Protein Conformational Search with Geometric Projections

Brian Olson[1], S. Farid Hendi[1], and Amarda Shehu[1,2*]
[1]*Department of Computer Science*
[2]*Department of Bioinformatics and Computational Biology*
George Mason University
Fairfax, VA, 22030, USA
bolson3@gmu.edu, shendi@gmu.edu, amarda@gmu.edu
∗Corresponding Author

***Abstract*—Protein structure prediction remains a central challenge in computational structural biology. Even at the coarse-grained level of detail, the protein conformational space is vast, and available energy functions contain many false local minima. In order to effectively characterize this space, a conformational search must sample a geometrically-diverse set of low-energy conformations. Our recently published FeLTr framework achieves this goal by employing a low-dimensional geometric projection layer to bias conformational sampling towards unexplored regions of the search space. In this work we present a new geometric projection layer based on the effective connectivity measure, which encapsulates interatomic distances within a conformation. Extensive analysis indicates that effective connectivity allows equipping the high-dimensional conformational search with an effective projection layer. On several target proteins, this layer improves significantly over our previous work, resulting in sampling of conformations with significantly lower lRMSDs to the known native structure.**

***Keywords*-protein native structure; near-native conformations; probabilistic conformational search; geometric projection; effective connectivity.**

I. Introduction

In the early 1970s it was discovered that the amino-acid sequence of a protein directly determines its biologically-active structure in vitro [1]. Forty years later, determination of this "native" structure remains one of the central challenges in computational structural biology [2]. The native structure of a protein can be experimentally determined though techniques, such as X-ray Crystallography, Nuclear Magnetic Resonance (NMR), and Cryo-electron Microscopy. These techniques are expensive, time-consuming, and are not keeping pace with the exponential growth of protein sequences deposited in protein databases. Computational methods present a complementary approach to elucidating native structures. Effective computational methods will not only advance our knowledge of protein function but also lead to novel drug development and assist in our understanding of protein-protein interactions in supramolecular assemblies [3]–[5].

The three-dimensional structure of a protein is determined by the dihedral angles defined over bonds connecting atoms in a protein chain. A protein conformation represents a particular arrangement of these angles. Each amino acid adds four or more dihedral angles and even small proteins contain up to 100 amino acids. The result is a vast conformational search space with many degrees of freedom. Summing over interatomic interactions in a particular conformation results in a potential energy value. Associating each protein conformation with a potential energy value results in an energy surface that underlies the protein conformational space. The native state consists of protein conformations associated with the lowest energies in the energy surface [6].

A popular strategy for reducing the complexity of the search space is to reduce the effective degrees of freedom through a coarse-grained representation of a conformation modeling only the protein backbone with two degrees of freedom per amino acid. The functions available to compute this coarse-grained energy surface, however, are semi-empirical, leading to a rugged energy landscape potentially rich in false local minima. The vast high-dimensional conformational space and its rugged energy surface make the problem of computing native protein conformations a challenging one [7]–[10]. and For this reason, Monte Carlo-based sampling techniques remain important in modern protein structure prediction algorithms [3], [11]–[14].

An emerging template to navigate this vast conformational space in search of the native structure splits the search into stages based on the level of granularity [3], [15]–[19]. The first stage is a coarse-grained search for a diverse set of low-energy decoy conformations. The second stage then refines these decoy conformations at the all-atom level of detail. The goal is to find near-native decoy conformations in stage one which are close enough that an all-atom refinement can reach the true native conformation. Shown successful in protein structure prediction, this template has very early origins in work by Scheraga and colleagues [20]–[22].

trajectories are often only the first stage in the search for the native structure. A popular approach for the coarse-grained search is to launch many independent Metropolis Monte Carlo (MMC) [3] or Molecular Dynamics (MD) [23] trajectories. While easily parallelizable, the independent trajectories cannot be guaranteed to explore different regions

978-1-4577-1613-3/11/$26.00 ©2011 IEEE

of the search space. Typically, some form of post-processing is performed to cluster computed conformations based on geometric similarity. The goal is to reveal a few representatives of different interesting regions of the conformational space. Conformations representative of these regions are then refined at all-atom detail. This process is time-consuming and can only be afforded on few conformations. For this reason, it is important that the coarse-grained search reveal geometrically-diverse conformations.

A different search template has recently been proposed by the Shehu lab [24], [25]. Instead of exploring the conformational space through independent trajectories and relying on analysis aposteriori, the search framework in [24] conducts a tree-based exploration of the protein conformational space. Analysis over lower-dimensional embeddings of the explored conformational space and its energy surface is conducted during the exploration in order to guide the tree towards low-energy regions of the energy surface and away from over-populated regions of the conformational space.

