Orientation-dependent potential of mean force for protein folding

Arnab Mukherjee, Prabhakar Bhimalapuram, and Biman Bagchi[a)]
Solid State and Structural Chemistry Unit, Indian Institute of Science, Bangalore, India 560 012

We present a solvent-implicit minimalistic model potential among the amino acid residues of proteins, obtained by using the known native structures [deposited in the Protein Data Bank (PDB)]. In this model, the amino acid side chains are represented by a single ellipsoidal site, defined by the group of atoms about the center of mass of the side chain. These ellipsoidal sites interact with other sites through an orientation-dependent interaction potential which we construct in the following fashion. First, the site–site potential of mean force (PMF) between heavy atoms is calculated [following F. Melo and E. Feytsman, J. Mol. Biol. **267**, 207 (1997)] from statistics of their distance separation obtained from crystal structures. These site–site potentials are then used to calculate the distance and the orientation-dependent potential between side chains of all the amino acid residues (AAR). The distance and orientation dependencies show several interesting results. For example, we find that the PMF between two hydrophobic AARs, such as phenylalanine, is strongly attractive at short distances (after the obvious repulsive region at very short separation) and is characterized by a deep minimum, for specific orientations. For the interaction between two hydrophilic AARs, such a deep minimum is absent and in addition, the potential interestingly reveals the *combined effect* of polar (charge) and hydrophobic interactions among some of these AARs. The effectiveness of our potential has been tested by calculating the *Z*-scores for a large set of proteins. The calculated *Z*-scores show high negative values for most of them, signifying the success of the potential to identify the native structure from among a large number of its decoy states. © *2005 American Institute of Physics*. [DOI: 10.1063/1.1940058]

I. INTRODUCTION

The protein folding problem, loosely defined as the determination of the native folded structure of a protein given its primary sequence, is a long-standing one with a huge literature of experiments,[1] simulation studies,[2–4] and studies of models.[5–10] A concise evaluation of the present status and recent advances in the problem has been presented by Dobson in a recent review.[11] A closely related problem is that of understanding the dynamics of folding, starting from an extended unfolded configuration of the protein. This understanding requires a thorough knowledge of the potential-energy landscape of the protein-solvent system and has a huge literature of its own.[12] The initial part of folding is well understood in terms of hydrophobic collapse and the drive for the secondary structure formation. However, it is the latter part of folding, formation of the native contacts, which has remained a subject of intense research. This latter (or, the final) part of folding is expected to depend critically on the interaction potential between the different amino acid residues and also the interaction of protein side chains with the surrounding solvent (mostly water). Clearly, this is a highly nontrivial problem because of the huge number of different interactions involved. One thus is forced to look for a simpler approach with an aim to understand some aspects of folding at least semiquantitatively. This has led to the development of solvent-implicit minimalistic models, which have greatly reduced the complexity of numerical approach.

The results of any computational study critically depends on the choice of the type of potentials (and their constituent parameters) and only a reasonably correct choice of these potentials to accurately describe the true potential surface of the experimental system assures that the dynamics probed by such a computational study are relevant and accurate.

The amino acid residues, which are the building blocks of the proteins, can be classified based on the properties of their side chain. In particular, a popular choice for such a classification is the polarity of the side chain, which can range from being nonpolar (hydrophobic) to polar (hydrophilic). The polar groups have charges (either positive or negative) and hence Coulomb potentials are used to account for these partial charges. The remaining atoms with no charges are usually described using the Lennard-Jones (LJ) potential. Such all atom simulations are computationally very expensive and—the 1-μs simulation of HP-36, a small 36 amino acid residue protein, by Duan and Kollmann[2] (which failed to fold the protein completely) is a testimony to the difficulty of carrying out fully atomistic simulations. Snow *et al.*, by introducing distributed computing procedure, were able to perform the very long protein folding studies in solvent-implicit models.[3]

Therefore, model studies using coarse-grained potentials, which remove the details of the less determinant inter-

[a)]Electronic mail: bbagchi@sscu.iisc.ernet.in

actions but retain the basic features are of particular interest and are shown to bring out several interesting aspects of protein folding, which are otherwise difficult to observe in all-atom simulations.

Potential of mean force (PMF) is one such approach to reduce the computational complexity of the protein folding problem. It can be used to coarse grain the system to obtain potential between groups of atoms by use of experimentally determined structures. Application of this idea to protein folding problem was pioneered by Sippl,[13] who developed a systematic technique to calculate the PMF from the experimentally determined native structures. Sippl[13] showed that the calculation of the interaction energy $E_{\alpha\beta}$ of a particular pair α, β with a distance separation r and a frequency $\rho_{\alpha\beta}(r)$ via the Boltzmann relation $E_{\alpha\beta}=-kT\ln(\rho_{\alpha\beta}(r)/\rho^{*}_{\alpha\beta})$ incorrectly uses the reference uniform density $\rho^{*}_{\alpha\beta}$. He suggested that the reference uniform density to be used should be the, $\rho^{*}(r)$, total pair frequency at that distance irrespective of the pairing. Determining the potential from equilibrium structure is an example of inverse problem and the goal of PMF approach is to solve this inverse problem satisfactorily so that the potentials thus obtained can be used to study a much wider class of proteins than used to determine these potentials.

Several minimalistic models have used the hydropathy scale to characterize the model interaction potential which is then used in computer simulations.[14,15] This scale uses the free energy of transfer of an amino acid from hexane to water. Therefore, such a scale does not do full justice to the protein environment which also needs to be considered in order to understand the role of hydrophobic effect in protein folding and stability, particularly because it is the late stage of folding (when the protein has already collapsed and partly folded) which controls the rate of protein folding. The statistical interaction potential model obtained from the analysis of the native folded structures deposited in the Protein Data Bank (PDB) provides a unique way to gauge the importance of hydrophobic interactions in the protein. In particular, one can study the distance and orientation dependence of interaction between, say, any two hydrophobic groups, or the same between hydrophilic and hydrophobic groups. Such an analysis can provide useful tests of the commonly assumed role of hydrophobicity in protein folding. We have carried out such an analysis in this work and the results are indeed interesting. We find that the interaction between two hydrophobic groups, such as phenylalanine–phenylalanine is strongly attractive at short distances while the same between two hydrophilic groups could be repulsive at all distances. The interaction potential shows substantial orientation dependence.

Such potentials based on the statistics of physically observable structures provide an excellent tool in understanding specific issues of particular interest to physical chemists in the field of protein folding. For example, the strength and the effect of interaction between a pair of aromatic residues is of particular relevance in understanding the issues of stacking. In addition, another relevant issue is that of cation-aromatic residue interactions. It is worth pointing out here that both the above-mentioned interactions and more specifically their strengths and a physical understanding of their effects can be quantified in our present calculation. The hydrogen bonds, both strong and weak, are thought to play an important role in providing kinetic pathways during the folding process. Such hydrogen bonds between various atoms present both in the side chain and in the backbone can also be observed and their strengths can be quantified in the present calculation.

