Development of a Four-Body Statistical Pseudo-Potential to Discriminate Native from Non-Native Protein Conformations

Bala Krishnamoorthy*
Dept. of Operations Research
CB 3180, UNC Chapel Hill NC 27599 USA

Alexander Tropsha†
Laboratory for Molecular Modeling
School of Pharmacy
UNC Chapel Hill NC 27599 USA

Abstract

Motivation: Most scoring functions used in protein fold recognition employ two-body (pseudo) potential energies. The use of higher-order terms may improve the performance of current algorithms.
Methods: Proteins are represented by the side chain centroids of amino acids. Delaunay tessellation of this representation defines all sets of nearest neighbor quadruplets of amino acids. Four-body contact scoring function (log likelihoods of residue quadruplet compositions) is derived by the analysis of a diverse set of proteins with known structures. A test protein is characterized by the total score calculated as the sum of the individual log likelihoods of composing amino acid quadruplets.
Results: The scoring function distinguishes native from partially unfolded or deliberately misfolded structures. It also discriminates between pre- and post-transition state and native structures in the folding simulations trajectory of Chymotrypsin Inhibitor 2 (CI2).
Availability : All codes are written in C/C++. Programs are available from the authors on request.
Contact : alex_tropsha@unc.edu

1 INTRODUCTION

Theoretical methods such as homology modeling, *ab initio* folding simulations and fold recognition (threading) are widely used to predict the protein structure from its primary sequence. The success of these methods ultimately depends on the accuracy of the underlying scoring function that should be capable of discriminating between correct (i.e. native) and incorrect configurations of the native polypeptide sequence. Development of such a function is not an easy task. For instance, Novotny and co-workers (Novotny et al.,1984) demonstrated that conventional molecular mechanics potential energy functions cannot accurately discriminate a native protein structure from its deliberately misfolded structure. More recently, free energy simulations that take into account explicit solvation effects have been used to successfully discriminate misfolded form native conformations for several proteins (Lazaridis & Karplus,1999;Vorobjev & Hermans,2001). Though reasonably accurate, these methods are relatively inefficient when its comes to the analysis of a large number of conformations. Thus, there is a need for a computationally efficient yet accurate scoring function that could distinguish native from non-native protein conformations.

Several empirical energy functions have been developed over the years (Miyazawa & Jernigan,1985; Sippl,1990; Park & Levitt,1996; Park et al.,1997). Most of these energy functions are derived from two-body interaction potentials where arbitrary distance criteria are used to define nearest neighbor or contact residues. Examples of such criteria include separation in the range of 4.5 to 7.5 Å (Miyazawa & Jernigan,1985), separation of C_α atoms by no more than 5.5 Å (Yee et al.,1994). Obviously, results of analyses using such definitions strongly depend on the chosen criteria of contact.

Earlier, we proposed a novel statistical pseudo-potential based on the four-body nearest neighbor propensities of amino acid residues (Tropsha et al., 1996; Tropsha et al.,1998). We used a computational geometry data structure known as Delaunay tessellation (DT) (Preparata & Shamos,1985), which uniquely partitions a set of points in 3D space into an aggregate of space-filling, irregular tetrahedra. Consequently, DT provides a robust and objective definition of nearest neighbors in sets of four points (in 3D space) identified as vertices of tetrahedra.

In this paper, we have developed a modified DT based scoring function. We have applied it successfully to discriminate native from the computer generated decoy structures listed in the Decoys 'R' Us database (Samudrala & Levitt,2000). The pseudo-potentials were also applied to a set of conformational samples from the

*email: kbala@unc.edu
†email: alex_tropsha@unc.edu (To whom correspondence should be addressed)

folding simulation trajectory of the protein CI2 (Li & Shakhnovich,2001), described as "pre-transition states" and "post-transition states" by the original authors. In our calculations, native, pre- and post-transition conformations were discriminated clearly.

This paper is organized as follows. We explain the ideas behind the formulation of the four-body pseudo-potentials in section 2. The methods used to develop the pseudo-potentials are discussed in section 3. Finally, the results from the application of the pseudo-potentials to various data sets are discussed in section 4. We discuss the application of the pseudo-potentials to random sequence decoys (section 4.1), single decoys (section 4.2.1), multiple decoys (section 4.2.2) and to the structures from the MD folding simulation of CI2 (section 4.3). Finally, we give the conclusions in section 5.

2 FOUR-BODY POTENTIALS

As the first step towards developing an objective four-body pseudo-potentials, we require a rigorous definition of the nearest neighborhood for each residue. Tropsha and coworkers (Tropsha et al.,1996) suggested the use of a computational geometry approach for this purpose. The properties of the Delaunay tessellation that make it ideal for the purpose of objectively defining nearest neighbors are discussed below.

