Journal of Bioinformatics and Computational Biology
Vol. 10, No. 3 (2012) 1242005 (17 pages)

DOI: 10.1142/S021972001242005X

Imperial College Press
www.icpress.co.uk

GUIDING PROBABILISTIC SEARCH OF THE PROTEIN CONFORMATIONAL SPACE WITH STRUCTURAL PROFILES

BRIAN OLSON*, KEVIN MOLLOY*, S. FARID HENDI*
and AMARDA SHEHU*,†,‡,§

*Department of Computer Science, George Mason University
4400 University Drive Fairfax, VA, 22030, USA

†Department of Bioengineering, George Mason University
4400 University Drive Fairfax, VA, 22030, USA

‡Department of Bioinformatics and Computational Biology
George Mason University, 4400 University Drive Fairfax
VA, 22030, USA
§ashehu@gmu.edu

Received 11 February 2012
Revised 6 April 2012
Accepted 7 April 2012
Published 4 June 2012

The roughness of the protein energy surface poses a significant challenge to search algorithms that seek to obtain a structural characterization of the native state. Recent research seeks to bias search toward near-native conformations through one-dimensional structural profiles of the protein native state. Here we investigate the effectiveness of such profiles in a structure prediction setting for proteins of various sizes and folds. We pursue two directions. We first investigate the contribution of structural profiles in comparison to or in conjunction with physics-based energy functions in providing an effective energy bias. We conduct this investigation in the context of Metropolis Monte Carlo with fragment-based assembly. Second, we explore the effectiveness of structural profiles in providing projection coordinates through which to organize the conformational space. We do so in the context of a robotics-inspired search framework proposed in our lab that employs projections of the conformational space to guide search. Our findings indicate that structural profiles are most effective in obtaining physically realistic near-native conformations when employed in conjunction with physics-based energy functions. Our findings also show that these profiles are very effective when employed instead as projection coordinates to guide probabilistic search toward undersampled regions of the conformational space.

Keywords: Protein conformational space; near-native conformations; structural profile.

1. Introduction

After nearly four decades, structural characterization of the native state for a novel protein sequence remains a central challenge in computational structural biology.[1]

§Corresponding author.

Experimental techniques are not able to keep pace with the exponential growth in the number of uncharacterized sequences deposited in protein databases. Computational methods present a complementary approach to provide a structural characterization of the protein native state and advance our understanding of the structure−function relationship in proteins.[1]

According to the thermodynamic hypothesis, the protein native state has lowest free energy and consists of low-energy conformations in the protein energy surface.[2] The surface emerges when associating with each conformation a potential energy value resulting from summation of interatomic interactions. The term conformation refers to a spatial arrangement of the atoms in a protein molecule and can be represented, for instance, through the dihedral angles over rotatable bonds connecting atoms in the protein chain. Each amino acid contributes two or more dihedral angles. The result is a space of at least $2n$ dimensions for a chain of n amino acids.

The dimensionality of the conformational space is one of the principal reasons that computing conformations of the native state is challenging. Many strategies are employed to simplify the conformational space. A popular strategy is to lower the dimensionality through coarse-grained representations. One such representation may model only the protein backbone with two dihedral angles per amino acid. Instead of independently modifying angles to sample conformations, techniques such as fragment-based assembly group angles in fragments of consecutive amino acids. Values for the angles in a fragment are obtained by copying those of configurations of that fragment in native structures deposited in experimental databases. The process is essentially assembly of protein-like conformations with fragment configurations and is currently among the most successful in *ab initio* structure prediction.[3,4]

Ruggedness of the protein energy surface is another reason that the search for native conformations is difficult. While modeling energy allows providing an energy bias to a search algorithm and leads to low-energy conformations, the energy surface is rich in local minima. Significant research is devoted to the design of coarse-grained energy functions that simplify the effective energy surface by essentially reducing its ruggedness to help guide search algorithms to regions near the native state.[5] Applications of probabilistic search algorithms that employ state-of-the-art not highly rugged potentials show that, in many cases, the search is led to low-energy regions that do not contain near-native conformations.[1]

Recent research advocates the use of structural profiles of the native state in investigating the role of additional information in guiding search to near-native conformations.[6,7] A structural profile reduces information about the three-dimensional structure of a protein in a one-dimensional vector representation.[8] Profiles that encapsulate information of a contact matrix have been shown to correlate well with hydrophobicity,[9] provide effective and efficient structural alignments,[10] distinguish decoy from near-native conformations,[11] and even assist with guiding Metropolis Monte Carlo (MMC) simulations toward near-native conformations on short protein chains (up to 50 amino acids).[6,7]

In this paper we investigate the applicability and effectiveness of structural profiles in a structure prediction setting for an extensive list of proteins of various sizes and folds (up to 123 amino acids). We pursue two directions. We first investigate the contribution of structural profiles in comparison to or in conjunction with physics−based energy functions in providing an effective energy bias that leads stochastic search to the native state. We do so in the context of an MMC search that employs fragment-based assembly. Second, inspired by the effectiveness of structural profiles in aligning and comparing protein structures, we pursue a novel non-energy−based employment of structural profiles in stochastic search. Specifically, we explore the effectiveness of structural profiles in organizing conformational space. We do so in the context of a robotics-inspired search framework recently proposed in our lab that employs projections of the conformational space to guide search.

We conduct a detailed analysis of computed conformations to determine their proximity to the native state through measures such as least Root-Mean-Squared Deviation (lRMSD) and percentage of native contacts (Q value). Our findings indicate that structural profiles are most effective in obtaining physically realistic near-native conformations when employed in conjunction with physics-based energy functions. Our findings also show that these profiles are equally effective when employed not as part of the energy bias but as projection coordinates instead to guide probabilistic search toward etc undersampled regions of the conformational space.