The effectiveness of statistical potentials constructed from the knowledge of three-dimensional native folded crystal structures has been critically examined by Thomas and Dill.[5] They used the lattice HP model in which amino acids are classified as being either hydrophobic (H) or hydrophilic (P), and hence only three interactions, namely, HH, PP, and HP, need to be considered. Because of this simplification, the full configuration phase space can be searched to find the global minimum, which in turn makes the model particularly amenable to a full analysis for the free-energy surface. They find that the basic premise of these statistical potentials which consider the pairwise interaction to be independent of other pairs does not take into consideration the excluded volume (and hence the many-body interaction) at a pair–pair interaction level. For this reason, they report that the then existing database of native structures may have limited value as a predictive tool. In wake of this criticism, considerable effort has been invested in finding a more firm base for this approach. In this paper, to check the effectiveness of our potential, we have calculated the *Z*-scores which provide a useful quantitative measure of the validity of the computed potential in reproducing the stability of the native fold. For a large set of proteins, the calculated *Z*-scores show high negative values for most of them, signifying the success of the potential to identify the native state from among a large number of decoy states. Finally, the potential developed in this paper is compared with the orientation-dependent potential developed by Buchete, Straub, and Thirumalai (BST).[16]

The rest of the paper is arranged as follows. Section II gives an overview of this approach. In Sec. III, we describe the use of the PMF approach to calculate the orientation-dependent interaction between the side chains. In Sec. IV, we demonstrate the salient features of the constructed potential by considering a few representative pairs of amino acid side chains. Section V presents the calculation of the scoring functions for a large set of proteins for which decoy states have been calculated by Samudrala and Levitt[17] from which is seen that this potential can successfully discriminate the native fold from the decoy states. This section concludes with a comparison to the orientation-dependent potential developed by BST (Ref. 16) and the results are found to be quite comparable. Section VI gives the conclusive remarks.

II. OVERVIEW OF THE PMF APPROACH TO PROTEIN FOLDING

Although solvent-implicit minimalistic models have been in vogue in the protein folding problem for quite some time, the idea of estimating the potential of mean force as a *statistical potential* using known protein structures was first proposed by Tanaka and Scheraga.[18] Miyazawa and Jernigan advanced this approach by explicitly considering the solvent effect.[19] Sippl and co-workers[13,20] and others[21] extended

these methods to include the dependence on pairwise separation of residues in space and along the primary sequence. Bryant and Amzel[22] developed a log-linear statistical model to analyze protein structures separately, rather than using simple sums over distributions of residues in all proteins.

Apart from residue only-based potential of mean force approach, there have been efforts to understand the interaction of the protein residues through other kinds of interactions. Godzik and Skolnick proposed a residue triplet term.[8] Nishikawa and Matsuo[23] and Kocher *et al.*[24] proposed the PMF of dihedral angles and Nishikawa and Matsuo proposed the solvent accessibility and hydrogen bonding.[23]

Reva *et al.* proposed a new method to estimate the energy functions of protein. They divided the interactions in short- and long-range terms and used the Boltzmann statistics to calculate the potentials.[25] Gatchell *et al.* calculated the free-energy functions and applied them to the decoy sets of a particular protein to show the validity of the potential.[26] In all the above cases, the calculation of PMF employed a reference state that can be characterized as a residue (atom)-averaged state. Zhou and Zhou calculated the PMF for proteins for the first time with respect to a distance-scaled finite ideal gas reference state.[27] The PMF approach has been used to calculate the protein–protein interaction also.[28]

Scheraga *et al.* calculated the long-range and local interaction energy terms of proteins from the crystal structures, and then they optimized the relative weights of the energy terms so that the native structures of the selected proteins become the lowest-energy structures.[29] The potential generated by Liwo *et al.* is known as united residue force field, where the coarse graining is performed on the lesser important degrees of freedom. Melo and Feytsman calculated potential of mean force for an all-atom (heavy atoms) protein system.[30] They categorized the total 167 heavy atoms into 40 groups depending on the bond connectivity or the chemical environment of the atoms. Naturally, the statistics of PMF calculated by this procedure is much higher and this kind of grouping is chemically more relevant than the grouping of atoms based on the names of the residues. The potential generated by Melo and Feytsman showed a high selectivity towards the native topology compared to the misfolded decoys.[30]

The above-mentioned approaches of calculating PMF (both all-atom and for coarse grained models) considered primarily the distance dependence of the potential. Perhaps Liwo *et al.* first used Gay–Berne potential[31] to mimic the orientation dependence of interaction among the amino acid side chains. Recently Buchete *et al.*[16] calculated an orientation-dependent PMF from the crystal structures of proteins. They generated a right-handed coordinate system with C_α, C_β, and C_γ atoms for each residue and thus calculated a five-dimensional potential. Then they split that to a sum of three-dimensional potential, which depends on the all three polar coordinates.[16] In a more recent development of the model, they showed that backbone–backbone and backbone–side chain interactions are important. They fit their potential to spherical harmonic functions, which seems to be even better in recognizing the native structures of the proteins.[32]

In this work, we calculated an orientation-dependent potential among the amino acid side chains by employing a completely different approach. We initially calculate the all-atom PMF following the approach of Melo and Feytsman. Then we coarse grained the potential to get a residue-based potential, which depends on both the distance and the orientation.

III. CONSTRUCTION OF THE POTENTIAL OF MEAN FORCE

Protein is a heteropolymer formed from 20 different naturally occurring amino acids and these amino acids differ only in the "side chain" (a slight exception is the amino acid proline). The carboxylic group and the amino group of the two consecutive amino acids in the primary sequence form an amide bond, also called the peptide bond, and the final heteropolymer formed by repeated addition of amino acids by peptide bond is the protein. As can be clearly seen, the protein has an amino and a carboxylic group at the opposite ends of this chain (called the *N* and the *C* termini, respectively) and the primary sequence starting from either one of the termini, conventionally the *N* terminus, exactly and entirely specifies the protein. A particular type of coarse-grained minimalistic model of protein considers the whole of the amino acid side chain as a single spherical "side-chain atom" and calculate the potential of interaction between the different such side-chain atoms using the statistics from the experimentally determined structure of the proteins (see work of Sippl[13,33] and Eyrich Friesner,[34] among others). As can be seen, the statistics can be poor leading to inaccurate potentials.

Combining the bonded hydrogen atoms with the heavy atoms, the 167 heavy atoms of the 20 amino acids, can be grouped into 40 different "sites" which depend only on their chemical environment. Of these 40 different sites, 35 sites are exclusively located in the side-chain part of the amino acid residues, and the remaining five exclusively belong to the backbone of the protein.[30,35] Melo and Feytsman[30,35] calculated the site–site potential of mean force (U_{ss}) between these various sites from statistics of distances between them calculated from the crystal structures of a set of proteins. We have followed the same procedure to first calculate the site–site potential using dynamical averaging method of Sippl[13] with a bin size of 0.5 Å. A clear advantage of site grouping of over the side-chain atom grouping is essentially the much better statistics for the PMFs (the former about 15 times better statistics[30]) and a more firm chemical basis of the site grouping. However, the implementation of the site potential in protein folding requires essentially an all atom model to be considered. So the increased accuracy of the PMFs is at the cost an increased computational effort required to fold the protein into its native state. To reduce this computational effort, calculation of the PMFs between side chains (which can possibly have include an orientation dependence) from the considerably more accurate site–site PMFs become an attractive choice.

With the assumption that the side chains of the amino acid residues in the proteins can be suitably represented by

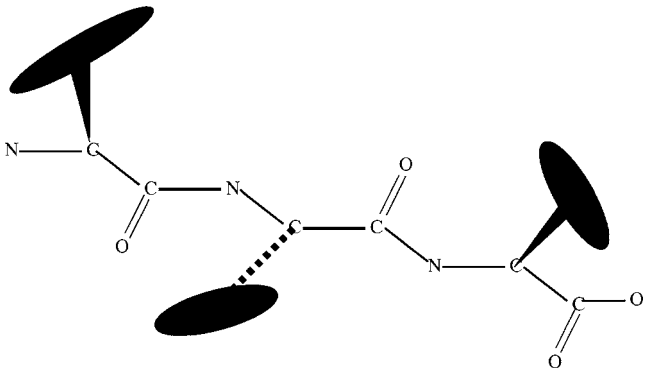

FIG. 1. The basic model of a protein is shown. All nonhydrogen backbone atoms are considered. All the atoms of the side chain are represented by a single ellipsoid.