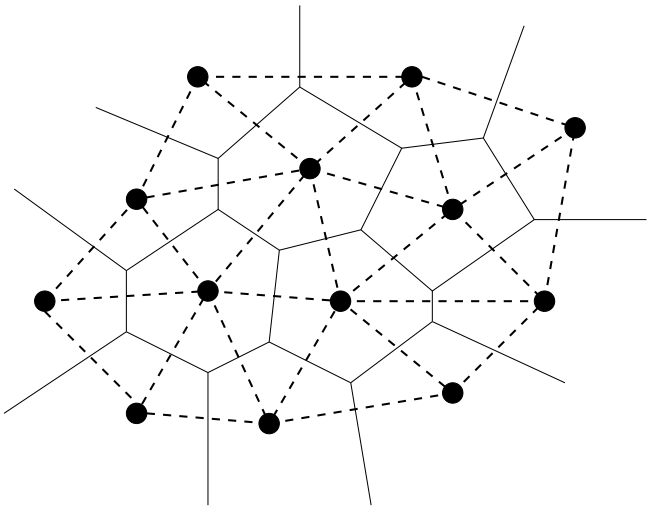

Figure 1: Delaunay triangulation (dotted lines) and the Voronoi diagram (solid lines) of a set of points in 2-D

2.1 Tessellation of Protein Structures

The nearest neighborhood for each point in a set of arbitrary points in space can be objectively identified using the Voronoi diagram of the points (Fig 1). The Voronoi diagram of a set of points consists of convex polytopes or polyhedra enclosing the points. In 3-D, the boundaries of each polyhedron are faces defined by planes normal to the lines joining the point enclosed and one of its nearest neighbors. In 2-D, the boundaries are straight lines (Fig 1). The region enclosed by the Voronoi polyhedron of a particular point is closest to that point than to any other point in the set. Hence, if the Voronoi polyhedra of two points share a common face (a line in 2-D), any point on that face is equidistant from both the original points. The Delaunay tessellation of the set of points is defined as the mathematical dual of the Voronoi diagram (Fig 1). The corresponding Delaunay tessellation is obtained from the Voronoi diagram by connecting all pairs of points whose Voronoi polyhedra share a common boundary (face). In 2-D, the Delaunay tessellation of a set of points would consist of a set of triangles while in 3-D, the Delaunay tessellation would consist of a set of space filling tetrahedra.

The tetrahedra given by the Delaunay tessellation in 3-D could be viewed as aggregates of four-body nearest neighbor clusters. We use the frequencies of the residue compositions of these clusters in the formulation of the four-body pseudo-potentials. This analysis ignores the presence of solvent molecules, metal ions, heme groups and other molecules complexed with the proteins.

2.2 Sequence-Structure Correlations for Proteins

In addition to the residue composition of the Delaunay tetrahedra, the chain separation of residues is an important factor in the definition of the four-body pseudo-potentials. In order to capture the correlation between the structure and the sequence of proteins, a classification of the Delaunay tetrahedra based on the relative positions of the vertex residues was introduced (Tropsha et al.,1996) Two residues are classified as *non-consecutive* if they are separated by at least one other residue in the protein primary sequence. The Delaunay tetrahedra occurring in the tessellations of protein structures belong to five classes (Fig 2). Class **{1,1,1,1}** has all the four residues non-consecutive in the primary sequence. In class **{2,1,1}** the tetrahedron has one pair of consecutive residues and the other two residues are both non-consecutive to the residues in the consecutive pair and to each other. In class **{2,2}** the four residues make two pairs of consecutive residues such that each residue is non-consecutive to both the residues in the other pair. Class **{3,1}** has three consecutive residues and the fourth residue is non-consecutive to these three. Finally, in class **{4}** all four residues in the tetrahedron are consecutive in the protein primary sequence. We denote these five classes of tetrahedra by a class value α that assumes the values $0, 1, 2, 3$ and 4 corresponding to the five classes in

the order listed above.

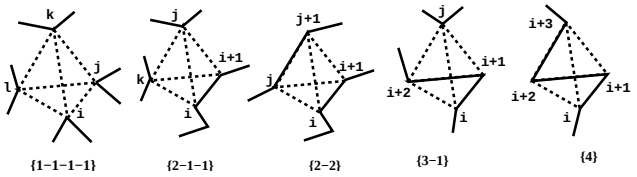

Figure 2: Five classes of Delaunay tetrahedra (see text for additional discussion)

From the training set of protein structures that we used to develop the four-body pseudo-potentials, the average frequencies of observation were 25.54 %, 35.61 %, 11.38 %, 22.05 % and 5.43 % for the classes {1,1,1,1}, {2,1,1},{2,2},{3,1} and {4}, respectively. The average number of occurrences of each of the 8855 quadruplet compositions recorded in the training set were approximately 30, 42, 13, 26 and 6 tetrahedra for the five classes, respectively. At the same time, some of the quadruplet compositions are unlikely to occur in nature due to the chemical structure of the contributing amino acids (e.g. four Lysines). For instance, 122 quadruplets are not observed at all for the class {1,1,1,1} in the training set protein chains. Detailed statistics on quadruplet frequencies as well as histograms of observations in the training set are available at http://www.unc.edu/∼kbala/DT/ .