1.1. *Related work*

Currently, the most successful algorithms for *ab initio* structure prediction rely on sampling a large number of low-energy conformations to obtain a broad view of local minima. The predominant framework involves launching many MMC trajectories with fragment-based assembly and organizing conformations in local minima through clustering.[3,4,12] Further exploration of select minima with fine-grained (computationally intensive) potentials is often pursued in a second stage to identify the global minimum and so reproduce the native structure with high fidelity.[13] Work in Ref. 4, attempts to address the possible issue of independent MMC trajectories providing redundant information through an iterative approach that periodically clusters conformations and launches new trajectories from the cluster centroids. This approach reapportions computational resources based on the clustering results but does not explicitly bias the exploration toward diverse conformations.

The probabilistic robotics-inspired search framework, FeLTr, integrates the conformational analysis in the search itself. The MMC trajectories are integrated in a tree search structure that maintains sampled conformations. Projection layers are employed to ensure energetic feasibility and structural diversity of sampled conformations.[14−16] Sampled conformations are projected onto a low-dimensional geometric space, where conformations can be grouped based on geometric similarity. This allows FeLTr to dynamically bias further sampling away from oversampled

regions. Our previous work employs the ultrafast shape recognition USR-based projection coordinates proposed in Ref. 17, which capture molecular shape. These coordinates are shown to effectively bias sampling in FeLTr toward near-native conformations.[15]

In contrast to the above approaches, the Basin Hopping (BH) framework explicitly samples local minima through repeated applications of a structural perturbation followed by an energy minimization. While most BH algorithms have been demonstrated with limited applicability in the context of protein structure prediction, they have been shown more effective at sampling local minima than simulated annealing with molecular dynamics.[18] Moreover, a recent effective realization of this template has been shown as effective in reproducing native structures of an extensive list of small- to medium-size proteins as leading fragment-based assembly methods.[19]

These frameworks rely at various degrees on energy to bias their exploration and converge to the native state. However, even with reasonably accurate energy functions, the energy bias often leads to lowest-energy regions that are not near the native state. In an effort to remedy this and assist search algorithms, structural profiles of the native state have been proposed for usage complementary to the energy bias. A one-dimensional profile is proposed in Ref. 8 based on the principal eigenvector of the contact matrix. This profile correlates well with sequence hydrophobicity,[9] which allows its prediction from sequence.[20,21]

Recent work has combined structural profiles with simple coarse-grained energy functions to form a pseudo-energy function by which to bias MMC simulations for structure prediction.[6,7] This work has shown limited success and applicability to small proteins (at most 50 amino acids in length).[7] It remains unclear what structural profiles offer over sophisticated coarse-grained energy functions, and whether they are more effective in comparison to or conjunction with such functions. We pursue this line of investigation here to further elucidate the contribution of structural profiles in guiding stochastic search toward near-native conformations.

Since structural profiles have also been shown effective in protein structure alignment,[10] it is relevant to explore their applicability beyond the energy bias setting. The FeLTr framework that employs projections of the conformational space provides a reasonable setting for this purpose. We pursue a novel direction and employ structural profiles to associate projection coordinates to conformations sampled by FeLTr and so construct a lower-dimensional projection layer, where it is easier to keep statistics for the purpose of guiding search to diverse conformations.

It is worth noting that our investigation of structural profiles in the context of energy bias here employs exact profiles extracted from the native structure. While in principle these profiles can be predicted from sequence, their investigation in the context of stochastic search is a necessary first step into providing further understanding of their role and effectiveness. Employment of these profiles as projection coordinates does not use any information about the native structure.

2. Methods

Since structural profiles in this paper are employed in two different settings, we first describe their computation. We then relate details on each of the two settings. We first describe how structural profiles are incorporated in a pseudo-energy function and employed in the context of MMC search. We then relate details on their employment instead as projection coordinates in the context of the FeLTr framework.

2.1. *One-dimensional structural profiles*

The structural profile employed here is the principal eigenvector (PE) of the contact matrix that can be associated with a conformation. The contact map C is an $N \times N$ binary symmetric matrix, where C_{ij} is 1 or 0 depending on whether or not amino acids i and j are in contact. Two amino acids are in contact 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 ($|i - j| > 3$). We tested a range of threshold values from 4.5 Å to 8.5 Å also employed in other studies[6,7,11] and found values in the 7.5−8.5 Å range to be equally effective. For consistency, all results in this paper use a threshold of 7.5 Å. The constraint on the sequence distance between amino acids i, j ensures that the structural profile captures non-local interactions.

C is a real symmetric matrix and so has N real eigenvalues. The PE is the eigenvector with the largest corresponding eigenvalue. PE encodes each amino acid's connectivity, essentially associating higher connectivity to amino acids with a larger number of contacts. The strong correlation of this structural profile to hydrophobicity, studied in detail in Ref. 9, follows from the fact that amino acids with high connectivity are often those buried in the hydrophobic interior of a structure.

Another equally expressive definition of a structural profile, referred to as effective connectivity (EC), employs a linear combination of all eigenvectors weighted by their corresponding eigenvalues. Research has shown that the main contribution to EC comes from PE; the correlation between the two is about 95% for single-domain proteins.[22] Multi-domain proteins have been shown to be captured better in terms of their contact map through EC rather than PE.[23,24] Since our focus is on single-domain proteins (multi-domain proteins remain beyond the applicability of *ab initio* structure prediction), we investigate PE in this paper.