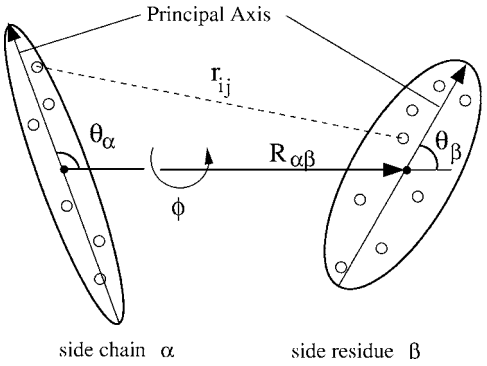

FIG. 2. The construction of four-dimensional ellipsoidal potential of mean force is shown schematically. $\mathbf{R}_{\alpha\beta}$ is the center-of-mass (COM) separation vector between the two ellipsoids. θ_α and θ_β are the corresponding angles between the principal axis of ellipsoids α and β and the COM separation vector $\mathbf{R}_{\alpha\beta}$ of the two ellipsoids. The circles on the ellipsoids denote heavy atom sites constituting the particular amino acid side chain represented by the ellipsoid. r_{ij} is the distance of separation between the sites i in ellipsoid α to another atom j in ellipsoid β. ϕ is the torsion angle between the principal axes of the two ellipsoids joined by their COM separation vector.

ellipsoidal sites (except for few discussed below), our model for the protein consists of the backbone represented by its constituent sites and the side chains by the ellipsoidal sites and is schematically shown in Fig. 1. We also assume that all interactions between various sites are pairwise additive and that higher-order interactions have no significant contributions. For the ellipsoidal side-chain sites, the specification of the principal axis and the various angles leave the orientation of the two minor axes unspecified and ambiguous. To remove this ambiguity in the model, the energy of interaction of the ellipsoidal side-chain site is calculated as an average of all the possible orientation of these minor axes with the Boltzmann factor as the weight function of such an average. For example, to calculate the interaction energy of a pair of side chains (denoted α and β), first the major axis is calculated by diagonalizing the moment of inertia matrix and its orientation of the axis by convention is chosen to approximately match the vector starting from C_β to the center of mass of the side chain (such a diagonalization needs to be performed only once for each side chain in a simulation). Next the angles made by the main axes of the ellipsoids with the intermolecular separation vector (θ_α and θ_β, respectively) and the torsion angle of the ellipsoids in consideration ϕ are calculated. For a particular orientation of the minor axes, which in turn fix the position of all the heavy atom sites of each side chain, the energy of this particular arrangement is calculated as pairwise sum of site–site interaction as (see Fig. 2)

$$U_{\alpha\beta}(\mathbf{R}_{\alpha\beta},\theta_\alpha,\theta_\beta,\phi)=\sum_{i=1}^{N_\alpha}\sum_{j=1}^{N_\beta}U_{s_\alpha(i)s_\beta(j)}(|\mathbf{r}_\alpha(i)-\mathbf{r}_\beta(j)|). \quad (1)$$

In this study, $U_{\alpha\beta}$'s are calculated for a distance separation of 0.2 Å and an angular separation of 30°. For all intermediate points, the values have been obtained by interpolation. Finally the ellipsoidal potential $U^e_{\alpha\beta}$ is calculated by averaging the energies with Boltzmann factor weight for all possible orientation obtained by rotation with respect to the principal axis, i.e.,

$$U^e_{\alpha\beta}(\mathbf{R}_{\alpha\beta},\theta_\alpha,\theta_\beta,\phi)=\frac{\sum U_{\alpha\beta}(\mathbf{R}_{\alpha\beta},\theta_\alpha,\theta_\beta,\phi)\exp[-U_{\alpha\beta}(\mathbf{R}_{\alpha\beta},\theta_\alpha,\theta_\beta,\phi)]}{\sum \exp[-U_{\alpha\beta}(\mathbf{R}_{\alpha\beta},\theta_\alpha,\theta_\beta,\phi)]}, \quad (2)$$

where the Σ indicates the summing over all possible rotations about the principal axes. This ensures that the chemical properties of the side chains become symmetrical about the principal axes and that the calculated $U^e_{\alpha\beta}$ are pertinent to the present ellipsoidal model.

For some amino acids, side chain cannot be represented as ellipsoids. For example, glycine does not have a side chain. Alanine has only C_β atom in the side chain, so it is modeled as a spherical atom. Serine and cystine have only two atoms in the side chain, and their principal axes are taken to be $C_\beta \rightarrow$terminal atom vector.

IV. PAIRWISE INTERACTIONS IN PROTEIN

We note here again that the usual "textbook" classification of amino acid residues is based on the hydrophobicity/hydrophilicity of the side chains. In this section, we demonstrate the salient features of the orientation-dependent PMF constructed in the previous section by presenting the interaction between pairs of side chains usually characterized as hydrophobic or hydrophilic. Hydropathy index, which combines the hydrophobicity and hydrophilicity of the side chains of the amino acids, is commonly used for such a classification. Hydropathy index for a few selected amino acids is given in Table I. Here we demonstrate the PMF constructed for the various kinds of interactions seen in the protein, namely, the interactions between the backbone atoms, the interactions of the backbone atoms and the side chains, and finally those between the side chains. Section IV A shows the interaction between the backbone atoms and Sec. IV B between the backbone atoms and the ellipsoidal side chains. For the pair interaction between side chains, of the 400 possible pairs, we choose the following amino acid pairs which form a representative sample demonstrating the

TABLE I. Hydropathy index for a few amino acids. The negative values indicate that the amino acid is hydrophilic and the positive values indicate that it is hydrophobic.

Amino acid	Hydropathy index
Phenylalanine	2.8
Lysine	−3.9
Arginine	−4.5
Glutamate	−3.5

essential features of the pairwise interactions of amino acid side chains in the protein environment: PHE–PHE, LYS–LYS, GLU–GLU, LYS–GLU, and ARG–PHE and are presented in consecutive subsections. For all the pairing of side chains and for all combinations of the orientation of pairs, we note couple of common features, the harsh short-ranged repulsion is essentially determined by the steric factors and the long-ranged interaction asymptotically decays to zero. For simplifying the following analysis, we fix the torsional angle ϕ at zero and look at the dependence of interaction on the distance of separation between the residues for some specific values of angles θ_α and θ_β.

A. Pair interaction between backbone atoms

The backbone atom of the protein can participate in hydrogen bonding with other atoms in the protein and the surrounding solvent water. The hydrogen bonding of backbone atoms with other intraprotein backbone atoms is the major reason for formation and stabilization of various secondary structures such as α helix, β sheet, etc. The stabilization of these secondary structures hence shows in the stabilization of the backbone atom pairs oxygen–nitrogen and nitrogen–oxygen in Fig. 3. The O–N and N–O interactions are *not* the same due to the asymmetry of the protein chain and while the O–N pair has a stability of about $2k_BT$ the N–O pair has a much larger stability of $3.5k_BT$. Sippl *et al.*[36] showed that in the construction of O–N and N–O PMF's, if only statistics of the residue separation of eight or more are considered, then the O–N and N–O interaction curves fall exactly on each other, thus showing that the chain asymmetry for the backbone atoms is present only for short sequence separation. In this work, the potential is taken to be asymmetric. This asymmetry becomes negligible for large separation, as shown in Fig. 3.