2.3 Revised Formulation of the Four-Body pseudo-Potentials

Based on the ideas presented earlier (Tropsha et al.,1996,Tropsha et al.,1998) Gan, Tropsha and Schlick (Gan et al.,2001) calculated the four-residue contact energies Q^{α}_{ijkl} using the formula

$$Q^{\alpha}_{ijkl} = -K_B T \, ln \left[\frac{f^{\alpha}_{ijkl}}{p_{ijkl}} \right] \tag{1}$$

Here f^{α}_{ijkl} is the observed frequency of the residue composition $(ijkl)$ in a tetrahedron of type α over a set of protein structures (the training set used to develop the pseudo-potentials). p_{ijkl} was defined as the expected random frequency of observing the same residue combination $(ijkl)$. The expected random frequency term defined here does not discriminate the five different classes of tetrahedra. At the same time, the observed frequency term in the numerator would be calculated separately for each tetrahedron class α. Hence the expected random frequency term was modified to p^{α}_{ijkl} and correspondingly defined as the expected random frequency of observing residues i, j, k and l in a type α tetrahedron. The pseudo-potentials are no longer considered as contact energies or potentials of mean force. Hence the term $-K_B T$ would just be a constant multiplier (temperature is not considered as a discriminating factor). Furthermore, conventional potentials typically take the logarithms to the base 10. Taking into account all these considerations, we define the four-body pseudo-potentials as a log-likelihood ratio given by

$$Q^{\alpha}_{ijkl} = log \left[\frac{f^{\alpha}_{ijkl}}{p^{\alpha}_{ijkl}} \right] \tag{2}$$

Any possible subset of all the five classes of tetrahedra could be used in combination to develop total log-likelihood scores. Using only the type $\alpha = 0$ tetrahedra (class {1,1,1,1} tetrahedra) gives non-bonded log-likelihood scores as reported elsewhere (Carter et al.,2001).

Given a representative set of protein structures (the training set) the observed nonrandom frequencies f^{α}_{ijkl} and the expected random frequencies p^{α}_{ijkl} are calculated as follows:

$$f^{\alpha}_{ijkl} = \frac{\text{number of type } \alpha \text{ quadruplets with composition } (ijkl)}{\text{total number of } \alpha \text{ quadruplets}} \tag{3}$$

and

$$p^{\alpha}_{ijkl} = C \, a_i a_j a_k a_l \, p_{\alpha} \tag{4}$$

where

$$a_i = \frac{\text{number of amino acids of type } i \text{ in the data set}}{\text{total number of amino acids in the data set}} \tag{5}$$

Here, a_i is the estimate of the probability (frequency) of observing a residue of type i in the data set and p_{α} is the estimate of the probability of observing a type α quadruplet in the data set, independent of the residue composition. It is calculated as follows:

$$p_{\alpha} = \frac{\text{number of type } \alpha \text{ quadruplets in the data set}}{\text{total number of quadruplets in the data set}} \tag{6}$$

The p_{α} values were listed as percentages at the end of section 2.2. C is a combinatorial factor that is required because we treat the tetrahedron with the composition $(ijkl)$ and the one with the composition $(iklj)$ or $(jilk)$ as having the same quadruplet composition. This combinatorial factor C is defined as (Tropsha et al.,1996;Tropsha et al.,1998)

$$C = \frac{4!}{\prod_{\nu=1}^{\eta} t_{i_\nu}!} \tag{7}$$

where η is the number of distinct residue types in the quadruplet ($1 \leq \eta \leq 4$) and t_{i_ν} is the number of residues of type i_ν in the quadruplet.

3 METHODS

All calculations reported in this paper were done using programs written in C/C++. The code titled *nnsort.c*

available from D. F. Watson (Watson,1992) was adapted to calculate the Delaunay tessellation of proteins. The four-body pseudo-potentials were developed for the complete set of 20 native amino acids as well as for a reduced 6-type definition as suggested earlier (Gan et al.,2001) using both C_α atoms and the sidechain centers of residues as representative points. Thus, four different sets of values for Q^{α}_{ijkl} were generated for different combinations of amino acid representations and the residue types used (C_α or sidechain centers and 20-type or 6-type).

3.1 Selecting the Training Set

A diverse set of proteins representing different protein families was selected from the Brookhaven Protein Data Bank (PDB) for developing the pseudo-potentials. The list of protein chains listed in the CulledPDB directory (Wang & Dunbrack,2002) with a resolution of 2.2 Å and a maximum sequence identity of 30 % was used. This list had 1653 chains. To make the training set more exhaustive, another set of protein chains with a resolution of 2.5 Å and a maximum sequence identity of 25 % was selected from the Sequence Unique Database used by WHAT IF (Whatif). This list had 967 chains. A combined list was obtained by taking the union of these two lists. This had 2069 chains with a total chain similarity of at most 30 % and a resolution of 2.5 Å.

The protein chains in the combined list were parsed rigorously for the following types of errors:

1. Residue numbers are not consecutive.
2. Individual residue composition is incorrect.
3. Gaps in the sequence (indicated by $C_\alpha - C_\alpha$ distances higher than 4.2 Å between consecutive residues).