2.2. *Employing structural profiles to bias MMC search*

2.2.1. *Defining a pseudo-energy term based on PE*

PE can easily be employed to define a pseudo-energy term that achieves its lowest value (0), when the PE of a computed conformation (let us refer to it as c) reaches the PE of the known native structure (let us refer to it as t). The term, which we refer to as E_{PE}, can essentially be the sum of differences in the vector entries per amino acid i as in: $E_{\mathrm{PE}} = \sum_i \min(|c_i - t_i|, 0.25)$. The cutoff of 0.25, suggested by previous research that tests E_{PE} on small proteins,[7] limits the contribution of each vector

entry. This cutoff has no significant effect on conformations that are close to the native structure, but it helps escape local minima at early stages in the search.

The above definition of PE considers all contacts before the eigendecomposition. Variations have emerged in literature, which do not consider all contacts.[6,7] For instance, a distinction is made between cooperative and non-cooperative contacts. Cooperative contacts are defined as those assisting in secondary structure formation. All other contacts, which essentially encapsulate tertiary structure, are considered non-cooperative. The MMC we employ in this work, described below, readily forms secondary structures through fragment-based assembly and does not need to limit PE to cooperative contacts. The non-cooperative contacts, on the other hand, are investigated here and compared to the utilization of all contacts in PE. We also investigate the role of restricted contacts, which are contacts only between amino acids in secondary structures. Restricted contacts still capture tertiary structure, namely the folding of secondary structures, but do not consider coil-like regions (these regions are somewhat easier to address once the fold has been found).

2.2.2. *Incorporating E_{PE} in a pseudo-energy function*

Employing E_{PE} as the pseudo-energy function may lead an MMC simulation to conformations that contain steric clashes and other unfavorable interatomic interactions not captured in this simplistic pseudo-energy function. For this purpose, it is worth investigating the role of E_{PE} not only in comparison to a physically realistic energy function but also in conjunction. Essentially, a pseudo-energy function can be defined by summing the terms of a physically realistic energy function with E_{PE}. Our analysis employs both a simple energy function whose sole purpose is to penalize steric clashes and a sophisticated physics-based energy function, the Associative Memory Hamiltonian with Water (AMW). The AMW has been developed by the Wolynes lab[25] and shown successful in *ab initio* structure prediction by us and others.[12,14–16,19,26] Our employment of AMW does not include local interactions, since fragment configurations are extracted from realistic structures in the Protein Data Bank (PDB)[27] and further idealized. The modified AMW is essentially a linear combination of the non-local terms $E_{\mathrm{Lennard-Jones}}$, $E_{\mathrm{H-Bond}}$, E_{contact}, E_{water}, and E_{burial}. The last three terms model water-mediated interactions in coarse-grained conformations. Details can be found in Ref. 25.

The range of energy values returned by a physically realistic energy function, such as E_{AMW}, and a pseudo-energy term, such as E_{PE}, can be quite different, and weighting each of them to form a pseudo-energy function essentially modulates their contributions. Here we employ a parameter α to essentially weight the contribution of E_{PE} relative to E_{AMW}. The value of α can be static and not change through the search, or it can change according to a dynamic schedule that reweights the contribution of E_{PE} depending on where in the energy surface the search is. Our analysis in Sec. 3 employs a static value for α, but a dynamic schedule is under investigation in ongoing work.

2.2.3. *Investigating structural profiles in an MMC search setting*

Our analysis of the role of PE and its variants in biasing toward near-native conformations does so in the context of an MMC search. It is worth noting that this setting is different from previous work that investigates structural profiles also in MMC simulations for small proteins of up to 50 amino acids.[6,7] Instead of sampling values for the dihedral angles to obtain consecutive conformations in the MMC trajectory, we employ fragment-based assembly. The fragments are of length 3, and their configurations are extracted from libraries that include configurations extracted from structures of sequence-homologous proteins. The configurations are limited, however, to a subset that contain secondary structures consistent with predictions from the sequence of the protein under consideration. More details about the construction of these libraries and their employment can be found in Ref. 14. Significant efforts in previous work on employing E_{PE} went to seeding conformations with secondary structures predicted from sequence. The fragment configurations we employ readily provide protein-like conformations with reasonable secondary structures.

The analysis in Sec. 3 investigates the role of PE in biasing an MMC simulation with fragment-based assembly when employing only E_{PE} as the pseudo-energy function, when combining it with a simplistic collision-avoidance energy term, and when combining it with a sophisticated physics-based energy function that is shown successful for structure prediction. All variants of PE are explored, whether considering all contacts, only non-cooperative contacts, or restricted contacts.

2.3. *Employing structural profiles to organize conformational space*

The employment of structural profiles to provide an energy bias through a pseudo-energy function relies on knowledge of the PE of the native structure. While this can be predicted from sequence with about 70% accuracy,[20] the employment of structural profiles as projection coordinates does not rely on the native PE. Instead, the PE can be calculated for each sampled conformation and regarded as a succinct representation of the topology of that conformation. Given a set of conformations, their PEs can be employed for clustering, for instance, to group together structurally similar conformations and organize the explored conformational space.