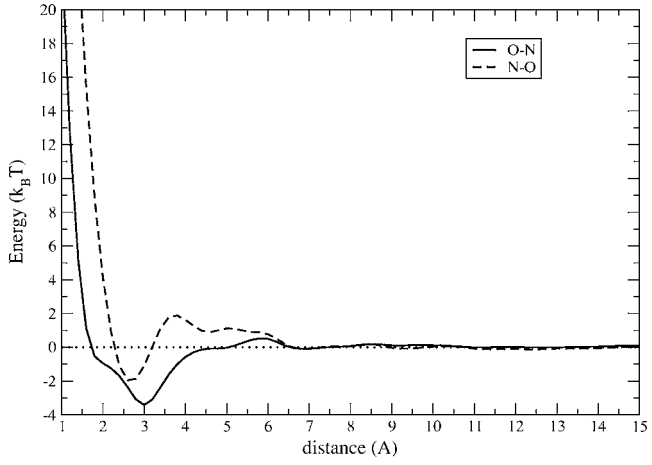

FIG. 3. The interaction between the backbone atoms of the protein chain.

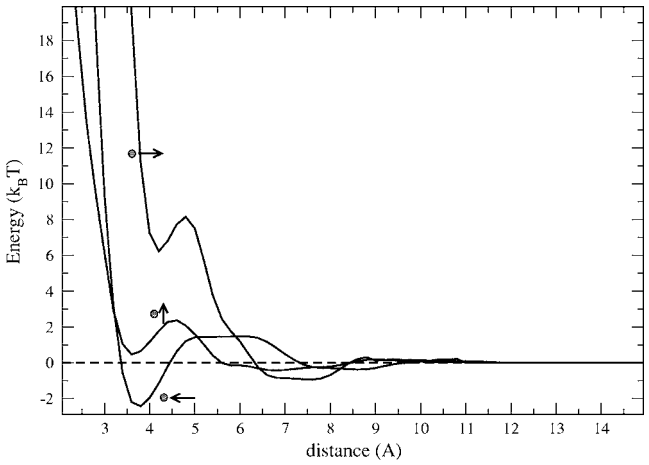

FIG. 4. Interaction between the ellipsoidal glutamate site and the nitrogen (N) atom of the backbone. The filled circle represents the N atom. The arrow indicates the direction of the major axis of the side chain and is approximately in line with the vector starting from the C_β to the COM of the individual side chain. The arrow, in this figure, represents the glutamate side chain.

B. Pair interaction between a backbone atom and a side chain

As described in Sec. IV A, the backbone atoms can form hydrogen bonds and few of the side chains share the same ability, too. While the hydrogen bonds formed among the backbone atoms are known to stabilize the local secondary structures, the hydrogen bonds formed by backbone atom *and* side chain can stabilize both the local secondary structures and also have a big contribution in interaction/association of two different secondary structures and hence to the formation of tertiary and quaternary structures. Figure 4 shows the interaction of the backbone nitrogen (N) atom with the glutamate side chain. It is clear from the figure that the orientation of the side chain in which the negatively charged oxygen atom of glutamate and the backbone N atom are close is stabilized due to the formation of a hydrogen bond with an energy stabilization of $2.5k_BT$. Similarly, between the arginine side chain and the backbone oxygen atom interaction is stabilized by $3.5k_BT$ in Fig. 5.

C. Phenylalanine–phenylalanine (PHE–PHE) pair interaction

First, we consider the PHE–PHE pair. Phenylalanine is a hydrophobic amino acid and the pair interaction PHE–PHE can be considered a major test for the "correctness" of the constructed potential. The interaction potential of this pair in a protein as a function of distance is shown in Fig. 6 for a range of orientations and has a few interesting features. Except for a single arrangement where the C_β's are packed too closely, at short separation the remaining arrangements have a pronounced minimum and hence a large stabilization. Most of the considered arrangements show a large stabilization, indicating that both the aromatic interaction[37] and the hydro-

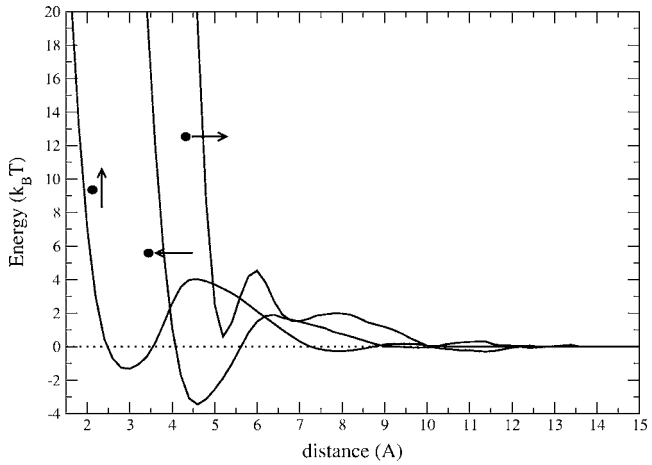

FIG. 5. Interaction between the ellipsoidal arginine side chain and the backbone oxygen atom. The filled circle represents the oxygen atom and the arrow represents the arginine side chain (see also caption of Fig. 4).

phobic interaction together play a significant part. It is also possible that the quadrupolar/induced-dipole interactions also contribute to this stabilization. Finally, we note that the magnitude of stabilization depends on the extent of contact between the pairs and the distance of shortest approach. It is also interesting to note that the arrangements which pack closely the C_β's of the side chains when compared to the arrangements that keep C_β's apart but have the same area of contacts, the latter arrangements have a stronger stabilization. This points to the importance of steric constraints and packing of the protein interior. In conclusion, to the discussion for this pair, we can say that the constructed PMF gives a physically intuitive understanding of the packing and its energetics for pairwise interactions between hydrophobic side chains in the protein.

D. Lysine–lysine (LYS–LYS) pair interaction

Figure 7 presents the results of the analysis of LYS–LYS arrangements. Firstly for all the considered arrangements for the two residues, the interaction is largely repulsive but with

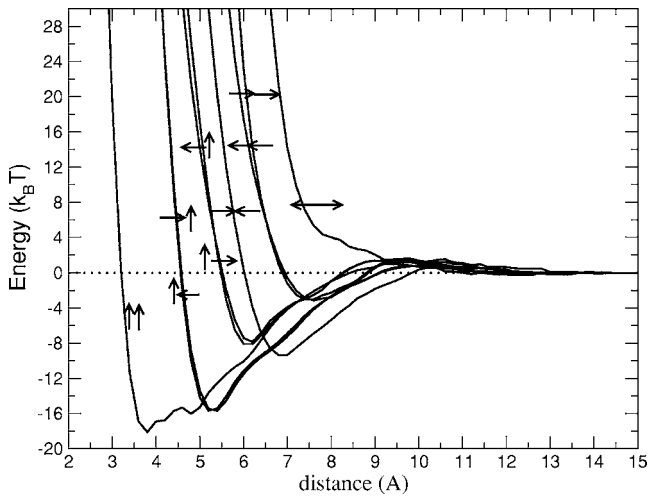

FIG. 6. The interaction of the phenylalanine–phenylalanine pair plotted as a function of the distance for different θ_α and θ_β orientations (also see caption for Fig. 4).

