMO method without any restraints leads to a 2 Å increase in the resolution of the homology models and enriches the native like conformations with more effective conformational search.

Besides the default dynamics of all-torsion, one of the major advantages of the GNEIMO method is that it allows the user to freeze and/or thaw any part of the protein as desired (8). This allows one to guide the dynamics along the low-frequency degrees of freedom. We have previously demonstrated the advantage of the “Freeze and Thaw” strategy for studying folding of small proteins (8). In this work, we show that the “Freeze and Thaw” clustering strategy can also be used to develop an effective structure refinement procedure with fewer degrees of freedom than all-torsion dynamics techniques. The focus of the paper is how sampling in the reduced coordinate system employed by GNEIMO is preferable to sampling the complete coordinate system for structure refinement applications.

METHODS

Protein Model Set

We have selected a test set of eight proteins for which a high resolution (ranging from 0.89 Å to 2.5 Å) X-ray crystal or NMR structure exists (28). These eight proteins have different secondary motifs: all- α , α/β , and all- β as shown in Figure 1.

Homology Modeling for Decoy Set Generation

As noted by the protein structure prediction community, the refinement of low-resolution decoys depends on the starting conformation of the decoy (29). We generated a low-resolution decoy set as follows: (i) performed homology modeling using MODELLER (19) choosing a template structure that has 60–70 % sequence identity to the target sequence (i.e. test set in Figure 1); (ii) the resulting top ranking 100 homology models were clustered by structural diversity into 5 clusters. We then chose a representative structure from each cluster with the most secondary content and performed simulated annealing with temperature ranging from 310 K to 1200 K in 50 K increments using all-torsion GNEIMO dynamics to swell the homology model to a lower resolution structure similar to other works (12),(30); (iii) we selected three swollen snapshots from the simulated annealing trajectory with a backbone RMSD range 2–5 Å with respect to the native structure; (iv) we then performed unconstrained Cartesian MD energy minimization using 1000 steps of steepest descent method followed by 1000 steps of conjugate gradient method. We used the AMBER force field (31) for all the simulations and the Generalized Born (GB) solvent model as implemented in the AMBER package (32) with non-bond cutoff of 20 Å. The structure refinement for each of the decoys was performed using the all-torsion GNEIMO method coupled with replica exchange molecular dynamics (REXMD) with eight replicas in the temperature range of 310 K to 415 K with 15 K intervals. Each replica was run for 5–15 ns totaling to 40–120 ns. All simulation times are tabulated in Table S1 of the Supporting Information.

Generalized Newton-Euler Inverse Mass Operator (GNEIMO) – Constrained Dynamics Method

The GNEIMO method is generalized constrained MD method in internal coordinates, for performing multibody dynamics of macromolecules. In the GNEIMO method the high frequency degrees of freedom are kept rigid as holonomic constraints and the macromolecule is modeled as a collection of rigid bodies (of varied sizes) connected by flexible hinges. The rigid bodies also known as “clusters” can vary in size, from a single atom to a whole domain of a protein. The hinges could be modeled with one to six degrees of freedom. When the hinges have one degree of freedom it becomes the torsion about the bond connecting the two clusters. The equations of motion thus become coupled in internal coordinate system and the computational cost of solving the coupled equations of motion in internal coordinates scales as cubic power of the number of degrees of freedom with conventional algorithms. This cost scales linearly with the number of degrees of freedom with the NEIMO algorithm (8)–(10). The current implementation in GNEIMO supports a whole range of dynamic models for the protein, ranging from an all-atom model, to rigid clusters containing a few atoms to rigid clusters constraining motifs or domains of proteins. We have implemented the algorithm for clustering large domains of a protein as rigid body and allow torsional motion between these large rigid bodies (8).

All-torsion GNEIMO Protocol for Refinement

The GNEIMO constrained MD simulations were carried out using the GNEIMO code we had developed (8) and the AMBER99 force field with the Generalized Born/Surface Area

(GB/SA) OBC implicit solvation model (33) using an interior dielectric value of 1.5 for the solute, and exterior dielectric constant of 78.3 for the solvent. We used a solvent probe radius of 1.4 Å for the non-polar solvation energy component of GB/SA. The non-bond forces were smoothly switched off at a cutoff radius of 20 Å (8). Constrained MD simulations were done using all torsional degrees of freedom and other hierarchical clustering schemes. The dynamics was done using a Lobatto integrator, and an integration step size of 5 fs. We implemented the temperature replica exchange MD (REXMD) algorithm into the GNEIMO MD method to expedite simulation time as well as to enhance conformational sampling. Temperature exchange occurred every 400 time steps (corresponding to 2 ps). The eight temperatures chosen for the REXMD simulations were 310–415 K in increments of 15 K.

Hierarchical GNEIMO Method for “Freeze and Thaw” Clustering

One of the major advantages of the GNEIMO method is that it allows the user to freeze and/or thaw any part of the protein as desired. We have demonstrated the advantage of the freeze and thaw strategy in an earlier publication studying folding of small proteins (8). We tested the “Freeze and Thaw” clustering strategy for the mixed motif proteins that contain mixed α -helix and β -sheet motif. Thus for the α/β mixed motif decoy set, we treated either the α -helix or β -sheet motif as a rigid body, freezing only the backbone atoms belonging to it and leaving the side chains sampled as all-torsion. We refer to this strategy as “hierarchical clustering”. The rest of the protein is treated as movable with all-torsion dynamics. Figure S4 (in the supporting information) illustrates the hierarchical clustering strategy used in this work, where the black colored regions of the protein are treated as rigid bodies connected by movable torsions to the rest of the protein.