Employment of Projections in the FeLTr Framework: Organization of the conformational space in a lower-dimensional embedding is central to the success of the robotics-inspired FeLTr framework proposed in our lab to enhance sampling of geometrically diverse low-energy conformations.[14,15] FeLTr essentially explores the conformational space by growing a tree. Branches are short MMC trajectories that employ fragment-based assembly. The tree biases its growth by selecting conformations from which to grow branches through a two-level projection layer. Conformations in the tree are projected onto a one-dimensional grid based on their potential energy. To select a vertex for expansion, FeLTr selects first an energy level in the grid through a weighting function that biases toward low-energy levels. The goal is to bias

the tree toward low-energy regions of the energy surface. Once an energy level is selected, FeLTr has access to all sampled conformations associated with the selected energy. These conformations are projected onto a three-dimensional (geometric) grid using three coordinates based on the ultrafast shape recognition (USR) features. A second weighting function is used to select a geometric cell that has not been selected many times and does not contain many conformations in it. The goal is to bias the tree away from oversampled regions of the conformational space. Further details on the FeLTr framework can be found in Refs. 14–16.

USR Coordinates: Our previous work employs projection coordinates based on the USR features proposed in Ref. 17. These features encode molecular shape as a vector of geometric descriptors based on a subset of interatomic distances. Only three descriptors are employed (see Ref. 15 for details), resulting in FeLTr using a three-dimensional grid. Each dimension is split into 30 cells, with the range of the grid calculated based on the minimum and maximum possible radii of gyration for a given protein sequence. The result is an efficient process for grouping together geometrically similar conformations sampled during the search.

2.3.1. *Employing PE to project conformational space in FeLTr*

While the PE associated with each conformation in FeLTr succinctly captures structural information in that conformation, its direct employment as a projection coordinate is infeasible. If one were to do so, the result would be an N-dimensional grid (the length of PE is the number of amino acids N). Such a high-dimensional grid would not be effective at organizing conformations. Since PE is already an approximation of a conformation's topology, we employ the Locally-Sensitive Hash (LSH) technique, which has been shown effective on high-dimensional data.[28]

The LSH function generates a hash key for each PE corresponding to a conformation in the FeLTr tree, mapping geometrically similar conformations to the same key. The PE of a conformation represents a single point in N-dimensional space. In LSH, $h < N - 1$ hyperplanes are randomly generated. A hash key is computed by calculating the normal vector to a given (PE) point from each hyperplane. If the direction of the normal vector is negative, the hash value is 1; otherwise it is 0. The result is a bit vector of length h, which can be represented as an integer, thus giving the hash key.

A large value for h results in too many cells, but a small h may group together dissimilar conformations. The analysis in Sec. 3 shows that $h = 15$ is a good compromise. It results in 2^{15} cells, which is close to the number of USR-based cells used in our previous work on FeLTr. More importantly, conformations in a PE-based cell are shown to be structurally very similar and more similar than conformations projected to the same USR-based grid cell.

3. Results

Systems of Study: We consider an extensive list of 13 proteins of varying sizes and folds with known native structures in the PDB. The list is shown in Table 1.

Table 1. PDB id, number of amino acids, and fold are shown for each protein.

Nr.	1	2	3	4	5	6	7	8	9	10	11	12	13
PDB	2i2v4	1isuA	1c8cA	1hz6A	1wapA	1fwp	1ail	4icb	1cc5	2ezk	1hhp	2hg6	2h5nD
N	38	62	64	67	68	69	70	76	83	93	99	106	123
Fold	β	α/β	α/β	α/β	β	α/β	α	α	α	α	β	α/β	α

Experiments: The effectiveness of PE is investigated in two experimental settings. (I) An MMC search is conducted on each protein sequence. Three different scenarios are analyzed: (a) employing only E_{PE} as the pseudo-energy function, considering its variants of non-cooperative contacts (nPE) and restricted contacts (rPE); (b) employing E_{PE} in conjunction with a collision-avoidance energy term; and (c) employing E_{PE} in conjunction with E_{AMW}. (II) PE is employed to project the conformational space through LSH in FeLTr, and results are compared to those obtained when the USR-based grid is used instead.

Measurements: For each experimental setting, three quantities are shown: the lowest lRMSD to the native structure, the lRMSD to the native structure of the lowest-energy conformation sampled, and the highest Q value sampled. It is worth noting that lRMSD and Q value are two complementary measures to determine the proximity of a conformation to the native structure. In particular, Q is an unforgiving measure. Often, low lRMSDs can be obtained on conformations that do not contain many native contacts. In this respect, showing higher Q values is a stronger result than showing lower lRMSDs. Additionally, while the interpretation of low lRMSDs depends on protein size, higher Q values are more indicative of higher-quality conformations. In particular, while most of cooperative folding occurs during the collapse of a protein chain until Q values around 0.4 are reached, values above this threshold are strong indicators that the native topology has been reproduced.[18]

Implementation Details: MMC search is run for a total of 10^6 energy function evaluations. This provides a fair setting, whether the sequence considered is short or long (energy function evaluations are more expensive on longer chains). FeLTr is run until a total of 100,000 conformations have been added to the tree. This allows direct comparison with previous work. For protein chains longer than 100 amino acids, the running time for MMC or FeLTr can exceed 48 hours of CPU time.

3.1. *E_{PE} guides MMC search*

In this setting, the only energy function guiding MMC is E_{PE}. Three variants of PE are compared to one another, the full PE, which considers all contacts, the one that considers only non-cooperative contacts, and the one that considers only restricted contacts. We refer to the respective energy terms as E_{PE}, E_{nPE}, and E_{rPE}. Table 2 shows the lowest lRMSD (lR), the lRMSD of conformation with lowest energy ($\mathrm{R_{lE}}$), and maximum Q (mQ) in each setting.

Inspection of the mQ values in Table 2 indicates that the MMC simulation in each setting has converged to conformations very similar to the native structure (all mQ

Table 2. lR and R_{lE} (in Å) and mQ (in %) are shown when using full PE, non-cooperative PE, and restricted PE to define the pseudo-energy function.