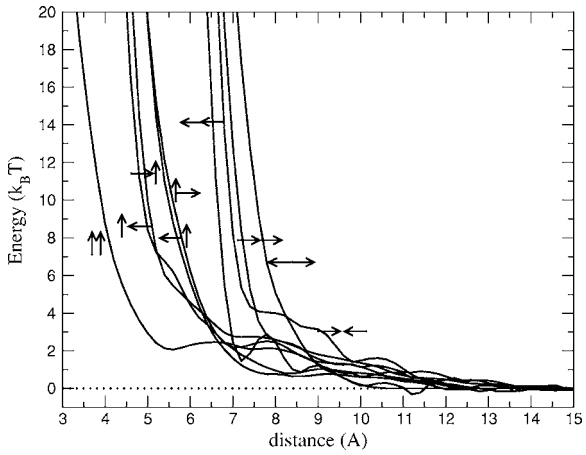

FIG. 7. Energy of interaction of lysine–lysine pair. Other details are same as in Fig. 6.

the long-range repulsion showing a much weaker dependence than $1/r$, indicating that the other residues in the surrounding protein core provide a very strong screening of the electrostatic interaction. The special case where the residues maximize the contact area (parallel configuration) has a smaller and a softer core and more interestingly shows a plateau at intermediate distances, before showing the typical long-ranged repulsion. This plateau clearly indicates that apart from the electrostatic interactions, the hydrophobic interactions are slightly repulsive at this distance contributes significantly to the softening of the repulsive core and the long-ranged behavior of the repulsion.

E. Glutamate–glutamate (GLU–GLU) pair interaction

The interaction between two glutamate side chains, a negatively charged hydrophilic residue, is shown in Fig. 8. The interaction shows qualitatively different behavior from the LYS–LYS pair interaction. The GLU–GLU pair interaction is energetically stabilized at short distances for some arrangements, while for the remaining configurations typical repulsive behavior is seen. Glutamate at physiological pH has its carboxylic group deprotonated making the side chain

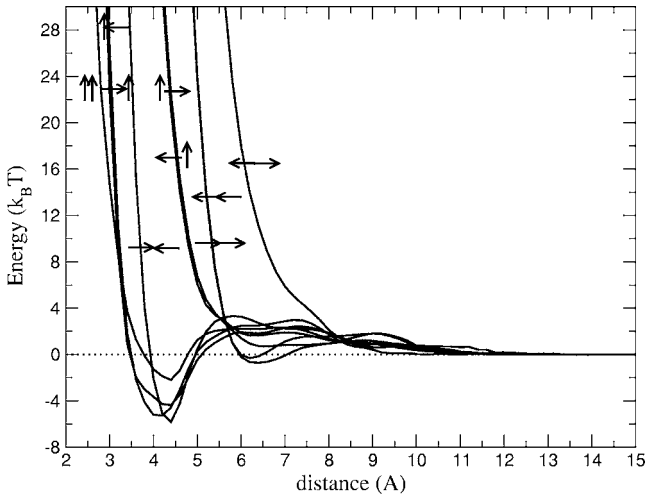

FIG. 8. Energy of interaction of glutamate–glutamate pair. Other details are same as in Fig. 6.

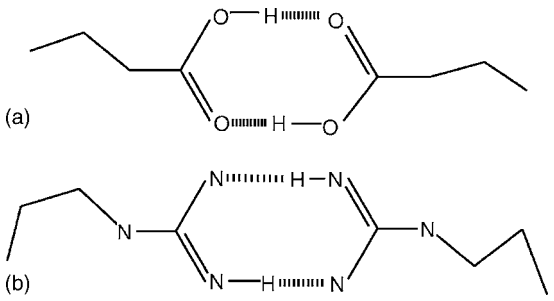

FIG. 9. Schematic diagram showing the possible hydrogen bonding for two side chains of same AAR when the ellipsoidal side-chain sites are arranged in antiparallel fashion. The hydrogen bonds are denoted by the dashed lines and the hydrogens not explicitly involved in the hydrogen bonding are not shown for clarity. (a) Between two glutamate side chains, assuming that both the COO^- groups are protonated, two hydrogen bonds are possible. (b) Between two arginine side chains, the two possible hydrogen bonds are shown.

of this residue negative in charge. Keeping this in mind, the stabilization of arrangements, which bring the negative charges to contact distances irrespective of the extent of the surface contact, is a slight anomaly. Since the stabilization is not considerably different for arrangements with varying surface contact areas of the side chains, the effect of hydrophobic interactions can be safely ruled out. This leaves the only possibility of hydrogen bonding between the two glutamate side chains,[38,39] which is possible only if at least one of the side chain carboxylic group is protonated (see Fig. 9). Also seen in this schematic diagram is the possibility of the hydrogen bonding between two positively charged hydrophilic arginine residues.

F. Lysine–glutamate (LYS–GLU) pair interaction

Next, we consider the LYS–GLU pair. In Fig. 10, the arrowheads indicate the relative position of the charges of the respective residues; we have the following interesting observations for the interaction of these oppositely charged hydrophilic residue side chains. The strong stabilization at short distances for the configurations, which bring the charges close, has two major contributions: hydrogen bond

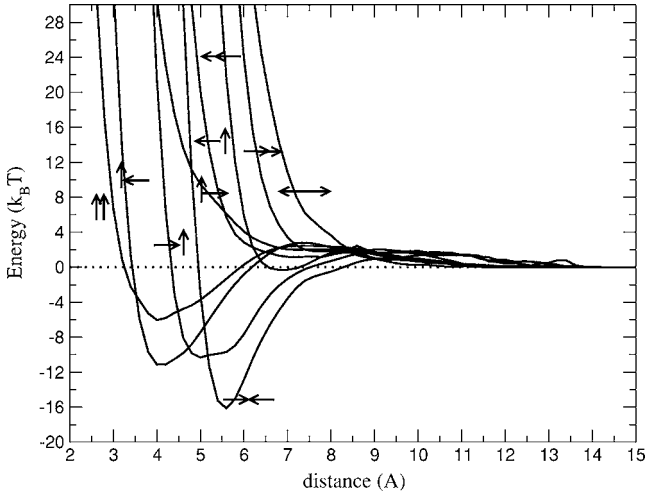

FIG. 10. Energy of interaction of lysine–glutamate pair. Other details are same as in Fig. 6.

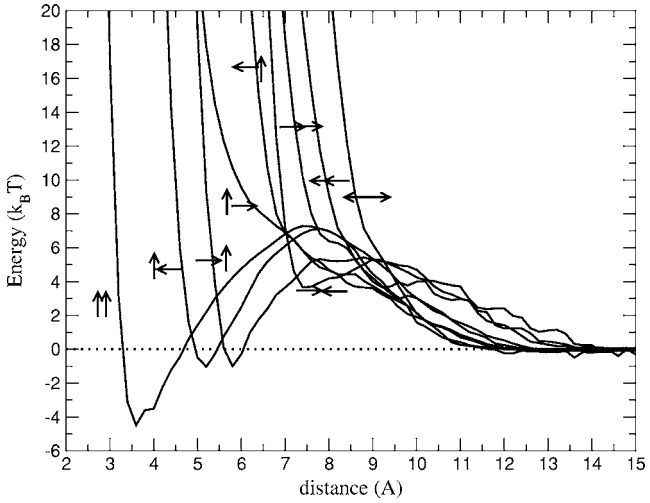

FIG. 11. Energy of interaction of arginine–phenylalanine pair. Other details are same as in Fig. 6.

and the Coulombic interaction (salt bridge). The hydrogen bond stabilization, the GLU–GLU pair, is weaker in comparison with the stabilization seen for the LYS–GLU pair, suggesting that the major contribution to this stabilization is due to the formation of the salt bridge between these two charged side chains. For the arrangements considered, the stabilization can be seen to be essentially determined by the electrostatic interaction between the residues. The configurations, which have a smaller separation between the charges, have a deeper minimum. However, at slightly longer distances, the interaction is slightly energetically destabilizing showing a behavior opposite to the expected trend when only electrostatic interactions are considered. This fact indicates that for larger separations, these charged side chains are successfully screened by the partial charges in the protein environment.