Unconstrained Cartesian REXMD

To compare the performance of the structure refinement protocol of GNEIMO-REXMD against unconstrained Cartesian REXMD, we used the AMBER10 package (32). We set up the same number of replicas with the same target temperature (310 to 415 K with 15 K interval) as used for the GNEIMO-REXMD simulations. Replica exchange was set to occur every 500 time steps (corresponding to 1 ps), with a time step of 2 fs (see Table S1 for simulation time comparison).

Calculation of RMSD, Percentage Native Contacts, and Number of Hydrogen Bonds

We have calculated several metrics to assess the native likeness of the GNEIMO-REXMD sampled conformations. We have calculated the RMSD of the backbone atoms to the respective X-ray or NMR structures. The RMSD was calculated using snapshots from the whole REXMD trajectory of eight replicas using the utility called “rms” in the AmberTool1.4 analysis suite (34). To examine whether the refinement in the models structure comes from the secondary structure regions or the loop regions, we further calculated the backbone RMSD for the secondary structure region as in the native structure, and the backbone RMSD for the whole structure including the loops and termini, both with respect to native structures.

To measure the refinement in the overall packing and fold, we calculated the percent of native contacts made in the GNEIMO and unconstrained MD simulation trajectories using the “cmap” protocol in the bio3d R statistical programming package. We calculated the $N \times N$ matrices consisting of pair-wise $\text{Ca}(i)\text{-Ca}(j)$ atom distances for the native structures, and for the whole trajectories of GNEIMO-REXMD, where N is a number of residues in a protein, and i and j are residue indices. A pair-wise $\text{Ca}(i)\text{-Ca}(j)$ distance shorter than 8 Å is considered a contact and given an index value of 1. Hence the calculation of the contact map results in the construction of an $N \times N$ contact matrix consisting of 0's (no contact farther

than the 8 Å cutoff or within the 4 Å neighbor cutoff in sequence) and 1's (a contact within the 8 Å cutoff). We then considered each C α pair in the simulation snapshots to be a contact if it was both within the proper distance, as well as present on the contact map of the native structure (35). We have calculated the percentage native contacts as the dot product of the N $\times$ N contact matrix calculated for the native structure and the corresponding matrix for each snapshot.

To generate the contact maps, and the residue-wise distance improvement toward the native structure, we computed the distance matrix as the difference between refined structure and the native structure. We used “*dist*” protocol of AmberTools1.4 to obtain the pair-wise C α distance matrix followed by distance matrix calculations and matrix plots using R. Finally, to count the total number of hydrogen bonds in each snapshot of the simulations, we used “*findhbond*” command in Chimera program (25).

Quantitative Assessment for the Refined Structure via Residue-wise PCA and Essential Dynamics

We carried out principal component analysis (PCA) on the GNEIMO-REXMD trajectories for all temperatures using the bio3d R statistical programming package (36). PCA was performed to quantify the extent of movement towards the native like structure in various parts of the protein. We calculated the N $\times$ N distance covariance matrix consisting of variations of Cartesian coordinates of C α -atoms over time (an entire trajectory) with respect to the average coordinates, where N is a number of atoms in the protein model. A set of Cartesian coordinates corresponding to the two most significant principal modes (namely, PC1 and PC2) and the average structure of the trajectory were extracted. To demonstrate the direction of the movement in these vectors PC1 and PC2, we generated a series of frames of atomic displacements by interpolating the snapshots in the trajectory from the mean structure along PC1 and PC2.

Refinement of the Torsion Angles

We assessed the extent of the structure refinement in the torsional angle space by measuring the deviation of the torsion angles from the high-resolution structures for backbone dihedrals ϕ (C $_i$ -N $_{i+1}$ -C α_{i+1} -C $_{i+1}$), ψ (N $_i$ -C α_i -C $_i$ -N $_{i+1}$), and ω (C α_i -C $_i$ -N $_{i+1}$ -C α_{i+1}) as well as side-chain dihedral angles χ through dihedral G-factor (20),(37). The dihedral G-factor is a measure of normality of the dihedral angles (ϕ - ψ , χ and ω) in the refined structures. This measure is particularly useful to assess the quality of structure refinement in a case where the native structure is not known. We calculated it in this work for proteins with known X-ray crystallographic or NMR structures to check if this measure varies similarly to the backbone RMSD and percentage native contact measures with respect to native structure. Details of this dihedral G-factor are published in references (38)-(40).

RESULTS and DISCUSSION

Structure Refinement by All-torsion GNEIMO-REXMD

We observed that the larger the RMSD of the starting decoy, the better the extent of refinement achieved. For the all α -helical motif decoys (left columns in Table 1), an average refinement of about 2.0 Å, 1.4 Å, and 0.6 Å was achieved for the decoys with initial RMSD 5.0 Å, 4.0 Å, and 3.0 Å away from the native structure, respectively. Similarly, for the mixed α/β and all- β motif decoys (right columns in Table 1), an average RMSD improvement was 1.9 Å, 1.6 Å, and 1.0 Å for the decoys with starting RMSD of about 4 Å, 3 Å, and 2 Å, respectively.