Many of the protein chains considered were found to be defective due to one or more of the above types of errors. All such chains were removed from the list. This left 1563 chains for the analysis using C_α's and 1167 chains for the analysis using sidechain centroids. The rigorous parsing ensured that the pseudo-potentials were developed using chains with no irregularities. On the other hand, parsing reduced the size of the training set. Nevertheless, leaving out certain chains from the training set did not affect the sequence identity of the training set too much as the errors in the PDB files were not spotted in any specific fold family alone. Also, combining the two lists did not affect the accuracy of the developed pseudo-potentials either. Log-likelihood scores developed using only the first subset of proteins were highly correlated with those developed using the combined list (the highest correlation coefficient of 0.98 was observed for class {1,1,1,1} and lowest was 0.92 for class {4}). At the same time, some of the quadruplet compositions that were not observed in the subset of proteins were observed in the combined list. Details of another experiment to validate the selection of the training set are available at http://www.unc.edu/∼kbala/DT/ .

3.2 Calculating the pseudo-Potential Values

The results of Delaunay tessellation as applied to the training set structures were additionally processed as follows. A cut-off distance of 10 Å was used to discard tetrahedra for which at least one of the edges was longer than 10 Å. The number of occurrences of $(ijkl)$ tetrahedra were counted for each amino acid residue composition ($(ijkl)$). Counts of the types of tetrahedra were also recorded. The four-body statistical pseudo-potentials were calculated using the formulae given before. The log-likelihood ratios were found to have the widest range of values for the type $\alpha = 0$ (class {1,1,1,1} or non-bonded) tetrahedra. At the same time, class {1,1,1,1} tetrahedra were only the second most frequently observed class of tetrahedra (as noted by the relative frequencies of occurrence for each class of tetrahedra mentioned earlier).

For any test protein, the log-likelihood ratios (Q^{α}_{ijkl}) corresponding to each tetrahedron in its Delaunay tessellation were added up to give the total score. This would be labeled as the **20L5T** total log-likelihood score when using the exhaustive 20-type definition for the residues and **6L5T** total log-likelihood score when using the reduced 6-type definition for the residues. Here, the term **5T** indicates that all the five classes of tetrahedra are used to calculate the total score for the protein (as opposed to using just a single type of tetrahedra or two types of tetrahedra for scoring). The *structure profile* of the protein can be obtained concomitantly by plotting the sum of the four-body pseudo-potential values corresponding to all the tetrahedra that the individual residue participates in as a function of the residue number in the primary sequence (labeled as **20L5T-profile**). (cf. 3D-1D profiles proposed by (Wilmanns & Eisenberg,1993)).

3.2.1 Weights for the Five Classes of Tetrahedra

Our initial calculations on decoys demonstrated that overall, using the sidechain centers of residues to obtain the tessellations produced better results as compared to the use of C_α's. Still, the **20L5T** pseudo-potentials were unable to discriminate clearly the native structure from its decoys in some cases. It was noticed that the frequencies of occurrence of the five classes of tetrahedra quite often varied from the average values that were observed for the whole training set. A deviant behavior from the average as far as the distribution of the classes of tetrahedra were concerned would be expected to pro-

duce unfavorable results. Hence we decided to penalize the deviations from the expected (average) frequencies of distribution for the five classes of tetrahedra. This was done by weighing the contributions to the total log-likelihood score from each class of tetrahedra by a factor that reflected the deviant behavior. For each individual protein, the fractions of each class of tetrahedra were calculated. Then, the ratios of these fractions to the overall average frequencies (obtained for the whole training set, listed before) were calculated. If any of these ratios turned out be larger than one, it was inverted. The log-likelihood scores from each class of tetrahedra were weighed by these ratios and added up to get the total score. This total score was labeled as the **W-20L5T** total log-likelihood score. Hence, if the fractions of the five classes of tetrahedra turned out to be exactly the same as the average values, the weights calculated would all be equal to one. So, the **W-20L5T** score and the **20L5T** score would be identical in this case. The weights calculated here would all be less than or equal to one. A smaller weight (much less than one) would indicate an observed frequency that is much different from (greater or smaller) the average frequency for that class of tetrahedra. The **W-20L5T** scores were observed to be much more effective than the **20L5T** scores (using the sidechain centers of the residues) in discriminating native structures from the corresponding decoys.

4 RESULTS and DISCUSSION

All the results discussed here used sidechain centers of the residues for tessellation. The total scores reported here are **W-20L5T** scores. The pseudo-potentials performed the best overall when using the aforementioned scores. Hence results using other scores (**20L5T** or **6L5T**) are not discussed here.

4.1 Testing the pseudo-Potentials on Random Sequence Decoys

As a first step towards testing the four-body pseudo-potentials, a series of random sequence decoys were generated. A separate diverse set of protein chains was chosen from the CulledPDB (Wang & Dunbrack,2002) database for this study. A set of 336 chains was chosen with maximum residue sequence identity of 20% and resolution of 1.6 Å. These chains were parsed in the same way as described in Section 3.1. This left a total of 217 chains in the set.