G. Arginine–phenylalanine (ARG–PHE) pair interaction

The results for the hydrophilic–hydrophobic ARG–PHE pair are presented in Fig. 11. This pair shows other additional aspects of pairwise interactions between side chains in the protein environment. All arrangements in which either the distance between the charge of ARG and the π cloud of PHE is too large, or the direction is unfavorable, there is no stabilization and for the arrangements which have the positive charge of the ARG pointing directly at the π-electron cloud of the PHE side chain have a modest energy stabilization. This indicates that the stabilization of some configurations is due to the polarizability of the aromatic charge cloud which is commonly called as the "cation-π" interaction in the literature.[40] For intermediate distances, we also note the presence of a huge barrier separating the small and the large separation arrangements. This feature is typical of solvent-separated ion pairs, but its origin in the present context is not very clear.

We conclude this section by noting that the above discussion clearly indicates that the constructed PMF has all the

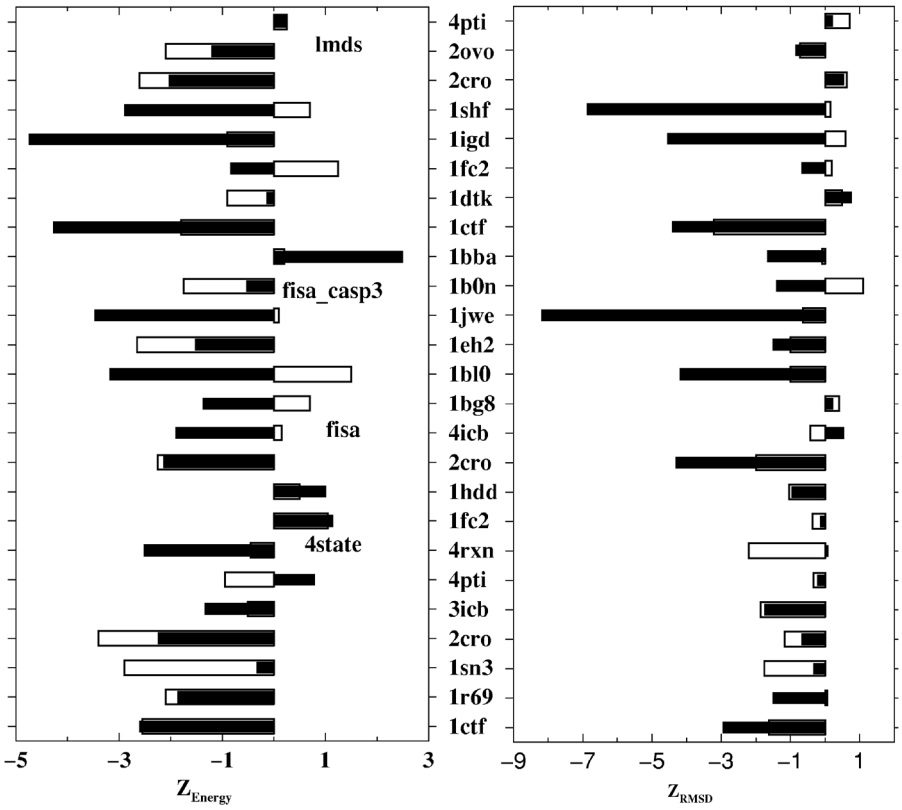

FIG. 12. Comparison of the *Z*-score between the potential of Buchete *et al.* (Ref. 16) (unfilled bars) and this work (filled bars).

salient features of pairwise interactions in proteins and thus is a strong candidate for understanding of the protein folding.

V. EFFICIENCY OF THE PMF IN RECOGNIZING THE NATIVE STRUCTURE FROM AMONG DECOYS

The total energy E of a particular configuration of the protein can be calculated as the sum of all the pairwise interactions between various "structural units" of the minimalistic model presented in the Sec. III as

$$E = \sum_{B_1} \sum_{B_2 > B_1} U_{B_1,B_2}(\mathbf{r}_{B_1} - \mathbf{r}_{B_2}) + \sum_{B} \sum_{\alpha > B} U^{e}_{B,\alpha}(\mathbf{R}_{B\alpha}, \theta_{\alpha}) + \sum_{\alpha} \sum_{\beta > \alpha} U^{e}_{\alpha\beta}(\mathbf{R}_{\alpha\beta}, \theta_{\alpha}, \theta_{\beta}, \phi_{\alpha\beta}), \tag{3}$$

where the B's are the backbone sites, α and β are the side chains of the residues (with α preceding β in the primary sequence by at least by two residues). In the above equation, the first summand U_{B_1,B_2} accounts for the pair interaction of the backbone sites B_1 and B_2 (taken to be spherically symmetric), the second summand $U^{e}_{B,\alpha}$ is the ellipsoidal interaction of a backbone site with the ellipsoidal side chain, and the third summand $U^{e}_{\alpha\beta}$ is for the interaction between two side-chain residues. We emphasize that all the interactions are asymmetric due to the directionality of the protein along its primary sequence and is explicitly taken into account in the above equation.

The ellipsoidal potential of mean force is tested against the orientational potential of mean force proposed by Buchete *et al.*[16] They calculated the orientational potential of mean force directly from PDB (Ref. 41) using the structural data set used by Lee *et al.*[7] On the other hand, we have used a completely different route to calculate the orientation-dependent potential of mean force. The protein data set used in our calculation is the same as that used by Liwo *et al.*[29] and Buchete *et al.*[16] However, the potential of mean force proposed by Buchete *et al.* does not contain backbone–backbone and backbone–side-chain interactions. They improved their potential subsequently in a later paper[32] by including the above interactions in their original model.

Z-score is one of the measures which is often used to test a potential. *Z*-score for a particular quantity is defined as

$$Z_x = \frac{x - \langle x \rangle}{\sigma_x}, \tag{4}$$

where $\langle x \rangle$ is the mean value and σ_x is the standard deviation for a certain distribution of x. Z_E is defined for the energy of the native structure (from the crystal structure or the NMR structure). Z_{RMSD} is defined with respect to the RMSD (root-mean-square deviation) of the lowest energy structure obtained from PMF calculation.

We have plotted the *Z*-score values for energy and RMSD in Fig. 12. Results have been compared to the recent work of BST.[16] Figure 12 shows the comparison of the *Z*-score values for several proteins and their corresponding decoy states. The filled bars show our work whereas the open bars show the results from BST.[16] We can see that the *Z*-score values for both energy and RMSD are quite comparable.