Improvement in RMSD to the native structure—As described in detail in the Methods section, we have used the temperature replica exchange (REXMD) in the GNEIMO method. We selected eight proteins with varying α -helical and β -sheet content and known native structures, experimentally resolved by either X-ray or NMR methods (Figure 1) for testing the GNEIMO-REXMD method for structure refinement. We used the homology based method MODELLER (19) to derive three decoys per protein as starting models for refinement. After GNEIMO-REXMD simulations, the backbone RMSD in coordinates with respect to the native structure was calculated for the whole protein and for the regions with secondary structure (see Method section). Table 1 shows the RMSD of the starting decoy for each protein and the best refined structure after the GNEIMO refinement procedure.

For example, decoys of 1ab1, a β -sheet containing protein, refined to very high resolution - up to 0.6 Å in RMSD of structured region. Although the refinement was greater for the α -helical and β -strand regions, showing a growth in the secondary structure regions (inset structures in Figure 2), it was not confined to the secondary structure regions alone. The loop regions showed a substantial improvement as seen in the overall packing of the whole protein structure in Table 1. Even without the use of additional constraints, the all-torsion GNEIMO-REXMD simulations led to substantial improvement in structure refinement over unconstrained Cartesian simulations.

To study whether the GNEIMO trajectory contains an enriched population of near native structures in comparison to the starting decoy, we analyzed the RMSD of the entire GNEIMO-REXMD trajectory (for all eight replicas). Figure 2 shows the population density distribution with RMSD of all the structures over the entire trajectory. The RMSD was calculated with respect to the initial decoy structure. It is seen that the population density of RMSD shifted toward the native like (RMSD < 3 Å) structures (indicated by the solid arrow) compared to the initial decoy RMSD (indicated in dotted blue line) for three decoys each in seven out of eight proteins. An exception where the population density shifts away from the starting decoy was for the protein 2f3nA. This is due to the unavoidable gap in the sequence alignment used for homology modeling this protein. The misalignment leads to a very low resolution structural model in the amino terminus of this protein that does not get resolved well during the refinement.

Improvement in percentage native contacts—Another measure to assess the extent of structure refinement, particularly the packing of the tertiary structure and formation of β -sheet motif, is the number of non-local contacts (i.e. long-range contact, defined as a contact made between $|i-j| \geq 4$, where i and j are residue numbers) that is observed in the native structure. The percentage of native contacts is also shown in Table 1. The average percentage of native contacts improved by about 30 % for the all- α helical and mixed α/β motif decoy set; 47 % for the 1bhp decoy set; 7 % for the all β -sheet motif (protein model 1cskA).

Furthermore, to gauge the overall similarity in the tertiary fold of the refined structure to the native structure, we constructed a C α -atom based pair-wise distance matrix for the initial decoys and the corresponding refined structure. We chose the best refined structures having the lowest backbone RMSD for the secondary structure region with respect to the native structure as listed in Table 1. We then computed their pair-wise distance from the native structure. This method forms a symmetric matrix in which the x- and y-axes correspond to C α -atom numbers (i.e. residue numbers). Short-range contacts lie near the diagonal, while long-range contacts are located in the far off-diagonal. As shown in Figure 2A and B, as the refined structure comes closer to the native structure, the color map becomes blue; as the structure moves away from the native structure, the color map becomes red. Overall, upon

refinement via all-torsion GNEIMO, long-range contacts were captured, thus the overall structure came closer to the native structure. Substantial refinement indicated by change of colors from yellow/red in the initial decoy map (left matrix) to the blue in the refined map (right matrix), occurred as a result of the packing of the entire structure followed by growth of secondary structure. In Figure 2B, for the mixed α/β motifs (1ab1, 1bxyA, and 1bhp), the loop motifs in the decoys became β -sheet motifs in the refined structures; also, their short β -strand motifs in the decoys grew to similar length as in the native structures (similar growth of β -strand was clearly observed in all- β motif in 1cskA decoys). Similarly, for the all decoy set of all- α (Figure 2A) and mixed α/β motifs (Figure 2B), the partially formed α -helix motifs were grown to a large extent.

In the distance matrix map of the refined structure (right matrix in Figure 2A and B), however, we still observed the persistent light blue region (still deviated from the native structure). This region mainly corresponds to N and C-termini as well as loop motifs. Due to their intrinsic flexibility in the dynamics of the structure, those regions have high B-factors even in the X-ray crystallographic structures. Given the dynamic natures of those floppy motifs, all-torsion GNEIMO refined the structures to the native-like structure by capturing the native-like packing and growth of secondary structures. Upon refinement the reduction in C α -atom distance was substantial as shown in the Figure 2, from more than 20 Å to less than 8 Å (usually used for cutoff distance of native contact).