This work examines the ability of a new geometric measure that discriminates between protein conformations, known as Effective Connectivity (EC), to improve sampling of a probabilistic search framework guided by low-dimensional projections. This measure is the first eigenvector of the interatomic distance matrix and has been proposed by the Vendruscolo group as a low-dimensional representation of a conformation's geometric structure [26], [27].

Here we employ EC to geometrically project conformations sampled during a search. EC is incorporated into our recently published FeLTr framework which uses a geometric projection layer to ensure structural diversity during a conformational search [24], [25], [28], [29]. Section III shows that, for many target proteins, EC is as effective as our previous approach at biasing sampling towards near-native conformations and is able to significantly improve the average lRMSD sampled for several proteins.

A. Related Work

A common approach for a coarse-grained structure prediction is to run my independent MMC trajectories and cluster the resulting conformations based on geometric similarity [3], [15]–[19]. Sampling diverse conformations is important, as the independent trajectories are often only the first stage in the search for the native structure. Typically, some form of post-processing is performed to cluster computed conformations based on geometrical similarity. The goal is to reveal a few representatives of different interesting regions of the conformational space. Conformations representative of these regions are then refined at all-atom detail. This process is time-consuming and can only be afforded on few conformations. For this reason, it is important that the coarse-grained search reveal geometrically-diverse conformations.

Performing a single clustering operation on the result of many MMC trajectories allows for the use of a wide range of existing clustering methods [30], [31]. However, there is no guarantee that the independent MMC trajectories will sufficiently diverge to provide a representative sample of the energy surface. Brunette and Brock propose an iterative approach which periodically clusters the results of MMC trajectories, using the cluster centroids to launch a new set of trajectories [16]. This approach periodically reapportions computational resources, but still does not explicitly bias the MMC trajectories towards diverse conformations.

Our group recently proposed a probabilistic search framework, FeLTr, which employs a geometric projection layer to ensure structural diversity during conformational search [24], [25], [28], [29]. Conformations sampled during the search are projected onto a low-dimensional geometric space where they can be efficiently grouped based on geometric similarity. This allows FeLTr to dynamically bias further conformational sampling away from regions of the space which have already been heavily explored. The Ultrafast Shape Recognition (USR) [32], [33] projection coordinates employed in FeLTr are shown to effectively bias sampling towards near-native conformations [24]. However, analysis shows that conformations with similar USR coordinates may have a least Root Mean Square Deviation (lRMSD) of more than 2Å.

Finding suitable geometric coordinates to summarize a protein conformation remains an open problem in biophysics [34]. Measures such as lRMSD and radius of gyration mask important differences between conformations. Computing lRMSD is also expensive, and studies show that important minima are missed when grouping conformations by radius of gyration [15].

Q-Score, employs the interatomic distances within a protein molecule to create an effective similarity measure [9], [35]. Q-Score computes the full distance matrix between C_α atoms for each conformation. The difference between two conformations is then the summation of the difference between the two distance matrices. Unlike lRMSD, Q-Score is very accurate at measuring the difference between not only similar conformations, but also more distant ones and does so without the need for structural alignment [9]. Q-Score can also be modified by comparing subsets of the distance matrix to measure only large differences in global tertiary structure or only local differences in secondary structure. Despite many advantages, however, computing the Q-Score between intermediate decoy conformations requires calculating and maintaining the entire $n*(n-1)^2/2$ distance matrix (where n is the number of amino acids in the protein).

The Vendruscolo group recently proposed Effective Connectivity (EC) as a new low-dimensional geometric representation of a conformation's geometric structure [26], [27]. Like Q-Score, EC is based on the interatomic distance matrix. However, EC reduces the complexity of the distance matrix to a threshold-based contact map. EC also reduces the dimensionallity by storing only the first eigenvector of the

contact map, which is shown to be a sufficient representation to reconstruct the full contact map [36].

Work in [26], [27] employs the difference between the EC of the known native structure and the EC of the current sampled conformation (let's refer to it as EC_diff) as another term in the pseudo-energy function minimized during a MMC search. Through extensive analysis (partial data shown in Table I), we have been able to replicate these results on short protein chains. We have found, however, that on chains longer than 50 amino acids, the use of EC_diff in a simple energy function does not offer improvements over a physically-realistic energy function. Later work employs EC_diff as an alternative to the Rosetta [37] potential energy function as post-processing filter over conformations resulting from independent MMC trajectories. Analysis reveals a high correlation between EC_diff and lRMSD to the native structure [38].