For each protein in the set, 1200 random permutations of the sequence were considered. Hence, the structure was not altered in the sense that the coordinates of the sidechain centers for each residue were kept the same. Since we are generating these sequences by randomly sampling from all the possible permutations of the residue identities, the resulting structures would be highly improbable to occur in nature. Hence the four-body pseudo-potentials would be expected to rank the native structure much higher than all its random sequence decoys. This indeed was the case except for three protein chains. We calculated the Z-scores for the native structure in each case averaged over its 1200 decoys using the **W-20L5T** total scores. The average Z-score was 10.85 and except in the case of three chains (out of the total of 217 chains studied) the Z-score was higher than 2.3. The smallest Z-score was 1.16.

4.2 Testing the pseudo-Potentials on Structural Decoys

The *Decoys 'R' Us* database (Samudrala & Levitt,2000) maintains a list of decoy structures whose main use is in testing energy or scoring functions for protein structures. *Decoys* are computer generated conformations of proteins that possess some characteristics of native proteins, but are not biologically real. The four-body pseudo-potentials were applied to the single and the multiple decoy sets available in the database. Single decoy sets have one incorrect conformation given for each native (correct) conformation. The multiple decoy sets list a range of conformations with varying root mean square deviations (RMSD) from the native conformation. The primary objective is to distinguish the non-native conformations from the native one.

4.2.1 The Single Decoy Set

The single decoys set has misfolded conformations listed for 26 native chains. The **W-20L5T** total scores for the native and the corresponding misfolded conformations are shown in Fig 3. The total scores for the native conformations were observed to be consistently higher than the total scores for the non-native conformations. On average, the total scores of the misfolded conformations were lower than that of the corresponding stable conformations by 29.6 % .

4.2.2 The Multiple Decoy Sets

The *Decoys 'R' Us* database provides multiple decoy structures for a set of proteins. For each native conformation, there are multiple non-native conformations which fall in a range of root mean square deviations (RMSD) from the native structure. The decoys generated using different methods are classified separately (labeled *lattice_ssfit, 4state_reduced, lmds, fisa* and so on). Decoys are generated for a series of native proteins using each method. Rank scores and Z-scores of the native structure

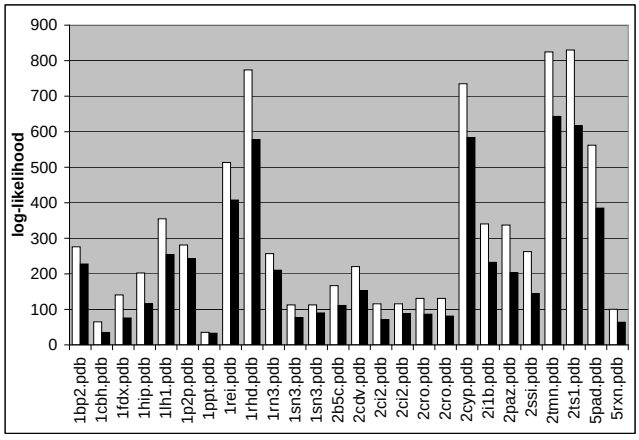

Figure 3: **W20L5T**-Total scores for the *mis_fold* single decoys (from the Decoys 'R' Us database). The white bars represent the scores for the native structures and the black bars give the total scores for the corresponding decoys.The native conformations score higher than their decoys in all cases

among its decoys as well as energy-RMSD plots for the native and decoy structures have been commonly used to test the effectiveness of potential functions (Park & Levitt,1996;Park et. al.,1997). The Rank scores and average Z-scores were calculated for the X-ray structure in each case using the **W-20L5T** total scores. The results are given in Tables 1, 2, and 3. Results are not given for the few sets of decoys for which the X-ray structure is not known. Also, for comparing the four-body potentials with a conventionally used two-body potentials, we list the native Z-scores for the *4state_reduced* decoy set calculated using contact Miyazawa-Jernigan (MJ) potentials as listed in (Park & Levitt, 1996). Since a negative Z-score is considered better there, we list those Z-scores with their signs reversed (so that they could be directly compared with the native Z-scores given by the four-body potentials). The reader is referred to the reference cited above for a discussion on the performance of other conventional potential functions on the same decoy set.

As shown by the results, the four-body pseudo-potentials are able to successfully distinguish the X-ray structure from its decoys in most of the cases. Except in the case of the local minima decoy sets (**lmds**), the native structure is ranked the best (or at least one of the topmost) among all the decoys for all the proteins considered. The average performance (as indicated by the Z-scores) is also good. Note that since a higher total score would indicate a more probable structure, a high positive Z-score is considered most favorable here. In order to study the variation of the total scores with the RMSD of the protein structures, we plotted the **W-20L5T** total scores for the native structure (which is assigned an RMSD value of zero Å) and its decoys against their RMSD values for each of the decoys considered. The plot for **1ctf** from the *4state_reduced* decoys set is shown in Fig 4. There is an observable trend in decreasing the total score with increasing RMSD values in most of the cases. Plots for other decoy sets are available at http://www.unc.edu/∼kbala/DT/ .