We also calculated the dihedral G-factor using the PROCHECK program (20)-(21), to assess the extent of refinement in the main chain and side chain torsion angles. Details of the calculation of the dihedral G-factor are given in the supporting material (Figure S1). The dihedral G-factor is a measure of how well the backbone and side chain torsional angles fall within the acceptable regions of the torsional angle space observed in the high-resolution structures in the Protein Data Bank. Thus it is a useful metric to assess the quality of structural refinement for cases where there is no prior knowledge of the native structure. Dihedral angle G-factors for the initial decoys and best refined structures in the 24 decoy set are shown in Figure S1. In all cases, the total G-factor of the refined structure improved to less than -0.5 except decoys of 1cskA compared to that of the initial decoy, as indicated by the change in the red bars became shorter blue bars in the bottom row of each decoy panel. The majority of the improvement in the total dihedral G-factor comes from the backbone dihedrals with some improvement in the side chain torsions.

Enrichment of native-like structures in the GNEIMO-REXMD trajectories—To assess the enrichment of native-like structures in the ensemble of conformations generated by the all-torsion GNEIMO-REXMD simulations we have calculated percentage native contacts and number of hydrogen bonds for all the structures in the entire trajectory. We then compared the distributions of the percentage native contacts and number of hydrogen bonds for the entire trajectories. It is evident from Figure 3, by measure of all the three metrics a large proportion of the conformational ensemble is close to the native structure compared to the starting decoy, denoted by dotted vertical line. For the refinement of decoy1 of 2f3nA (in Figure 4A) the percentage native contacts and hydrogen bonds showed substantial enrichments even though this was not reflected in the RMSD.

Potential energy reduction upon refinement and its correlation with native-likeness—We analyzed the all-torsion GNEIMO refinement trajectory of decoy1-2cr7A as a case study to see if the potential energy becomes more favorable as the native contacts are formed and the structure gets refined. To understand the dynamics of the NMR structure, we performed all-torsion GNEIMO MD at 310 K with time steps of 5 fs, starting from the NMR structure of the protein 2cr7A. This trajectory generates a native structure ensemble that can be used for comparison with the energies of the ensemble generated by the all-torsion

GNEIMO-REXMD refinement simulations. The black curve in the Figure S2A shows the distribution of the native state potential energy obtained from GNEIMO MD at 310 K. The potential energy of the starting decoy is denoted by red dotted line. The potential energy distribution of all the eight replicas obtained from all-torsion GNEIMO-REXMD simulations is shown in various colors in Figure S2A. It is observed that upon refinement, the potential energy decreased toward the native ensemble energies (black curve). The distribution of various components of the potential energy is shown in the supporting material in Figure S3.

To check if the potential energy can be used to identify the native like structures we have plotted the potential energies and various components of the potential energy as a function of RMSD to the native structure for the protein 2cr7A in Figure 4. The potential energies of all snapshots of all-torsion GNEIMO MD at 310 K starting from the native structure are shown as pink dots and those generated during the all-torsion GNEIMO-REXMD refinement simulations are shown in blue dots. The top panel (“PE_noHighF” in Figure 4) is the potential energy excluding the contribution from the high frequency bond length and bond angle energy components. The van der Waals energy (vdw) and the torsional energy, that dominantly contribute to packing and strain respectively in the structures, showed a funnel shape correlation with RMSD, whereas electrostatic energy (ELS) spread widely in energy, resulting in no clear correlation with RMSD. The limit of the refinement of decoyl-2cr7A at the lowest RMSD of 2.3 Å and the maximum of 82 % native contacts may have also weakened the correlation between potential energy and RMSD. Therefore one could use van der Waals and torsional forces while performing GNEIMO-REXMD structure refinement to achieve a better correlation between energy and native-likeness. The focus of this work is, however, to examine the use of GNEIMO as a conformational search method in conjunction with existing force fields, for protein structure refinement. In principle the GNEIMO conformational search MD method can be used with any type of force field fitted for protein structure refinement (22)-(24).

Analysis of the ensemble of conformations generated during the refinement—

To identify the refined regions of the protein structure based on the entire trajectory sampled by all-torsion GNEIMO-REXMD, we analyzed the ensemble of GNEIMO snapshots for all eight replicas using principal coordinate analysis (PCA). We then extracted a set of Cartesian coordinates corresponding to the two most significant principal coordinate modes (namely, PC1 and PC2). Figure 5A and B show the comparison of the best refined structure to that obtained by PCA. The width of each C α -atom in Figure 5B indicates the extent of variation/fluctuation with respect to the average structure of the entire GNEIMO-REXMD trajectory.

As can be seen from Figure 5, the GNEIMO-REXMD trajectories contain structures refined both in the packing interactions and the secondary structure regions. PC1 in Figure 5B shows movement in the amino terminus while PC2 represents movement in the loop connecting the two helices, as marked by the dotted ovals. These movements are overlaid on the time-averaged structure in Figure 5B. As marked with dotted ovals, the PC1 mode shows the packing and growth of the blue-helix region against the rest of the protein while the PC2 shows both packing and growth of the green-helix region. Thus the structures generated during the GNEIMO refinement procedure are all enriched towards more refined structures than the starting decoy.

Comparison of Refinement using GNEIMO MD with Unconstrained Cartesian MD

To compare the performance of all-torsion GNEIMO MD to unconstrained Cartesian MD in structure refinement, we performed unconstrained Cartesian REXMD simulations under the same conditions as the previously discussed GNEIMO simulations, except that a time-step

size of 2 fs was used instead of the 5 fs used with GNEIMO. It is seen in Figure 6 that the population density of the structures sampled by unconstrained Cartesian MD simulations shifts away from the starting decoy thus leading to little or no structure refinement compared to the all-torsion GNEIMO simulations.