The work by the Vendruscolo group demonstrates EC's effectiveness as part of a pseudo energy function comparing decoy conformations to the native structure. The work we present here examines the ability of EC to differentiate between two decoy conformations generated during a search of the conformational space. EC is employed to project conformations sampled during a search onto a geometric projection layer. This projection is incorporated into our recently published FeLTr framework which uses the geometric projection layer to ensure structural diversity during a conformational search. Section III shows that EC is effective not only as EC_diff, but also as a general purpose geometric projection layer.

II. Methods

We first provide a brief overview of FeLTr. A detailed description can be found in previous publications [24], [28].

FeLTr explores the protein conformational space though a tree-based search. Pseudo code is provided in Algo. 1. FeLTr takes a protein sequence α as input and outputs an ensemble Ω_α of low-energy conformations. The root of the search tree is initialized with an extended conformation (Algo. 1, lines 1-2). At each iteration, FeLTr selects a vertex in the tree for expansion through a short MMC trajectory that employs fragment-based assembly [39]. The effectiveness of FeLTr relies on a two-level projection layer which biases selection towards conformations which are both low-energy and geometrically-diverse.

To select a vertex for expansion, first an energy level ℓ is selected (Algo. 1, line 4). This level projects each conformation onto a one-dimensional grid based on potential energy, grouping conformations in increments of 2kcal/mol. The selection is made by weighting each energy level such that $w(\ell) = E_{\text{avg}}(\ell) \cdot E_{\text{avg}}(\ell)$. A level ℓ is then selected with probability $w(\ell)/\sum_{\ell' \in \text{Layer}_E} w(\ell')$.

Once an energy level is selected, FeLTr chooses a geometric cell within ℓ (Algo. 1, line 5). Conformations are projected into grid cells using USR(section II-A). In this work, we explore the use of EC to project conformations onto a Locally Sensitive Hash (LSH) embedding (sections II-B and II-C, respectively). A second weighting function is used to select a cell according to the formula $w(cell) = 1.0/[(1.0 + \texttt{nsel}) \cdot \texttt{nconfs}]$. The variable $nsel$ is the number of times the cell has been selected in the past and $nconfs$ is the total number of conformations which project into the cell. A conformation C corresponding to both the selected energy level and geometric cell is then selected uniformly at random (Algo. 1, line 6). This process biases selection of C towards under explored regions of the conformational space.

The selected conformation C is expanded with a short MMC trajectory that results in a new conformation C_{new} (Algo. 1, line 7). The length of the trajectory is $n - 2$, where n is the number of amino acids in the target protein. Each MMC move performs a single trimer replacement as described in [28]. The energy function used to evaluate each C_{new} is a modified version of the AMW [40] energy function which combines both local and long-range energy terms as described in previous work [24]. The result of this expansion is finally added as a new vertex in the search tree (Algo. 1, line 8) and in the associated output ensemble Ω_α (Algo. 1, line 9).

Algo. 1 A high-level description of the FeLTr framework is given as pseudo code.

Input: α, amino-acid sequence
Output: ensemble Ω_α of conformations

1: $C_{\text{init}} \leftarrow$ extended coarse-grained conf from α
2: AddConf($C_{\text{init}}, \text{Layer}_E, \text{Layer}_{\text{Proj}}$)
3: **while** Time AND $|\Omega_\alpha|$ do not exceed limits **do**
4: $\quad \ell \leftarrow$ SelectEnergyLevel(Layer_E)
5: $\quad$ cell $\leftarrow$ SelectGeomCell($\ell.\text{Layer}_{\text{Proj}}$.cells)
6: $\quad C \leftarrow$ SelectConf(cell.confs)
7: $\quad C_{\text{new}} \leftarrow$ ExpandConf(C)
8: $\quad$ AddConf($C_{\text{new}}, \text{Layer}_E, \text{Layer}_{\text{Proj}}$)
9: $\quad \Omega_\alpha \leftarrow \Omega_\alpha \cup \{C_{\text{new}}\}$

A. Ultrafast Shape Recognition (USR)

Our previous implementation of FeLTr employs coordinates taken from the USR method as the geometric projection layer [32], [33]. USR encodes the shape of a molecule as a vector of geometric descriptors which contain a subset of interatomic distances. USR avoids calculating the entire distance matrix by selecting only a few critical points from which to measure interatomic distances. This work employs three of these points: the center of mass of the conformation (ctd), the farthest C_α atom from the ctd (fct), and the farthest C_α atom from the fct (ftf). For each of these points, the mean Euclidian distance to every C_α atom in the protein conformation is calculated.