		E_{PE}			E_{nPE}			E_{rPE}		
Nr.	PDB ID	lR	R_{lE}	mQ	lR	R_{lE}	mQ	lR	R_{lE}	mQ
1	2i2v4	3.2	4.9	91	3.2	5.9	86	3.8	6.1	82
2	1isuA	5.7	10.9	65	5.3	12.9	63	5.3	11.7	60
3	1c8cA	5.8	10.4	79	6.5	19.2	69	5.0	13.5	76
4	1hz6A	4.6	12.4	77	6.5	17.1	60	4.4	5.3	74
5	1wapA	5.9	12.1	67	7.0	16.3	49	5.6	12.2	71
6	1fwp	6.2	9.8	64	6.8	13.9	52	6.2	11.8	65
7	1ail	4.2	8.7	78	4.8	10.0	69	4.0	10.0	76
8	4icb	4.3	11.8	79	4.7	8.7	77	5.2	19.9	76
9	1cc5	5.5	10.0	64	6.1	12.2	57	6.0	11.2	70
10	2ezk	5.0	15.8	88	6.2	17.2	82	5.2	13.4	83
11	1hhp	7.5	11.5	64	8.6	14.3	42	7.7	12.8	53
12	2hg6	8.5	13.6	56	9.6	34.0	48	9.2	25.1	56
13	2h5nD	7.5	21.1	69	7.5	18.7	63	7.4	14.7	69

values are $> 40\%$). Overall, mQ values $> 70\%$ correspond well to lRMSDs ≤ 6.0 Å to the native structure, with few exceptions. Comparison of the lowest lRMSD and maximum Q values suggests that neither PE variant has a clear advantage.

3.2. *E_{PE} and steric clash avoidance guide MMC search*

The experiments above are repeated with the pseudo-energy function $E_{\text{steric}} + \alpha E_{\text{PE}}$. E_{steric} is set to a high value when two atoms are in a collision and 0 otherwise, and $\alpha = 10$, as in Ref. 7 (this results in E_{PE} being in the order of one per residue for non-native conformations). The goal is to obtain collision-free conformations with similar contact topology to the native structure. The results are shown in Table 3.

A comparison of the lowest lRMSD and maximum Q values between Tables 2 and 3 shows that similar results are obtained. These findings suggest that E_{steric} does not prohibit E_{PE} from leading the MMC search toward near-native conformations in terms of lRMSD and contact topology, but further improves the quality of these conformations by removing steric clashes.

3.3. *E_{PE} in conjunction with E_{AMW} guides MMC search*

Optimizing a physics-based energy function can often lead to non-native yet low-energy conformations (determined, for instance, through high lRMSDs from the native structure or low Q values). On the other hand, the results above show that biasing stochastic search toward conformations with structural profiles similar to the native structure leads to near-native conformations. In this experiment, we investigate whether combining E_{PE} with E_{AMW}, a physics-based energy function, will lead to both physically realistic and near-native conformations.

Here E_{steric} is replaced with E_{AMW}. The range of values for E_{AMW} is shifted by the minimum energy value that an MMC trajectory reaches, so that the minimum

Table 3. lR and R_{lE} (in Å) and mQ (in %) are shown when using full PE, nPE, and rPE in conjunction with a steric clash avoidance energy term.

		$E_{steric} + \alpha E_{PE}$			$E_{steric} + \alpha E_{nPE}$			$E_{steric} + \alpha E_{rPE}$		
Nr	PDB ID	lR	R_{lE}	mQ	lR	R_{lE}	mQ	lR	R_{lE}	mQ
1	2i2v4	3.6	7.2	73	3.6	7.7	70	3.8	10.1	72
2	1isuA	6.1	9.9	46	6.1	9.3	51	6.7	11.6	44
3	1c8cA	6.5	14.7	68	6.5	18.7	60	6.2	13.5	67
4	1hz6A	5.3	7.7	57	6.8	17.6	50	5.8	12.3	61
5	1wapA	6.9	11.7	46	7.7	18.7	43	7.3	14.2	47
6	1fwp	6.8	11.3	48	7.6	19.7	43	5.9	8.9	47
7	1ail	4.1	5.0	74	3.9	8.7	73	5.5	9.8	71
8	4icb	4.1	9.2	72	4.9	11.0	72	4.6	9.7	72
9	1cc5	6.0	11.7	54	5.7	11.5	51	6.2	12.2	52
10	2ezk	5.7	12.2	84	6.9	26.0	76	5.3	13.3	84
11	1hhp	9.9	13.4	35	10.4	16.4	28	10.6	19.5	28
12	2hg6	9.3	27.1	43	10.0	23.6	40	9.9	17.2	49
13	2h5nD	8.2	20.6	60	NA	NA	NA	9.3	27.6	56

energy value for E_{AMW} is 0, as for E_{PE}. The value of α is kept as above, as it allows both E_{AMW} and E_{PE} to cover the same range of energy values. Table 4 repeats the analysis employing all three different versions of PE. The results are compared with those obtained when using only E_{AMW} for the energy bias as a point of reference.

The results in Table 4 allow drawing the following conclusions: (i) the incorporation of E_{AMW} does not prohibit E_{PE} and its variants in obtaining similar near-native conformations in terms of lowest lRMSDs and maximum Q values (in comparison to results shown in Table 2 obtained employing only E_{PE}); (ii) comparison of the maximum Q values shows that combining PE and its variants with a physics-based energy function, such as E_{AMW}, leads to conformations that are both physically

Table 4. lR and R_{lE} (in Å) and mQ (in %) are shown when using full PE, nPE, and rPE in conjunction with AMW.