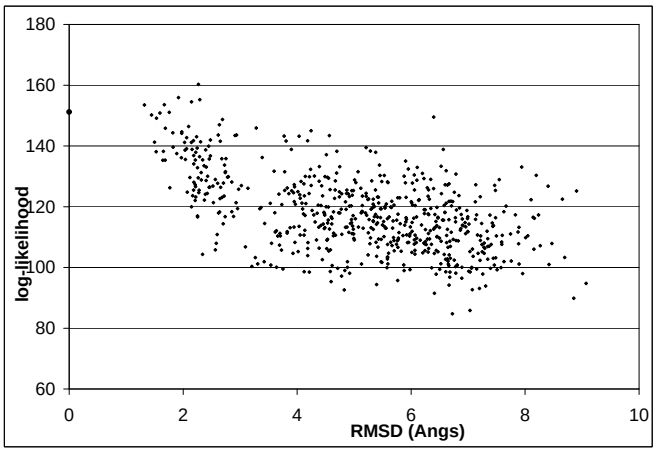

Figure 4: Total log-likelihood scores for 1ctf (indicated by the mark on the Y-axis) and its 630 decoys (from the *4state_reduced* decoys set).

4.3 Application to Folding Simulations

The four-body pseudo-potentials were applied to a set of conformational samples obtained from the molecular dynamics (MD) simulations of protein folding. Dokholyan et al. have applied the MD simulation technique along with the Gō model to identify a *folding nucleus* for the off-lattice protein model (Dokholyan et al.,2000). They reported the formation of a few well-defined contacts with high probability in the transition state ensemble of conformations. These contacts determine folding cooperativity and drive the model protein into its folded conformation.

4.3.1 Pre- and Post-Transition State Conformations

With the aim of identifying the folding nucleus, Dokholyan et. al. studied the conformations that appear in various types of folding $\rightleftharpoons$ unfolding fluctuations. From among a series of transition state conformations with comparable energies, two classes of conformations were distinguished - *pre*-transition state structures and *post*-transition state structures. Both these types of conformations are used as starting structures from which the simulations proceed further. Starting with a pre-transition structure gave an unfolded conformation with probability almost equal to one. On the other hand, starting with a post-transition structure almost always

Table 1: Native rank scores and Z-scores for the **4state_reduced** multiple decoy set. The number of decoys considered for each protein is given in parenthesis along with the native rank score. Except in the case of 1sn3, the four-body pseudo-potentials clearly distinguish the native structure from its decoys. It also outperforms the conventional two-body contact MJ potentials for each protein (as indicated by the corresponding worse native MJ Z-scores listed in the last column)

protein	native rank score	native Z-score	-(MJ Z-score)
1ctf	7 (630)	2.62	0.38
1r69	3 (675)	2.90	1.35
1sn3	113 (660)	1.04	0.13
2cro	1 (674)	3.04	0.52
3icb	1 (653)	2.90	-0.69
4pti	1 (687)	3.18	-0.58
4rxn	5 (677)	2.58	1.02

Table 2: Native rank scores and Z-scores for the **lattice_ssfit** multiple decoy set. 2000 decoy structures were considered for each protein.

protein	native rank score	native Z-score
1beo	1	5.35
1ctf	1	4.18
1dkt-A	89	1.67
1fca	1	4.91
1nkl	1	4.38
1pgb	14	2.58
1trl-A	1179	-0.23
4icb	1	5.47

gave a fully folded (very close to the native structure) conformation. The possible explanation was that the formation of the folding nucleus is highly unfavorable in the pre-transition structures while it is almost formed already in the post-transition conformations.

The interesting fact to note here is that both pre- and post-transition structures had comparable total energies and were indistinguishable by the Gō potential (Dokholyan et al.,2000). A sample of six pre-transition state conformations and twenty post-transition state structures was selected from the MD simulation of **CI2** (chymotrypsin inhibitor) (the coordinates were kindly provided by Dr. Eugene Shakhnovich). The **W-20L5T** total scores for these conformations were compared with each other and with that of the actual X-ray structure (2CI2 was used). The results are given in Fig 5.

The total log-likelihood scores are observed to be lowest for the pre-transition conformations, higher for the post-transition structures and the maximum for the native structure. Among the six pre-transition structures sampled, one conformation (identified as the fifth one in Fig 5) has a markedly higher total score than the rest. This particular conformation actually had a higher probability of folding into a native-like structure (around 0.15) while the corresponding probabilities for all the other pre-transition conformations were observed to be zero (or very close to zero). On the other hand, all the post-transition structures had higher than 0.5 probabilities of folding into a native-like structure. For comparison, the above analysis was repeated using the conventional two-body Miyazawa-Jernigan (MJ) potentials. Though the native structure was distinguished from the rest, the MJ potentials could not clearly distinguish between the pre- and post-transition structures. The corresponding plot (using MJ potentials) is available at http://www.unc.edu/∼kbala/DT/ .