Using decoy1-2cr7A as a case study, we also compared the conformational space sampled by all-torsion REXMD and unconstrained Cartesian REXMD using the PCA from the combined trajectories; the two REXMD trajectories have the same duration of simulation time to allow for fair comparison. As shown in Figure 7A, conformations sampled via all-torsion GNEIMO-REXMD completely overlapped with the native structure of 2cr7A denoted by pink, filled circles. In contrast, in Figure 7B, conformations sampled by unconstrained Cartesian REXMD covered a wide range of PC1 and PC2 axes, and only a small fraction of snapshots (shown as cyan dot in Figure 7B) of unconstrained Cartesian REXMD overlapped with native structure.

In addition, we generated a series of frames of atomic displacements from the trajectories, by interpolating from the mean structure along PC1 (inset structures in Figure 7A and B). All-torsion GNEIMO-REXMD refined the structure of the starting decoy toward native-like structure, whereas the conformations generated in the unconstrained Cartesian REXMD simulations (inset in Figure 7B) unraveled the existing secondary structure elements in the decoy structure, and pushed the ensemble away from the native-like structure, perhaps due to friction stemming from high-frequency modes. Owing to constraining high-frequency modes, conformational search in all-torsion GNEIMO-REXMD was enhanced towards native-like conformations while maintaining stable structures at high temperatures.

Structure Refinement by “Freeze and Thaw” (Hierarchical) Clustering Scheme

One of the major advantages of the GNEIMO method is that it allows the user to freeze and/or thaw any desired part of the protein (8). This allows one to guide the dynamics along the low-frequency degrees of freedom. We had demonstrated the advantage of “freeze and thaw” clustering in reference (8) for studying the folding of small proteins. It allows dynamic models of proteins that range from unconstrained Cartesian MD, to coarser ones with large sections of the protein frozen as rigid bodies while optimizing the torsions connecting these rigid bodies (also known as clusters). We can therefore use various clustering strategies and study their effect on structure refinement. The all-torsion GNEIMO method, where covalent bond lengths and angles of high-frequency degrees of freedom can be frozen, was discussed in the previous section. In this section we describe the “Freeze and Thaw” clustering scheme (also called Hierarchical clustering) for freezing large clusters containing secondary structure elements in the decoy for structure refinement.

We used various hierarchical clustering schemes as described in the supporting information in Figure S4. Given the difficulty in formation of β -sheet motifs, which require capturing of non-local contacts via a near perfect match of inter-residue contacts, we froze the backbone atoms of β -strand motifs of varying length as it is in the decoy. In case of decoys without β -strand motif being formed; we froze the backbone atoms of α -helix motif instead. Other than the frozen motifs, the rest of regions of the decoy were treated as all-torsion GNEIMO. We used the Chimera program (25) to identify the secondary structure in the decoy. However, to avoid artificial distortion in rigid clusters, we fixed a small portion, and not the whole predicted secondary structure regions.

We have compared the ensemble of conformations generated by all-torsion GNEIMO-REXMD, Hierarchical GNEIMO-REXMD and unconstrained Cartesian REXMD in Figure 8. The overall refinement results via the hierarchical GNEIMO method were comparable to those from all-torsion GNEIMO, although the RMSD (with respect to the native structure)

of the secondary structure regions for the hierarchical GNEIMO results was not as good as that of the all-torsion GNEIMO. The population density in the hierarchical simulations did show a shift towards the native structure. Again, similar to the Figure 6 (decoys with all- α motifs), unconstrained all-atom REXMD showed the conformations that are further off from the native structure.

We compared the hierarchical schemes of fixing α -helices (decoy1-1ab1, decoy3-1ab1, and decoy1-1bhp) and fixing β -strands (all other decoys). Fixing shorter length of β -strands showed similar refinement as the all-torsion GNEIMO; whereas fixing α -helix results showed worsening of the RMSD compared with all-torsion GNEIMO. These hierarchical GNEIMO refinement results imply that choosing the right portion of the structure to cluster is important in structure refinement. We are currently developing automated schemes of clustering that will be the subject of another publication.

Nevertheless, as shown in Table 2, using the hierarchical clustering scheme in GNEIMO reduces the simulation time needed to achieve similar levels of structure refinement as all-torsion GNEIMO. We observed that fixing the β -strand motif in the decoy would be beneficial to achieve a refined structure in a short simulation time. For example, the time gain was up to 250 times faster via hierarchical GNEIMO in the decoy3-1bxyA case.

To examine how rapidly the various methods (all-torsion GNEIMO, hierarchical GNEIMO, and unconstrained MD) explore the available conformational space in terms of optimizing the native-likeness, we have analyzed the time series of the enrichment of the native like structures. We calculated the backbone RMSD with respect to native structure from eight replicas of REXMD at various time points by dividing the total time into five time windows shown as orange bar in the Figure 9. It is evident from these enrichment plots that the enrichment of the near native structures happened early in the GNEIMO-REXMD simulations for the all-torsion as well as the hierarchical simulations. The all-atom unconstrained Cartesian dynamics with implicit solvent model disrupts the starting decoy early in the simulations.