		E_{AMW}			$E_{AMW} + \alpha E_{PE}$			$E_{AMW} + \alpha E_{nPE}$			$E_{AMW} + \alpha E_{rPE}$		
Nr.	PDB ID	lR	R_{lE}	mQ	lR	R_{lE}	mQ	lR	R_{lE}	mQ	lR	R_{lE}	mQ
1	2i2v4	4.2	9.9	56	3.8	7.7	68	3.8	5.2	65	4.0	8.7	67
2	1isuA	5.3	11.3	49	5.0	10.5	56	5.5	10.0	50	5.2	11.6	46
3	1c8cA	6.5	17.8	53	6.8	11.1	64	7.3	17.3	63	6.2	11.6	72
4	1hz6A	6.4	12.3	47	5.6	12.5	54	6.1	14.3	51	4.3	11.8	73
5	1wapA	7.6	13.5	40	5.7	12.7	60	7.1	15.2	44	6.3	10.3	47
6	1fwp	7.5	10.7	40	6.0	9.0	51	4.8	5.8	62	5.4	10.4	46
7	1ail	3.8	8.3	85	4.8	7.2	80	4.0	6.6	82	4.5	5.8	84
8	4icb	4.7	11.1	69	4.3	12.2	72	3.6	9.4	76	4.0	8.1	76
9	1cc5	6.5	10.8	52	6.5	10.9	55	6.1	14.8	56	5.0	10.5	54
10	2ezk	4.7	16.8	87	5.9	14.4	89	6.7	17.1	79	5.2	16.2	88
11	1hhp	9.9	14.2	43	9.1	15.1	36	9.4	13.8	37	9.3	15.4	36
12	2hg6	9.7	21.7	43	8.8	17.1	46	9.4	29.3	43	8.6	15.3	46
13	2h5nD	7.7	15.4	63	6.0	14.5	70	6.4	21.1	65	7.0	11.7	68

realistic and closer to the native structure (in percentage of native contacts) than conformations obtained only with E_{AMW} in the MMC search; (iii) comparison of the lowest lRMSDs to the native structure shows that combining PE and its variants with E_{AMW} also leads to comparable or lower lRMSDs than when using only E_{AMW} for the majority of proteins; (iv) in particular, the lowest lRMSDs are most improved when the restricted PE is employed in combination with E_{AMW} over the other variants. Taken together, these results strongly suggest that combining a contact-based structural profile with a sophisticated physics-based energy function improves both conformation quality and proximity to the native state.

3.4. *PE-based projection of conformational space in FeLTR*

We compare the ability of FeLTr to sample near-native conformations when using USR (F_{USR}) over PE with LSH (F_{PE}). Table 5 reports lR, $\mathrm{R_{le}}$, and mQ in each setting. It is worth noting that the number of conformations sampled in FeLTr is an order of magnitude smaller than that sampled in the above MMC runs. Previous research on FeLTr has shown that in a fair setting, FeLTr obtains more native-like conformations that MMC.[14–16] Our goal, here, however, is simply to compare the USR-based to the PE-based projections. Only the full PE vector is employed rather than its variants, since PE is now part of the projection and not energy. Two main conclusions can be drawn from the results shown in Table 5. First, employing PE with LSH to project the conformational space is just as effective as USR in allowing FeLTr to obtain low lRMSDs to the native structure. Comparable and in some cases even lower lRMSDs are obtained through F_{PE} over F_{USR}. Second, even though PE is used here to project the conformational space rather than provide an energy bias, comparable or higher maximum Q values are reached over F_{USR}. This is an interesting result, as it shows that the quality of near-native conformations is improved even when using PE to project conformations.

A detailed analysis of the PE-based projection actually shows that this projection is more effective at decomposing the conformational space. The grouping of similar conformations is tighter than when employing USR-based projection coordinates. Figure 1 summarizes the results as follows. The conformations in the FeLTr tree that

Table 5. PDB id, number of amino acids, and fold are shown for each protein.

Nr.	1	2	3	4	5	6	7	8	9	10	11	12	13
PDB	2i2v4	1isuA	1c8cA	1hz6A	1wapA	1fwp	1ail	4icb	1cc5	2ezk	1hhp	2hg6	2h5nD
$\mathrm{F_{USR}}$													
lR	5.0	7.2	7.5	6.0	7.7	6.7	4.2	5.0	7.0	4.9	10.7	10.8	8.8
$\mathrm{R_{lE}}$	8.5	10.7	9.7	10.5	10.7	14.0	6.7	10.1	10.9	16.3	15.2	15.9	15.4
mQ	37	23	36	37	22	29	65	48	39	65	13	25	43
$\mathrm{F_{PE}}$													
lR	4.2	6.8	6.5	6.5	6.9	7.3	3.8	5.0	6.1	6.4	9.6	11.3	9.6
$\mathrm{R_{lE}}$	8.0	9.6	11.7	12.9	9.2	10.9	9.8	9.2	9.4	12.9	14.6	16.1	13.5
mQ	41	31	48	33	26	27	70	56	39	73	13	23	41