Once a conformation is projected into these three USR coordinates, it is stored in a three-dimensional grid, with each dimension representing one of the coordinates. Each dimension is split into 30 cells and the range of the grid is calculated based on the minimum and maximum possible radius of of gyration for the protein. The result is an efficient method for grouping similar decoy conformations generated during the search.

B. Effective Connectivity (EC)

In this work we explore the employment of EC as the geometric projection layer over USR. EC represents the internal connectivity of a conformation as a vector of length n. The contact map M is an $n \times n$ symmetric matrix, where n is the number of amino acids in the protein chain. Each matrix cell, M_{ij} is 1 or 0 depending on whether or not the two amino acids i and j are in contact. Two amino acids are in contact if and only if the Euclidean distance between their C_α atoms is less than a given threshold and they are more than three amino acids apart in the protein sequence ($|i - j| > 3$).The distance between amino acids which are close in sequence defines the secondary structure of the protein. Since we employ EC as a measure of global structure, these secondary structure contacts only add noise. EC is then the principle (or first) eigenvector of M, providing a compact representation of the contact map.

We tested a range of threshold values from 4.5Å to 8.5Å employed by other studies [26], [27], [38] and found that there was no consistently optimal value. Additional work is needed to determine an optimal threshold value based on protein size and fold characteristics. For consistency, the results shown in section III all use a threshold value of 7.5Å.

C. Locally Sensitive Hash (LSH)

In our previous implementation of FeLTr, the problem of grouping similar conformations is solved by mapping the three USR coordinates onto a three-dimensional grid. Conformations are then considered to be geometrically similar if they lie in the same grid cell. The size, n, of an EC array, however, is equal to the number of amino acids in the target protein chain. With proteins over 100 amino acids in length, the curse of dimensionality dictates that an n-dimensional grid will not be effective at categorizing conformations by EC. Since EC is already an approximation of the conformation's structure, we employ an LSH as an efficient approximate similarity search technique which has been shown to be effective on high-dimensional data [41].

The LSH function generates a hash key for each EC array, mapping geometrically similar conformations to the same key. An LSH consists of a set of h randomly generated hyper planes of n dimensions, where n the length of the EC array. An EC array then represents a single point in the n dimensional space. A hash key can be computed by calculating the normal vector to a given point from each hyper plane. If the direction of the normal vector is negative, then hash value is 1; otherwise it is 0. The result is a bit vector of length h, which can be represented as an integer. In this study, we use 15 as the value of h, since 2^{15} is approximately the number of grid cells used in our previous work which employs USR.

III. Results

We compare the results of employing FeLTr using EC, referred to as FeLTR-EC, over those employing USR, referred to as FeLTr-USR. We do so on an extensive set of target proteins with known native structure in the Protein Data Bank(PDB) [42] . In this work, we report a representative subset of 10 targets which range in length from 38 to 93 amino acids and cover α, β, and α/β folds. Section III-A briefly describes our experimental procedure. Section III-B compares EC as a pseudo-energy function (EC_diff) to a physically realistic energy function (AMW). Section III-C then compares FeLTr-EC and FeLTr-USR with respect to the minimum lRMSD from the known native structure. Finally, section III-D presents the distribution of lRMSDs of the conformations sampled during each experiment.

A. Experiments and Measurements

Each experiment using the FeLTr framework is run until $50,000$-$100,000$ conformations are added to the output ensemble. In practice, this takes about two days of CPU time on a 2.66 GHz Opteron processor with 8 GB of memory. USR is simpler to calculate than EC, however, in this context the difference is insignificant compared to the cost of an energy function evaluation. Therefore, holding the number of conformations which must be evaluated constant provides a fair comparison.

The Experiments comparing MMC using EC_diff and AMW as the energy function are each run for 12 hours of CPU time on a 2.66 GHz Opteron processor. EC_diff is significantly faster to compute than AMW, and thus the MMC-EC_diff search is able to sample an order of magnitude more conformations than MMC-AMW for the same amount of CPU time.

We calculate the lRMSD from the native structure for each conformation sampled in both FeLTr and MMC using the heavy backbone atoms N, C_α, C and O.