CONCLUSIONS

Protein structure refinement with unconstrained Cartesian MD simulations with explicit solvent did not show a positive trend towards refinement (14). We have observed that unconstrained Cartesian MD generates large clusters of structures that indicate the unraveling and reformation of secondary structure, hence it is not effective in searching in the low-frequency degrees of freedom. Starting from decoy structures of varied resolution generated from the homology method MODELLER, we have clearly demonstrated that the GNEIMO clustering scheme effectively enriches the conformational search in the low-frequency modes. This leads to refined structures without using experimental information in the refinement procedure. More importantly we derived a robust protocol to generate the decoys using homology methods and subsequent refinement using GNEIMO MD methods, which is very useful for the larger protein modeling community. We also showed that using the “Freeze and Thaw” clustering strategy can yield similar magnitudes in structure refinement with less computational time.

This opens up the possibility of using GNEIMO-REXMD methods, in conjunction with force fields better designed for this purpose and/or using known experimental distance restraints, as a high-throughput structural refinement method. We have not examined the effect of force fields in this study since the focus of this work was to examine the effect of various dynamic molecular models on the structure refinement. We will incorporate the other body of work available on effective force fields for structure prediction/refinement in

the near future (22)-(24),(26)-(27). The GNEIMO-REXMD method also needs to be tested for refinement of larger proteins and protein complexes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This project has been supported by Grant Number RO1GM082896 from the National Institutes of Health. We thank Dr. Karin Remington and Dr. Paul Brazhnik for their support and encouragement. We thank Simbios for providing us with the GB/SA solvation module, Mark Friedrichs for his help with validating our GB/SA module. The Simbios software was made freely available on <https://simtk.org/home/openmm> by the Simbios NIH National Center for Biomedical Computing. We used the force module of LAMMPS, an open-source code available from LAMMPS WWW Site. Part of the research described in this paper was performed at the Jet Propulsion Laboratory (JPL), California Institute of Technology, under contract with the National Aeronautics and Space Administration.

References

1. Zuckerman DM. *Annu Rev Biophys.* 2011; 40:41–62. [PubMed: 21370970]
2. Hamelberg D, Mongan J, McCammon JA. *J Chem Phys.* 2004; 120:11919–11929. [PubMed: 15268227]
3. Schlitter J, Engels M, Kruger P. *J Mol Graphics Modell.* 1994; 12:84–89.
4. Isralewitz B, Baudry J, Gullingsrud J, Kosztin D, Schulten K. *J Mol Graph Modell.* 2001; 19:13–25.
5. Torrie GM, Valleau JP. *J Comput Phys.* 1977; 23:187–199.
6. Sugita Y, Okamoto Y. *Chem Phys Lett.* 1999; 314:141–151.
7. Abagyan RA, Argos P. *J Mol Biol.* 1992; 225:519–532. [PubMed: 1593634]
8. Balaraman GS, Park IH, Jain A, Vaidehi N. *J Phys Chem B.* 2011; 115:7588–7596. [PubMed: 21591767]
9. Jain A, Vaidehi N, Rodriguez G. *J Comput Phys.* 1993; 106:258–268.
10. Vaidehi N, Jain A, Goddard WA. *J Phys Chem.* 1996; 100:10508–10517.
11. Schwieters CD, Clore GM. *J Magn Reson.* 2001; 152:288–302. [PubMed: 11567582]
12. Chen J, Im W, Brooks CL. *J Comput Chem.* 2005; 26:1565–1578. [PubMed: 16145655]
13. MacCallum JL, Perez A, Schnieders MJ, Hua L, Jacobson MP, Dilly KA. *Proteins Struct Funct Bioinf.* 2011;1097-0134.
14. Rigden, DJ. *From Protein Structure to Function with Bioinformatics.* Springer; 2009.
15. Lee MR, Tsai J, Baker D, Kollman PA. *J Mol Biol.* 2001; 313:417–430. [PubMed: 11800566]
16. Poon BK, Chen X, Lu M, Vyas NK, Quirocho FA, Wang Q, Ma J. *Proc Natl Acad Sci.* 2007; 104:7869–7874. [PubMed: 17470791]
17. Rice LM, Brünger AT. *Proteins.* 1994; 19:277–290. [PubMed: 7984624]
18. Güntert C, Mumenthaler C, Wüthrich K. *J Mol Biol.* 1997; 273:283–298. [PubMed: 9367762]
19. Sali A, Potterton L, Yuan F, van Vlijmen H, Karplus M. *Proteins.* 1995; 23:318–326. [PubMed: 8710825]
20. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. *J Appl Crystallogr.* 1993; 26:283–291.
21. Laskowski RA, Rullmann JA, Macarthur MW, Kaptein R, Thornton JM. *J Biomol NMR.* 1996; 8:477–486. [PubMed: 9008363]
22. Chopra G, Summa CM, Levitt M. *Proc Natl Acad Sci U S A.* 2008; 105:20239–20244. [PubMed: 19073921]
23. Simons KT, Kooperberg C, Huang E, Baker D. *J Mol Biol.* 1997; 268:209–225. [PubMed: 9149153]
24. Kussell E, Shimada J, Shakhnovich EI. *Proc Natl Acad Sci U S A.* 2002; 99:5343–5348. [PubMed: 11943859]