Table 3: Native rank scores and Z-scores for the **lmds** multiple decoy set. The number of decoys considered for each protein is given in parenthesis along with the native rank score. The pseudo-potentials do not do a spectacular job in all cases here.

protein	native rank score	native Z-score
1shf-A	28 (437)	1.48
1b0n-B	488 (497)	-1.93
1bba	205 (500)	0.20
1ctf	1 (500)	2.63
1dkt	4 (215)	2.06
1fc2	372 (500)	-0.71
1igd	189 (501)	0.32
2cro	1 (500)	3.88
2ovo	46 (352)	0.99
4pti	7 (343)	1.98

The **W-20L5T** structure profiles of these conformations were also studied. These profiles for a typical pre-transition conformation, a typical post-transition structure and the native structure are compared in Fig 6. A small number of residues are seen to have markedly high log-likelihood scores in the post-transition structure as compared to the pre-transition structure. The native residue profile dominates both the pre- and post-transition structure profiles with these particular residues having almost identical scores in the case of post-transition and the native structure profiles. The most significant among such residues seem to be 8L,13V,16A,20I,29I,47V,49L and 57I. Previous studies identified the residues 16, 49 and 57 as the most important ones in forming the nucleus (Itzhaki et al.,1995;Ladurner et al.,1997;Otzen & Fersht,1998). We propose that a few critical tetrahedra formed by these residues (identified by the log-likelihood profiles) act as *core tetrahedra* and once they are formed, the rest of the protein folds to a stable structure.

5 CONCLUSIONS

A revised four-body pseudo-potential that aims to distinguish correct structures from incorrect conformations has been developed. This potential is given as log-likelihood ratios for specific quadruplets of residues that fall into one of the five classes of tetrahedra (classified by the composing residue proximity in the primary sequence). The log-likelihood ratios are related to the probability of observing nonrandom clusters of nearest neighbor residues in the protein structures. The four-body pseudo-potentials were developed using a diverse data set of representative structures selected from the Protein Data Bank (PDB). The four-body pseudo-potentials

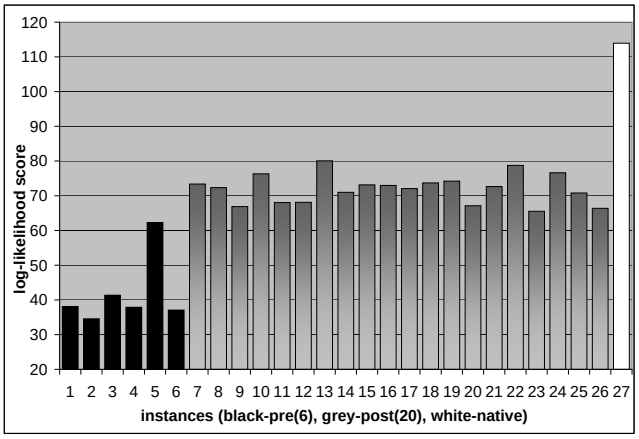

Figure 5: Total log-likelihood scores for six pre-transition(shown in black), 20 post transition conformations(shown in grey) and the native structure of CI2 (shown in white). The total scores for the pre-transition structures is lower than the total scores for the post-transition structures which in turn is lower than the total score for the native structure.

were tested on various sets of decoy structures. The non-native or misfolded structures were clearly distinguished from the stable native conformations for all the instances in the single decoy set. In the case of multiple decoys the pseudo-potentials successfully ranked the native structure higher than its multiple decoys in a majority of cases. The total scores also showed a decreasing trend with increasing RMSD values (from the native structure) for the decoy conformations in these cases.

The four-body pseudo-potentials were able to distinguish between pre- and post-transition conformations sampled from the molecular dynamics folding simulation of **CI2** despite the fact that all these conformations had comparable total energies (with the energy function used in the simulation). All these conformations were clearly distinguished from the most stable native conformation as well. The nucleation could be observed in terms of the formation of a few critical tetrahedra. This result could have far reaching implications as far as the dynamics of

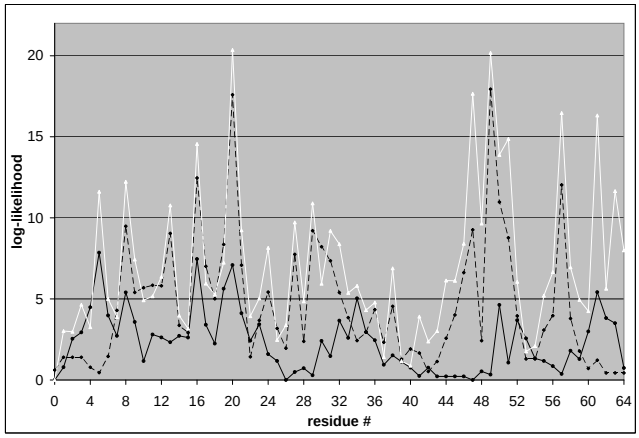

Figure 6: **W-20L5T** log-likelihood profile scores for a typical pre-transition structure (shown in black line), a typical post transition structure (shown in dotted lines) and the native structure of CI2 (shown in white line). The residues 8,13,16,20,29,47,49 and 57 are observed to have markedly high profile scores in the post-transition structure and in the native structure as compared to the pre-transition structure.

protein structure and folding are concerned. We are currently doing similar studies on other molecules. Preliminary results from the study of the MD simulation conformations of another protein (SH3) are available from http://www.unc.edu/∼kbala/DT/ .