B. EC_diff

We initially performed a series of experiments to judge the effectiveness of EC as an energy function. An MMC search of the protein conformational space was conducted, using EC as part of the energy function. To analyze this, we carefully compensated for the difference in range between the physically-realistic AMW energy function and the EEC pseudo-energy function. This was done through weights to scale the value of AMW versus that of EEC. Based

Table I: The lowest lRMSD from the native structure obtained by each algorithm is shown. Columns 5 and 6 show the results for MMC using both EC_diff and AMW as the energy function. Columns 7 and 8 show FeLTr employing both USR and EC as the geometric projection layer.

				avg(min) lowest lRMSD in Å			
	PDB ID	length	fold	MMC-AMW	MMC-EC_diff	FeLTr-USR	FeLTr-EC
1	1ail	70	α/β	2.2	4.1	4.0	4.6
2	1c8cA	64	α/β	6.8	6.4	6.1	7.7
3	1cc5	76	α	6.6	6.3	5.7	7.0
4	1fwp	69	α/β	6.7	7.0	7.4	7.5
5	1hz6A	67	α/β	5.7	5.6	6.4	6.7
6	1isuA	62	α/β	6.0	6.1	6.5	7.2
7	1wapA	68	β	7.0	7.0	8.7	8.4
8	2ezk	93	α	7.5	5.6	7.4	7.3
9	2i2v4	38	β	4.3	4.1	4.2	3.9
10	4icb	76	α	4.8	5.5	5.3	5.2

on the results, we did not observe dramatic differences between an MMC exploration with AMW as the energy function or an MMC exploration with EEC as the energy function. These results, juxtaposed in Table I, suggest that EC can work as well as AMW, and its next employment as a geometric measure between two conformations merits further investigation.

C. EC versus USR

We compare the ability of FeLTr to sample near-native conformations using USR and a 3d grid (FeLTr-USR) and EC with LSH (FeLTr-EC). Table I shows the minimum lRMSD from the native structure reached by each method on the 10 target proteins. The results show that the minimum lRMSD found by FeLTr-USR and FeLTr-EC is within 0.5Å for 6 of the 10 target proteins, with 4 cases where FeLTr-EC performs significantly worse than FeLTr-USR.

D. Histogram Analysis

Table I shows the minimum lRMSD from the native structure achieved in each setting. This is a useful metric to compare each method, but it can be distorted by outlying data. Therefore, Figure 1 plots the ensemble of conformations produced as a distribution of the lRMSDs from the native structure for each conformation sampled. Figure 1 shows that, while FeLTr-EC and FeLTr-USR result in similar minimum lRMSDs, the average lRMSD of sampled conformations is lower for FeLTr-EC on four target proteins. This suggests that the use of EC and LSH for the geometric projection layer has a significant effect on specific proteins.

IV. Conclusion

This work investigates the effectiveness of EC combined with LSH as a new low-dimensional projection layer in a probabilistic search framework guided by projections. Results show that EC works as well as USR in many cases. Analysis of the ensemble distribution in Figure 1 shows some bias towards lower-lRMSD structures on a few selected proteins. While this effect fails to improve the minimum lRMSD sampled, it is an interesting observation that merits further investigation. Since the effectiveness of EC does not appear to correlate to protein length or fold topology, further study is needed to determine why EC is more effective on certain proteins. An interesting observation elucidated by this work is that the employment of EC is more powerful as a projection coordinate than as part of an energy function. Physically-realistic energy functions are more effective than EC on medium protein chains.

Acknowledgment

This work is supported in part by NSF CCF No. 1016995.

References

[1] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, no. 4096, pp. 223–230, 1973.

[2] K. A. Dill, B. Ozkan, M. S. Shell, and T. R. Weikl, "The protein folding problem," *Annu. Rev Biophys.*, vol. 37, pp. 289–316, 2008.

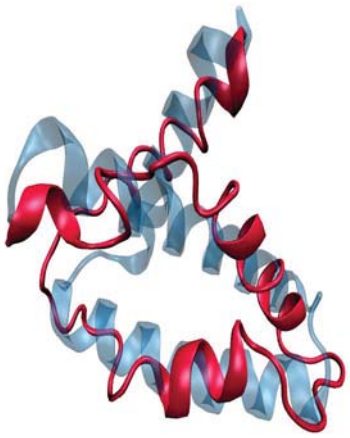
[3] P. Bradley, K. M. S. Misura, and D. Baker, "Toward high-resolution de novo structure prediction for small proteins," *Science*, vol. 309, no. 5742, pp. 1868–1871, 2005.

[4] S. Yin, F. Ding, and N. V. Dokholyan, "Eris: an automated estimator of protein stability," *Nat Methods*, vol. 4, no. 6, pp. 466–467, 2007.