25. Huang CC, Couch GS, Pettersen EF, Ferrin TE. Chimera: Pacific Symposium on Biocomputing. 1996; 1:724.
26. Katritch V, Totrov M, Abagyan R. J Comput Chem. 2003; 24:254–265. [PubMed: 12497604]
27. Arnaudova YA, Abagyan RA, Totrov M. Proteins. 2011; 79:477–498. [PubMed: 21069716]
28. Wroblewska L, Skolnick J. J Comp Chem. 2007; 28:2059–2066. [PubMed: 17407093]
29. Jagielska A, Wroblewska L, Skolnick J. Proc Natl Acad Sci U S A. 2008; 105:8268–8273. [PubMed: 18550813]
30. Robustelli P, Kohlhoff K, Cavalli A, Vendruscolo M. Cell. 2010; 18:923–933.
31. Wang J, Cieplak P, Kollman PA. J Comput Chem. 2000; 21:1049–1074.
32. Case, DA., et al. AMBER10. University of California; San Francisco: 2008.
33. Onufriev A, Bashford D, Case DA. Proteins. 2004; 55:383–394. [PubMed: 15048829]
34. Case DA, Cheatham TE III, Darden T, Gohlke H, Luo R, Merz KM Jr, Onufriev A, Simmerling C, Wang B, Woods R. J Computat Chem. 2005; 26:1668–1688.
35. Gromiha M. Prog Biophys Mol Biol. 2004; 86:235–277. [PubMed: 15288760]
36. Grant BJ, Rodrigues AP, ElSawy KM, McCammon JA, Caves LS. Bioinformatics. 2006; 22:2695–2696. [PubMed: 16940322]
37. Abseher R, Nilges M. Proteins: Structh, Funct, Bioinf. 2000; 39:82–88.
38. Engh RA, Huber R. Acta Cryst. 1991; 47:392–400.
39. Yu N, Li X, Cui G, Hayik S, Merz KM. Protein Science. 2006; 15:2773–2784. [PubMed: 17132861]
40. Priestle JP. J Appl Cryst. 2003; 36:34–42.

Figure 1. Protein models used in the refinement study: all α -helical proteins (A); mixed α/β and all β -sheet motifs (B). Structures are color-coded from blue to red in the order of N-termini to C-termini.

Figure 2.

Refinement of eight protein decoy structures listed in Table 1. The population density of the RMSD of all the structures in the entire GNEIMO-REXMD trajectory with respect to the native structure is shown in the figures to the left of each figure. Ca -atom based distance matrix maps of the initial decoys and the refined structures (lowest backbone RMSD for the secondary structure region) with respect to the native structures (right).

Figure 3. Enrichment of native-like structures in the GNEIMO-REXMD trajectories is demonstrated in the population density distributions of the backbone RMSD with respect to native structure for the secondary regions, enrichment of percentage of native contacts, and total number of hydrogen bonds. The dotted line denotes the position of the starting decoy. The order of the decoy set is the same as shown in Figure 2.

Figure 4. Correlation between backbone RMSD and total potential energy excluding high-frequency modes such as bonds and angles (PE_noHighF panel), total potential energy (PE panel), and various energy components (van der Waals, torsional, electrostatic and GBSA from top to bottom panels) for de-coy1-2cr7A at target temperature of 310 K.

Figure 5. Initial decoy and the best refined structures depicted both in ribbon representation (A). Residue-wise attributions to PC1 and PC2 modes (essential dynamics) in C α -trace representation, analyzed from the entire trajectory of all-torsion GNEIMO-REXMD (B). As marked with dotted ovals, the PC1 mode shows the packing and growth of the blue-helix region; the PC2 also shows both packing and growth of the green-helix region. Structures are color-coded from blue to red in the order of N-termini to C-termini.

Figure 6.

Comparison of structure refinement performance of all-torsion GNEIMO-REXMD to unconstrained Cartesian REXMD for the four all- α motif decoys. All-torsion GNEIMO refinements showed an RMSD shift (shown in blue arrows) in the secondary structure region, from the initial decoy RMSD (dotted black lines) toward native structure in blue lines in contrast to the unconstrained Cartesian REXMD in pink lines. We used a logarithmic scale on the x-axis of RMSD to zoom into the low RMSD region.

Figure 7. PCA projection onto the PC1-PC2 axes constructed from the combined trajectory of all-torsion and unconstrained Cartesian REXMD of decoy1-2cr7A. The projected snapshots are color-coded based on the RMSD (in Å) based k-mean clustering. The magenta dot denotes the 2cr7A native structure. Inset structures in (A) and (B) showed the corresponding PC1 modes in Ca-atom trace representation.

Figure 8.

The refinement results via the “Freeze and Thaw” (hierarchical) clustering scheme (see Figure S4) using the GNEIMO method (blue lines) in comparison to all-torsion GNEIMO (pink lines) and unconstrained Cartesian MD (green lines) for decoy1 of 1ab1 (A), 1bxyA (B), 1bhp (C) and 1cskA (D).

Figure 9.

Comparison of enrichment of the near native structures (in terms of backbone RMSD) at various time points for decoy1-1bxyA case via all-torsion GNEIMO (A), hierarchical GNEIMO (B) and unconstrained Cartesian MD (C). The total time was divided into five time windows shown as orange bar in each sub-panel.