Figures 13–15 show the *Z*-score values (both energy and RMSD) for three different decoy sets—hg_structal, ig_structal_hires, and ig_structal—respectively. Figure 13 (for the hg_structal set) shows only six positive *Z*-score values for energy and two small positive Z_{RMSD} values. So in this case

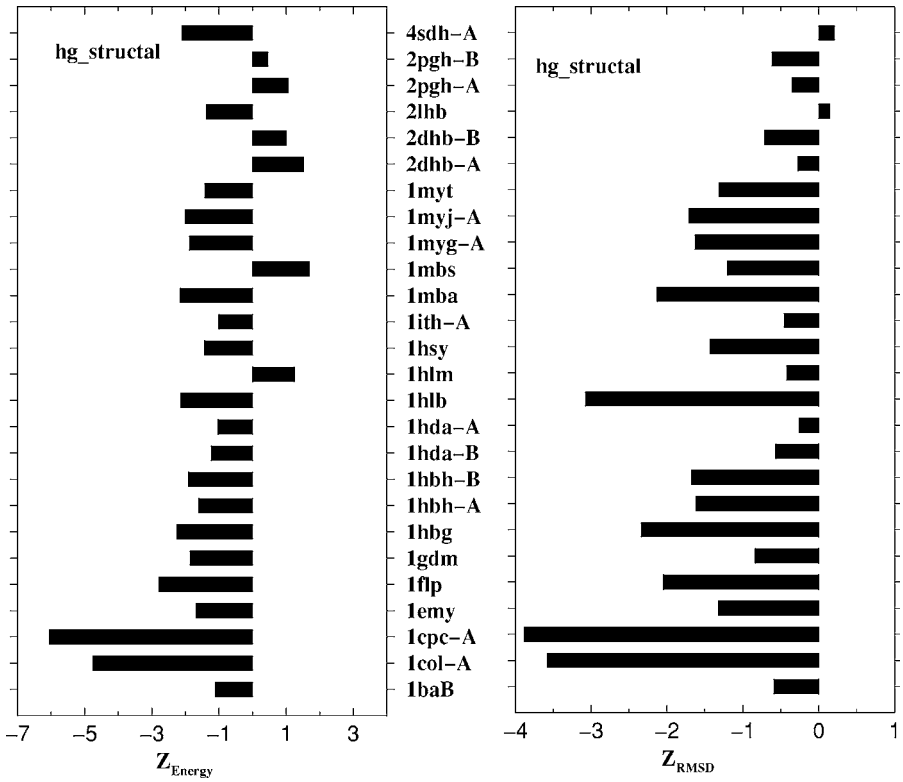

FIG. 13. *Z*-scores for energy and RMSD are shown for hg_structal set of decoys.

the results are better than the previous decoy sets shown in Fig. 12. Figure 14 shows the *Z*-score values for ig_structal_hires (upper half), where all the Z_E values are negative, and only one value of Z_{RMSD} is positive. Figures 14 (lower half) and 15 show the *Z*-score results for ig_structal decoy sets. Except three, all the Z_E values are negative, although Z_{RMSD} shows some positive values.

From the energy values of 136 proteins and their corresponding decoys, it is seen that in 54 proteins, native structure is the lowest-energy state. Considering the potential is derived from the native structures and the model considered being a coarse-grained one, the orientation-dependent potential can be considered as an effective tool for structure prediction.

VI. CONCLUSION

Let us first summarize the main results of this work. We have constructed a solvent-implicit, coarse-grained, minimal-

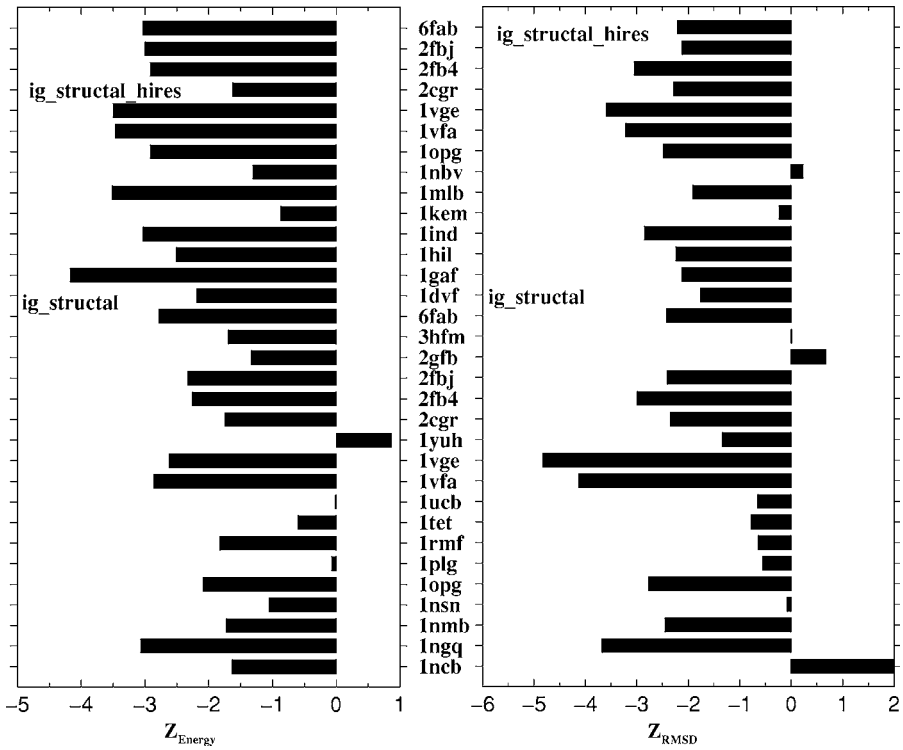

FIG. 14. *Z*-scores for energy and RMSD are shown for ig_structal_hires set of decoys.

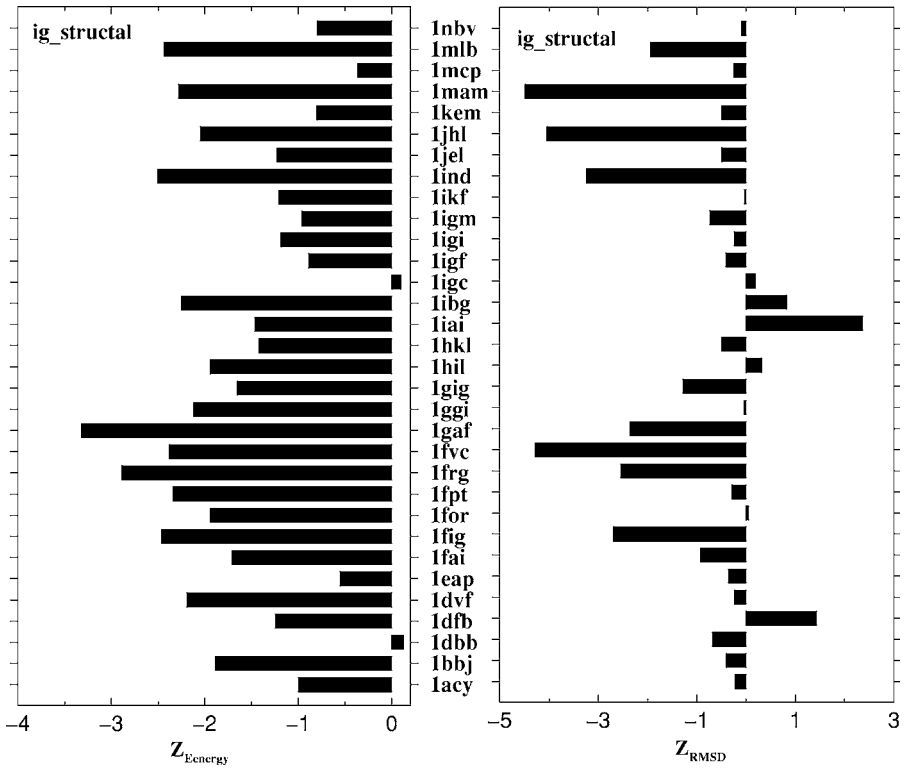

FIG. 15. *Z*-scores for energy and RMSD are shown for ig_structal set of decoys.

istic model of interaction potential between the different amino acid groups in a protein. The side chains of the amino acids have been characterized as ellipsoids. The potential is then four dimensional, distance and the three Euler angles being the four coordinates. The calculated potential is statistical and has been calculated from the native structure by summing over all the heavy atoms in a given amino acid, by following the procedure pioneered by Sippl and Feytsman.