[5] T. Kortemme and D. Baker, "Computational design of protein-protein interactions," *Curr. Opinion Struct. Biol.*, vol. 8, no. 1, pp. 91–97, 2004.

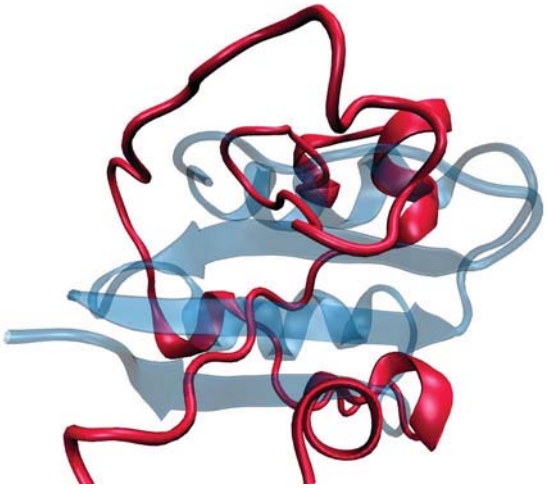
[6] K. A. Dill and H. S. Chan, "From levinthal to pathways to funnels," *Nat. Struct. Biol.*, vol. 4, no. 1, pp. 10–19, 1997.

[7] Y. Zhang, "Progress and challenges in protein structure prediction," *Curr. Opinion Struct. Biol.*, vol. 18, no. 3, pp. 342–348, 2008.

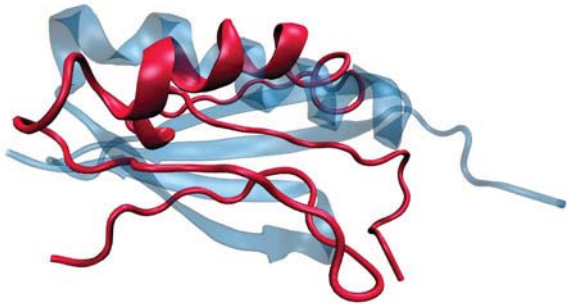


(a) 1ail (4.6 Å)

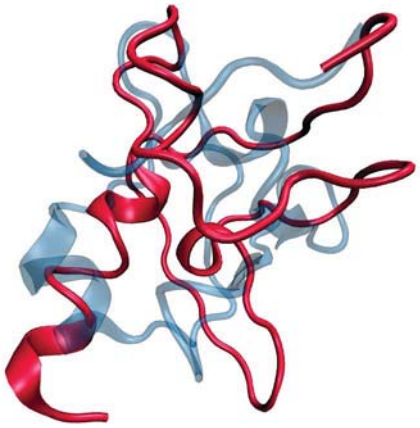
(b) 1cc5 (7.0 Å)



(c) 1fwp (7.5 Å)



(d) 1hz6A (6.7 Å)



(e) 1isuA (7.2 Å)

(f) 4icb (5.2 Å)

Figure 2: (a-f) show the lowest-lRMSD structure discovered by FeLTr-EC in red superimposed over the known native structure in transparent blue.