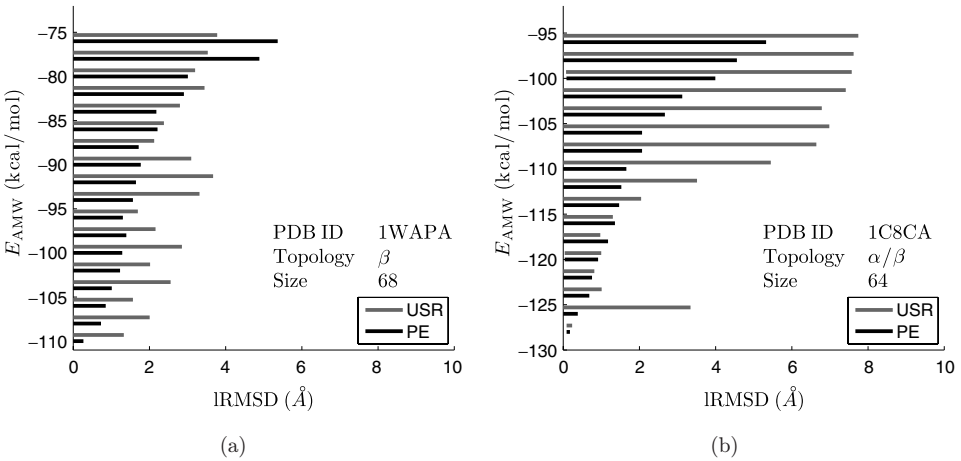

Fig. 1. Segments in black for statistics computed over the PE-based projection are superimposed over those in gray for the USR-based projection.

correspond to energy levels below the average energy sampled are analyzed in terms of their diversity. By using a USR-based or PE-based projection, the lRMSD between any two conformations in a cell is computed, and the minimum and median pairwise lRMSDs are recorded for each cell. The distribution of these values over all cells belonging to an energy level is summarized with the minimum and median statistics. Figure 1 plots the minimum and median statistics for each energy level below the average energy. A segment is drawn connecting the two statistics per energy level to provide a visual representation of conformational diversity. Figure 1 superimposes statistics obtained from the analysis of the PE-based projection over conformations obtained with F_{PE} in black over the statistics obtained from the analysis of the USR-based projection over conformations obtained with F_{USR}. Results are shown on two selected proteins.

Figure 1 shows that the difference between the statistics grows narrower with lower energy values. While this is expected, as the protein energy surface is funnel-like, the PE-based projections capture this feature better. A comparison of the statistics between F_{PE} and F_{USR} shows that the range is narrower for F_{PE}, suggesting that PE with LSH is more sensitive than USR and groups together conformations that are overall structurally more similar. This observation is in agreement with the fact that PE captures more detailed information about the topology of a conformation than the coarse information that the USR features capture about molecular shape.

4. Discussion

This work has conducted a comprehensive investigation of the effectiveness of contact-based structural profiles in assisting stochastic search of the protein conformational

space in sampling near-native conformations. Two settings have been explored, one in which the structural profile provides an energy bias to the search through a pseudo-energy function, and another where the profile is employed instead to organize the protein conformational space. A detailed analysis in the context of an MMC search shows that structural profiles are most effective in producing both near-native and physically realistic conformations when employed in conjunction with physics-based energy functions. The combination allows obtaining high-quality conformations closer to the native structure than when employing only physics-based energy functions.

The analysis shows that structural profiles are just as effective in leading to near-native conformations when employed to project the conformational space. These profiles allow better grouping together structurally similar conformations than coarse descriptors of molecular shape, mainly because contact-based structural profiles capture more details about the topology of a conformation. In light of their effectiveness in organizing conformational space, it is worth exploring different values of h in the LSH mapping to directly control cell width.

The detailed investigation of the role of contact-based structural profiles in assisting stochastic in this paper is a first step into providing further understanding. Future work will investigate profiles predicted from sequence in the context of energy bias. The current literature on predicting the native PE from sequence data requires some modifications to how contact matrices are interpreted. For instance, the contact matrix predicted from sequence is a real-valued symmetric matrix, where an entry records the probability of a contact rather than the presence of the contact or not. In this case, interpretation of the matrix is needed so that a binary symmetric one can be computed for the eigendecomposition that follows. We are currently pursuing this line of investigation.

Additionally, since the combination of E_{PE} with E_{AMW} was shown in this paper to be more powerful than each of the terms alone, as demonstrated in an MMC search setting, the pseudo-energy function that combines both terms can be employed instead of E_{AMW} in the FeLTr framework. In this way, the structural profile would provide both an energy bias to FeLTr and organize conformations in the tree through the LSH-based projection described in the paper. Future work will also pursue a dynamic schedule to weight the contribution of these profiles relative to a physics-based energy function as the search proceeds.

Acknowledgments

This work is supported in part by NSF CCF No. 1016995 and NSF IIS CAREER Award No. 1144106.

References

1. Prentiss MC, Hardin C, Eastwood MP, Zong C, Wolynes PG, Protein structure prediction: The next generation, *J Chem Theory Comput* **2**(3):705−716, 2006.