The calculated potential shows several expected features. The interaction between any two hydrophobic groups is attractive at short separations. The magnitude of the potential at the minimum seems to correlate with the hydrophobicity of the amino acid. For example, the minimum is deep for phenylalanine. Similarly, the interaction is mostly repulsive between hydrophilic groups.

There are also several rather unexpected (although understandable) features of the potential. For some orientations, there is a repulsive region in the interaction potential at distances larger than the minimum, giving rise to a shape akin to solvent-separated ion-pair (SSIP) potential. This is most clear for interaction between arginine and phenylalanine, for parallel orientation. There are several detail features of the statistical potential, which are rather satisfying, for example, the deep potential minimum between lysine and glutamate for face-to-face configuration. We also find evidence of the quadrupolar interaction between phenyl rings and charge-ring interaction, which are impressive given that the constructed potential is a statistical potential.

The new orientation-dependent potential successfully recognizes the native state for several proteins, as evident from the *Z*-score. The potential is obtained from atomistic potential of mean force which relies on the chemical environment of an atom rather that to which amino acid it belongs to. We believe that this ellipsoid potential of mean force will prove to be successful in finding out the native state of a protein with unknown sequence. The work in this direction is under progress. Note, in addition, that the smoothness of the calculated PMF may allow Brownian dynamics simulation to study aspects of the dynamics of protein folding.

Solvent-implicit minimalistic models of course suffer from the inherent drawback due to the absence of water whose detailed role in any given protein folding has not been elucidated yet. Given this lacuna, one desirable feature of any minimalistic model is that it should be sufficiently simple that it is easy to implement in any calculation and simulation. This requires coarse graining and this approach can only be successful if specific details are not too important.

[1] C. B. Anfinsen, Science **181**, 223 (1973); M. Sela, F. H. White, and C. B. Anfinsen, *ibid.* **125**, 691 (1957).
[2] Y. Duan and P. Kollman, Science **282**, 740 (1998).
[3] C. D. Snow, H. Ngyen, V. S. Pande, and M. Gruebele, Nature (London) **420**, 102 (2002).
[4] See the articles in (a) *Advances in Protein Chemistry*, edited by F. M. Richards, D. S. Eisenberg, J. Kuriyan, and V. Daggert (Academic, San Diego, 2003), Vol. 66. (b) *Advances in Protein Chemistry; Protein Folding Mechanisms*, edited by C. R. Matthews (Academic, San Diego, 2000), Vol. 53.
[5] P. Thomas and K. Dill, J. Mol. Biol. **257**, 457 (1996).
[6] M. Levitt and A. Warshel, Nature (London) **253**, 694 (1975).
[7] J. Lee, A. Liwo, and H. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **96**, 2025 (1999).
[8] A. Godzik and J. Skolnick, Proc. Natl. Acad. Sci. U.S.A. **89**, 12098 (1992).
[9] A. Sali, E. Shakhnovich, and M. Karplus, Nature (London) **369**, 248 (1994).
[10] R. Zwanzig, Proc. Natl. Acad. Sci. U.S.A. **92**, 9801 (1995).
[11] C. M. Dobson, Nature (London) **246**, 884 (2003).
[12] R. Zhou and B. J. Berne, Proc. Natl. Acad. Sci. U.S.A. **99**, 12777 (2002).
[13] M. J. Sippl, J. Mol. Biol. **213**, 859 (1990).

[14] G. Srinivas and B. Bagchi, J. Chem. Phys. **118**, 4733 (2003).
[15] A. Mukherjee and B. Bagchi, J. Chem. Phys. **116**, 6220 (2001).
[16] N.-V. Buchete, J. E. Straub, and D. Thirumalai, J. Chem. Phys. **118**, 7658 (2003).
[17] R. Samudrala and M. Levitt, Protein Sci. **9**, 1399 (2000).
[18] S. Tanaka and H. Scheraga, Macromolecules **9**, 945 (1976).
[19] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985).
[20] M. Hendlich, P. Lackner, S. Weitckus, H. Floechner, R. Froschauer, K. Gottsbacfher, G. Casari, and M. J. Sippl, J. Mol. Biol. **216**, 167 (1990).
[21] D. T. Jones, W. R. Taylor, and J. M. Thornton, Nature (London) **358**, 86 (1992).
[22] S. H. Bryant and L. M. Amzel, Int. J. Pept. Protein Res. **29**, 46 (1987).
[23] K. Nishikawa and Y. Matsuo, Protein Eng. **6**, 811 (1993).
[24] J. P. Kocher, M. J. Rooman, and S. J. Wodak, J. Mol. Biol. **235**, 1598 (1994).
[25] B. A. Reva, A. V. Finkelstein, M. F. Sanner, and A. J. Olson, Protein Eng. **10**, 865 (1997).
[26] D. W. Gatchell, S. Dennis, and S. Vajda, Proteins **41**, 518 (2000).
[27] H. Zhou and Y. Zhou, Protein Sci. **11**, 2714 (2002).
[28] L. Jiang, Y. Gao, F. Mao, Z. Liu, and L. Lai, Proteins **46**, 190 (2002).
[29] A. Liwo, S. Oldziej, M. Pincus, R. Wawak, S. Rackovsky, and H. Scheraga, J. Comput. Chem. **18**, 849 (1997).
[30] F. Melo and E. Feytsman, J. Mol. Biol. **267**, 207 (1997).
[31] J. G. Gay and B. J. Berne, J. Chem. Phys. **74**, 3316 (1981).
[32] N.-V. Buchete, J. E. Straub, and D. Thirumalai, Protein Sci. **13**, 862 (2004).
[33] M. J. Sippl, J. Mol. Biol. **224**, 725 (1992).
[34] V. A. Eyrich and R. A. Friesner, Adv. Chem. Phys. **120**, 22 (2002).
[35] F. Melo and E. Feytsman, J. Mol. Biol. **277**, 1141 (1998).
[36] M. J. Sippl, M. Ortner, M. Jaritz, P. Lackner, and H. Flöckner, Folding Des. **1**, 289 (1996).
[37] S. Scheiner, T. Kar, and J. Pattanayak, J. Am. Chem. Soc. **124**, 13257 (2002).
[38] I. Y. Torshin, R. W. Harrison, and I. T. Weber, Protein Eng. **16**, 16 (2003).
[39] S. O. Smith, C. S. Smith, and B. J. Bormann, Nat. Struct. Biol. **3**, 252 (1996).
[40] T. Steiner and G. Koellner, J. Mol. Biol. **305**, 535 (2001).
[41] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, Nucleic Acids Res. **28**, 235 (2000).