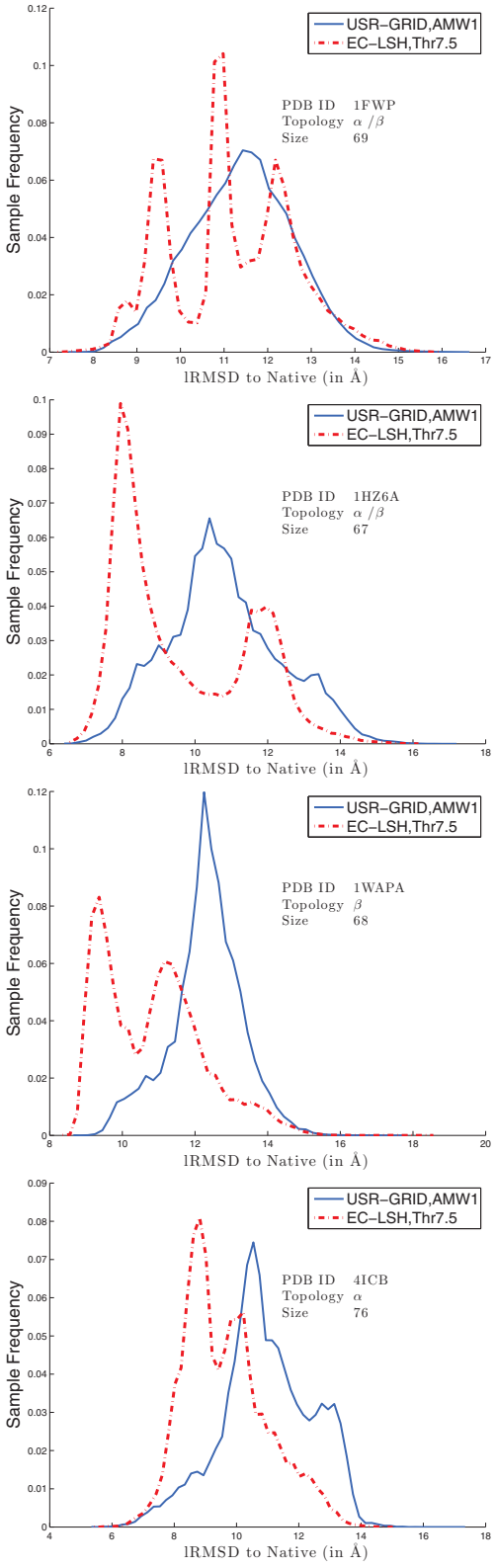


Figure 1: (a)-(d) show the distribution of conformations in the output ensemble for a given lRMSD from the native structure. FeLTr-USR is shown as a solid blue line and FeLTr-EC is shown as a dashed red line.

[8] J. Lee, S. Wu, and Y. Zhang, "Ab initio protein structure prediction," in *Ab Initio Protein Structure Prediction*, D. Rigden, Ed. Springer Science + Business Media B.V., 2009, ch. 1.

[9] M. Ben-David, O. Noivirt-Brik, and A. Paz, "Assessment of CASP8 structure predictions for template free targets," *Proteins: Structure*, Jan 2009.

[10] A. Shehu, "Conformational search for the protein native state," in *Protein Structure Prediction: Method and Algorithms*, H. Rangwala and G. Karypis, Eds. Fairfax, VA: Wiley Book Series on Bioinformatics, 2010, ch. 21.

[11] R. Abagyan and M. Totrov, "Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins," *J. Mol. Biol.*, vol. 235, no. 3, pp. 983–1002, 1994.

[12] J. S. Evans, A. M. Mathiowetz, S. I. Chan, and W. A. Goddard, "De novo prediction of polypeptide conformations using dihedral probability grid Monte Carlo methodology," *Protein Sci.*, vol. 4, no. 6, pp. 1203–1216, 1995.

[13] W. F. van Gunsteren and et al., "Biomolecular modeling: Goals, problems, perspectives," *Angew. Chem. Int. Ed. Engl.*, vol. 45, no. 25, pp. 4064–4092, 2006.

[14] Y. Okamoto, "Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations," *J. Mol. Graph. Model.*, vol. 22, no. 5, pp. 59–64, 2004.

[15] A. Shehu, L. E. Kavraki, and C. Clementi, "Multiscale characterization of protein conformational ensembles," *Proteins: Struct. Funct. Bioinf.*, vol. 76, no. 4, pp. 837–851, 2009.

[16] T. J. Brunette and O. Brock, "Guiding conformation space search with an all-atom energy potential," *Proteins: Struct. Funct. Bioinf.*, vol. 73, no. 4, pp. 958–972, 2009.

[17] R. Bonneau and D. Baker, "De novo prediction of three-dimensional structures for major protein families," *J. Mol. Biol.*, vol. 322, no. 1, pp. 65–78, 2002.

[18] J. DeBartolo, A. Colubri, A. K. Jha, J. E. Fitzgerald, K. F. Freed, and T. R. Sosnick, "Mimicking the folding pathway to improve homology-free protein structure prediction," *Proc. Natl. Acad. Sci. USA*, vol. 106, no. 10, pp. 3734–3739, 2009.

[19] A. Shehu, L. E. Kavraki, and C. Clementi, "Unfolding the fold of cyclic cysteine-rich peptides," *Protein Sci.*, vol. 17, no. 3, pp. 482–493, 2008.

[20] M. H. Lambert and H. A. Scheraga, "Pattern recognition in the prediction of protein structure. ii. chain conformation from a probability-directed search procedure," *J. Comput. Chem.*, vol. 10, no. 6, pp. 798–816, 1989.

[21] ——, "Pattern recognition in the prediction of protein structure. iii. an importance-sampling minimization procedure," *J. Comput. Chem.*, vol. 10, no. 6, pp. 817–831, 1989.