In conclusion, we have demonstrated using a number of tests that the four-body pseudo-potential appears to be an accurate pseudo-energy function. A natural use of the four-body pseudo-potentials would be in fold recognition and folding simulation studies, which are currently ongoing in our laboratory. Certain geometrical parameters such as tetrahedrality, volume and shape of the Delaunay tetrahedra are also being studied to see whether they could make the pseudo-potentials more accurate.

6 ACKNOWLEDGMENTS

The authors thank Dr. Stephen Cammer and David Bostick for helpful suggestions and Dr. Eugene Shakhnovich and Dr. Nikolay Dokholyan for sharing the results of MD simulations. This project is supported by the grants from NSF (ITR/MCB 0112896) and North Carolina-Israel Binational Research Foundation (NCI-SRP # 1999032).

7 REFERENCES

Carter C.W. Jr, LeFebvre B.C., Cammer S., Tropsha A. and Edgell M.H. (2001), Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations, J. Mol. Biol.,**311**,625-638

Dokholyan N.V., Buldurev S., Stanley G.E. and Shakhnovich E.I.(2000), Identifying the protein folding nucleus using molecular dynamics, J. Mol. Biol., **296**,1183-1188

Gan H.H., Tropsha A. and Schlick T.(2001), Lattice protein folding with two and four-body statistical potentials, PROTEINS: Structure, Function and Genetics, **43**, 161-174.

Itzhaki L.S., Otzen D.E. and Fersht. A.R.(1995), The structure of the transition state for folding CI2 analyzed by Protein Engg.,J. Mol. Biol.**254**, 260-268.

Ladurner A.G., Itzhaki L.S. and Fersht A.R.(1997), Strain in the folding nucleus of chymotrypsin inhibitor 2, Folding and Design,**2**,363-368.

Lazaridis T. and Karplus M.(1999), Discrimination of native from misfolded protein models with an energy function including implicit solvation, J. Mol. Biol. **288**,477-487.

Li L. and Shakhnovich E.I.(2001), Constructing, verifying and disecting the folding transition state of CI2 with all-atom simulations, Proc. Natl. Acad. U.S.A.,**98**, 13014-13018.

Miyazawa S. and Jernigan R.L.(1985), Estimation of effective inter-residue contact energies from protein crystal structures: A quasi-chemical approximation, Macromolecules,**18**, 534-552.

Novotny J., Bruccoleri R. and Karplus M.(1984), An analysis of incorrectly folded protein models - Implications for structure predictions, J. Mol. Biol. **177**,787-818.

Otzen D.E. and Fersht A.R.(1998), Folding of circular and permuted CI2: Retention of the folding nucleus, Biochemistry, **37**, 8139-8146.

Park B. and Levitt M. (1996), Energy functions that discriminate X-ray and near-native folds from well-constructed decoys,J. Mol. Biol. **258**,367-392.

Park B., Huang E.S. and Levitt M.(1997), Factors affecting the ability of energy functions to discriminate correct from incorrect folds, J. Mol. Biol. **266**,831-846.

Preparata F.P. and Shamos M.I.(1985), Computational Geometry: An Introduction, New York: Springer-Verlag

Samudrala R. and Levitt M.(2000),Decoys 'R' Us: A database of incorrect conformations to improve protein structure prediction, Protein Sci., **9**, 1399-1401.

Sippl M.J.(1990), Calculation of conformational ensembles from potentials of mean force, J. Mol. Biol., **213**, 859-883.

Thomas P.D. and Dill K.A.(1996), Statistical Potentials extracted from protein structures: How accurate are they?, J. Mol. Biol. **257**, 457-469.

Tropsha A., Singh R.K. and Vaisman I.I.(1996), Delaunay tessellation of proteins: Four body nearest neighbor propensities of amino acid residues, J. Comput. Biol. **1996,3**, 213-222.

Tropsha A., Vaismann I.I. and Zheng W. (1998), Compositional preferences in quadruplets of nearest neighbor residues in protein structures: statistical geometry analysis, IEEE Symposia on Intelligence and Systems, 1998

Vorobjev Y.N. and Hermans J.(2001),Free energies of protein decoys provide insight into determinants of protein stability, Protein Sci. **12**,2498-2506.

Wang G. and Dunbrack R.L. Jr.(2002), PISCES: a protein sequence culling server, http://www.fccc.edu/research/labs/dunbrack/pisces/

Watson D.F.(1992), CONTOURING: A guide to the analysis and display of spatial data, http://members.iinet.net.au/w̃atson/software.htm, Pergamon Press ISBN 0 08 040286 0

Protein Data Bank - http://www.rcsb.org

Wilmanns M. and Eisenberg D.(1993), Three-dimensional profiles from residue-pair preferences: identification of sequences with beta/alpha-barrel fold, Proc Natl Acad Sci USA, Feb 15 **90(4)**, 1379-83.

Yee D.P., Chan H.S., Havel T.F. and Dill K.A.(1994), Does compactness induce secondary structures in proteins? J. Mol. Biol. **241**, 557-573.

Sequence unique database used by *What If* - http://www.cmbi.kun.nl/swift/whatif/select/.