2. Dill KA, Chan HS, From levinthal to pathways to funnels, *Nat Struct Biol* **4**(1):10−19, 1997.
3. Bradley P, Misura KMS, Baker D, Toward high-resolution *de novo* structure prediction for small proteins, *Science* **309**(5742):1868−1871, 2005.
4. Brunette TJ, Brock O, Guiding conformation space search with an all-atom energy potential, *Proteins: Struct Funct Bioinf* **73**(4):958−972, 2009.
5. Clementi C, Coarse-grained models of protein folding: Toy-models or predictive tools? *Curr Opinion Struct Biol* **18**:10−15, 2008.
6. Wolff K, Vendruscolo M, Porto M, A stochastic method for the reconstruction of protein structures from one-dimensional structural profiles, *Gene* **422**(1−2):47−51, 2008.
7. Wolff K, Vendruscolo K, Porto M, Stochastic reconstruction of protein structures from effective connectivity profiles, *PMC Biophys* **26**(1):5, 2008.
8. Porto M, Bastolla U, Roman HE, Vendruscolo M, Reconstruction of protein structures from a vectorial representation, *Phys Rev Lett* **92**(21):218101, 2004.
9. Bastolla U, Porto M, Roman HE, Vendruscolo M, Principal eigenvector of contact matrices and hydrophobicity profiles in proteins, *Proteins: Struct Funct Bioinf* **58**(1):22−30, 2005.
10. Teichert F, Bastolla U, Porto M, SABERTOOTH: Protein structural alignment based on a vectorial structure representation, *BMC Bioinf* **8**(425):1−17, 2005.
11. Wolff K, Vendruscolo M, Porto M, Efficient identification of near-native conformations in *ab initio* protein structure prediction using structural profiles, *Proteins: Struct Funct Bioinf* **78**(2):249−258, 2010.
12. Shehu A, Kavraki LE, Clementi C, Multiscale characterization of protein conformational ensembles, *Proteins: Struct Funct Bioinf* **76**(4):837−851, 2009.
13. Shehu A, Conformational search for the protein native state, in Rangwala H, Karypis G, (eds.), *Protein Structure Prediction: Method and Algorithms*, Chap. 21, Wiley Book Series on Bioinformatics, Fairfax, VA, 2010.
14. Olson B, Molloy K, Shehu A, In search of the protein native state with a probabilistic sampling approach, *J Bioinf Comp Biol* **9**(3):383−398, 2011.
15. Shehu A, Olson B, Guiding the search for native-like protein conformations with an *ab initio* tree-based exploration, *Int J Robot Res* **29**(8):1106−11227, 2010.
16. Shehu A, An *ab initio* tree-based exploration to enhance sampling of low-energy protein conformations, *Robot: Sci Sys*, Seattle, WA, USA, pp. 241−248, 2009.
17. Ballester PJ, Richards G, Ultrafast shape recognition to search compound databases for similar molecular shapes, *J Comput Chem* **28**(10):1711−1723, 2007.
18. Prentiss MC, Wales DJ, Wolynes PG, Protein structure prediction using basin-hopping, *J Chem Phys* **128**(22):225106, 2008.
19. Olson B, Shehu A, Populating local minima in the protein conformational space, *IEEE Int Conf Bioinformatics and Biomedicine (IEEE BIBM)*, pp. 114−117, 2011.
20. Vullo A, Walsh I, Pollastri G, A two-stage approach for improved prediction of residue contact maps, *BMC Bioinf* **7**:180, 2006.
21. Kinjo R, Nishikawa K, CRNPRED: Highly accurate prediction of one-dimensional protein structures by large-scale critical random networks, *BMC Bioinf* **7**:401, 2006.
22. Bastolla U, Ortiz AR, Porto M, Teichert F, Effective connectivity profile: A structural representation that evidences the relationship between protein structures and sequences, *Proteins: Struct Funct Bioinf* **73**(4):872−888, 2008.
23. Teichert F, Porto M, Vectorial representation of single- and multi-domain protein folds, *Eur Phys J B* **54**:131−136, 2006.

24. Bastolla U, Porto M, Roman HE, Vendruscolo M, A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the protein data bank, *BMC Evol Biol* **6**:43, 2006.
25. Papoian GA, Ulander J, Eastwood MP, Luthey-Schulten Z, Wolynes PG, Water in protein structure prediction, *Proc Natl Acad Sci USA* **101**(10):3352−3357, 2004.
26. Hegler JA, Laetzer J, Shehu A, Clementi C, Wolynes PG, Restriction vs. guidance: Fragment assembly and associative memory hamiltonians for protein structure prediction, *Proc Natl Acad Sci USA* **106**(36):15302−15307, 2009.
27. Berman HM, Henrick K, Nakamura H, Announcing the worldwide Protein Data Bank, *Nat Struct Biol* **10**(12):980−980, 2003.
28. Gionis A, Indyk P, Motwani R, Similarity search in high dimensions via hashing, *Intl Conf on Very Large Databases VLDB '99*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 518−529, 1999.

Brian Olson received his B.S.E. in Computer Science from Princeton University in 2005. In 2011 he received his M.S. in Computer Science from George Mason University, where he is currently pursuing his Ph.D., also in Computer Science. His research interests include high-dimensional search and optimization, evolutionary algorithms, and clustering analysis. His work applies these interests to problems in computational biology. He is a member of the ACM and the IEEE.

Kevin Molloy received his B.S. and M.S. in Computer Science from George Mason University in 1998 and 2011. He is currently pursuing his Ph.D. in Computer Science. His research interests include computational biology, analytical performance modeling, and parallel computation. He is a member of the ACM.

S. Farid Hendi is pursuing a Ph.D. in Computer Science at George Mason University. He received his B.S. in Computer Engineering in 2009 at Isfahan University of Technology. His research interests include Computational Structural Biology and Robotics. His work focuses on robotics-inspired search algorithms that apply motion-planning concepts to problems related to protein modeling.

Amarda Shehu is an Assistant Professor in the Department of Computer Science at George Mason University. She holds affiliated appointments in the Department of Bioinformatics and Computational Biology and the Department of Bioengineering at George Mason University. She received her B.S. in Computer Science and Mathematics from Clarkson University in Potsdam, NY, and her Ph.D. in Computer Science from Rice University in Houston, TX, where she was an NIH fellow of the Nanobiology Training Program of the Gulf Coast Consortia. Shehu's research contributions are in computational structural biology, biophysics, and bioinformatics with a focus on issues concerning the relationship between sequence, structure, dynamics, and function in biological molecules. Shehu is the recent recipient of an NSF CAREER award for her research on probabilistic search algorithms for protein conformational spaces.