Table 1

Structure Refinement Results Achieved by All-torsion GNEIMO-REXMD

PDB model	Backbone RMSD (Å) for secondary structural region ^a	Backbone RMSD (Å) for whole structure ^a	Percentage Native contacts (%) ^a	PDB model	Backbone RMSD (Å) for secondary structural region ^a	Backbone RMSD (Å) for whole structure ^a	Percentage Native contacts (%) ^a
	Init. Decoy --> Refined	Init. Decoy --> Refined	Init. Decoy --> Refined		Init. Decoy --> Refined	Init. Decoy --> Refined	Init. Decoy --> Refined
1b72A (all-α motif)							
decoy1	4.3 --> 2.4	5.1 --> 2.7	48 --> 78	decoy1	4.0 --> 0.6	4.1 --> 2.1	39 --> 81
decoy2	3.5 --> 1.8	4.1 --> 2.3	52 --> 78	decoy2	2.8 --> 0.6	2.8 --> 1.0	58 --> 90
decoy3	2.7 --> 1.7	3.1 --> 2.1	55 --> 83	decoy3	2.0 --> 0.6	2.1 --> 1.2	68 --> 90
Avg. Δ ^b	1.5	1.7	28	Avg. Δ	2.3	1.5	32
1r69 (all-α motif)							
decoy1	4.6 --> 2.4	5.0 --> 3.0	49 --> 76	decoy1	3.6 --> 1.6	4.2 --> 2.2	42 --> 74
decoy2	3.8 --> 2.8	4.0 --> 3.5	51 --> 69	decoy2	2.9 --> 1.1	3.5 --> 1.8	52 --> 84
decoy3	3.0 --> 2.2	3.1 --> 2.6	57 --> 75	decoy3	2.4 --> 1.1	3.1 --> 1.8	61 --> 88
Avg. Δ	1.4	1.0	21	Avg. Δ	1.7	1.7	30
2c77A (all-α motif)							
decoy1	4.6 --> 2.3	4.9 --> 2.7	38 --> 82	decoy1	3.2 --> 1.0	4.0 --> 1.7	27 --> 75
decoy2	3.7 --> 2.4	4.1 --> 2.8	42 --> 79	decoy2	2.8 --> 1.1	3.3 --> 1.5	37 --> 82
decoy3	3.0 --> 2.2	3.1 --> 2.3	53 --> 83	decoy3	2.7 --> 1.0	2.8 --> 1.4	38 --> 86
Avg. Δ	1.5	1.4	29	Avg. Δ	1.7	1.9	47
2f3nA (all-β motif)							
decoy1	3.2 --> 2.2	5.0 --> 3.9	44 --> 82	decoy1	2.9 --> 1.3	3.5 --> 2.4	79 --> 91
decoy2	3.0 --> 2.1	4.5 --> 2.8	48 --> 80	decoy2	2.2 --> 1.4	3.0 --> 2.1	86 --> 91
decoy3	2.5 --> 2.2	3.1 --> 2.9	52 --> 70	decoy3	2.0 --> 1.3	2.7 --> 2.4	88 --> 91
Avg. Δ	0.7	1.0	29	Avg. Δ	1.0	0.8	7

^aThe best RMSD and best percentage native contact values are tabulated. See the method section for the calculation of backbone (Cα, C, and N atoms) RMSD of different regions and percentage of native contacts.

^b Average improvement in RMSD and percentage native contacts over all three decoys of the same protein model.

Table 2

Comparison of the best refined structure in RMSD and the simulation time via all-torsion and hierarchical GNEIMO

	All-torsion GNEIMO		Hierarchical GNEIMO ^a	
	Best RMSD ^b (Å)	Number of Time Step ^c (Time in ns)	Best RMSD ^b (Å)	Number of Time Step ^c (Time in ns)
1ab1 (α/β motif)				
Decoy1 ^d	0.6	1,480,000 (7.4)	2.2	440,000 (2.2)
Decoy2	0.6	2,080,000 (10.4)	0.6	1,340,000 (6.7)
Decoy3 ^d	0.6	1,000,000 (5.0)	0.8	2,200,000 (11.0)
1bhp (α/β motif)				
Decoy1 ^d	1.0	1,640,000 (8.2)	2.2	2,120,000 (10.6)
Decoy2	1.1	2,300,000 (11.5)	1.5	180,000 (0.9)
Decoy3	1.0	1,480,000 (7.4)	1.0	20,000 (0.1)
1bxyA (α/β motif)				
Decoy1	1.6	20,000 (0.1)	1.9	20,000 (0.1)
Decoy2	1.1	780,000 (3.9)	1.3	60,000 (0.3)
Decoy3	1.1	1,480,000 (7.4)	1.3	6,000 (0.03)
1cskA (all- β motif)				
Decoy1	1.3	20,000 (0.1)	1.9	2,000 (0.01)
Decoy2	1.4	880,000 (4.4)	1.9	640,000 (3.2)
Decoy3	1.3	1,000,000 (5.0)	1.7	20,000 (0.1)

^aThe hierarchical clustering scheme used is demonstrated in Figure S4.

^bThe best backbone RMSD for the secondary structure region; the all-torsion RMSD is the same as that shown in Table 1.

^cThe number of time steps to achieve the best RMSD via all-torsion and hierarchical GNEIMO-REXMD.

^dFor this decoy, the α -helix motif was fixed in the hierarchical GNEIMO simulations.