[22] M. Dudek and H. J. Scheraga, "Protein structure prediction using a combination of sequence homology and global energy minimization. i. global energy minimization of surface loops," *J. Comput. Chem.*, vol. 11, no. 1, pp. 121–151, 1990.

[23] Y. Duan and P. Kollman, "Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution," *Science*, pp. 740–744, 1998.

[24] A. Shehu and B. Olson, "Guiding the search for native-like protein conformations with an ab-initio tree-based exploration," *Int. J. Robot. Res.*, vol. 29, no. 8, pp. 1106–11 227, 2010.

[25] A. Shehu, "An ab-initio tree-based exploration to enhance sampling of low-energy protein conformations," in *Robot: Sci. and Sys.*, Seattle, WA, USA, 2009, pp. 241–248.

[26] K. Wolff, M. Vendruscolo, and M. Porto, "A stochastic method for the reconstruction of protein structures from one-dimensional structural profiles," *Gene*, vol. 422, no. 1-2, pp. 47 – 51, 2008.

[27] ——, "Stochastic reconstruction of protein structures from effective connectivity profiles," *PMC Biophys*, vol. 1, no. 1, p. 5, Jan 2008.

[28] B. Olson, K. Molloy, and A. Shehu, "In search of the protein native state with a probabilistic sampling approach," *J. Bioinf. and Comp. Biol.*, vol. 9, no. 3, pp. 383–398, 2011.

[29] ——, "Enhancing sampling of the conformational space near the protein native state," in *BIONETICS: Intl. Conf. on Bio-inspired Models of Network, Information, and Computing Systems*, Boston, MA, December 2010.

[30] J. Hartigan, *Clustering Algorithms*. New York: John Wiley and Sons, 1975.

[31] K. Molloy, "Variable-length fragment assembly within a probabilistic protein structure prediction framework," Fairfax, Virginia, 2011.

[32] P. J. Ballester and G. Richards, "Ultrafast shape recognition to search compound databases for similar molecular shapes," *J. Comput. Chem.*, vol. 28, no. 10, pp. 1711–1723, 2007.

[33] P. J. Ballester, I. Westwood, N. Laurieri, E. Sim, and W. G. Richards, "Prospective virtual screening with ultrafast shape recognition: the identification of novel inhibitors of arylamine n-acetyltransferases," *Journal of The Royal Society Interface*, vol. 7, no. 43, pp. 335–342, 2010. [Online]. Available: http://rsif.royalsocietypublishing.org/content/7/43/335.abstract

[34] C. Clementi, "Coarse-grained models of protein folding: Toy-models or predictive tools?" *Curr. Opinion Struct. Biol.*, vol. 18, pp. 10–15, 2008.

[35] M. P. Eastwood, C. Hardin, Z. Luthey-Schulten, and P. G. Wolynes, "Evaluating protein structure-prediction schemes using energy landscape theory," *IBM Journal of Research and Development*, vol. 45, no. 3.4, pp. 475 –497, may 2001.

[36] M. Porto, U. Bastolla, H. E. Roman, and M. Vendruscolo, "Reconstruction of protein structures from a vectorial representation," *Phys. Rev. Lett.*, vol. 92, p. 218101, May 2004. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevLett.92.218101

[37] K. W. Kaufmann, G. H. Lemmon, S. L. DeLuca, J. H. Sheehan, and J. Meiler, "Practically useful: What the rosetta protein modeling suite can do for you," *Biochemistry*, vol. 49, no. 14, pp. 2987–2998, 2010, pMID: 20235548. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/bi902153g

[38] K. Wolff, M. Vendruscolo, and M. Porto, "Efficient identification of near-native conformations in ab initio protein structure prediction using structural profiles," *Proteins: Structure*, Jan 2010.

[39] N. Haspel, C. Tsai, H. Wolfson, and R. Nussinov, "Reducing the computational complexity of protein folding via fragment folding and assembly," *Protein Sci.*, vol. 12, no. 6, pp. 1177–1187, 2003.

[40] G. A. Papoian, J. Ulander, M. P. Eastwood, Z. Luthey-Schulten, and P. G. Wolynes, "Water in protein structure prediction," *Proc. Natl. Acad. Sci. USA*, vol. 101, no. 10, pp. 3352–3357, 2004.

[41] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *Proceedings of the 25th International Conference on Very Large Data Bases*, ser. VLDB '99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 518–529. [Online]. Available: http://dl.acm.org/citation.cfm?id=645925.671516

[42] H. M. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank," *Nat. Struct. Biol.*, vol. 10, no. 12, pp. 980–980, 2003.