Available online at www.sciencedirect.com
SCIENCE DIRECT®
Polymer 45 (2004) 597–608
polymer
www.elsevier.com/locate/polymer

Orientation-dependent coarse-grained potentials derived by statistical analysis of molecular structural databases

N.-V. Buchete[a,1], J.E. Straub[a,*], D. Thirumalai[b]

[a]*Department of Chemistry, Boston University, Boston, MA 02215, USA*
[b]*Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742, USA*

Received 15 July 2003; received in revised form 29 August 2003; accepted 29 August 2003

Abstract

We present results obtained for anisotropic potentials for protein simulations extracted from the continually growing databases of protein structures. This work is based on the assumption that the detailed information on molecular conformations can be used to derive statistical (a.k.a. 'knowledge-based') potentials that can describe on a coarse-grained level the side chain–side chain interactions in peptides and proteins. The complexity of inter-residue interactions is reflected in a high degree of orientational anisotropy for the twenty amino acids. By including in this coarse-grained interaction model the possibility of quantifying the backbone–backbone and backbone–side chain interactions, important improvements are obtained in characterizing the native protein states. Results obtained from tests that involve the identification of native-like conformations from large sets of decoy structures are presented. The method for deriving orientation-dependent statistical potentials is also applied to obtain water–water interactions. Monte Carlo simulations using the new coarse-grained water model show that the locations of the minima and maxima of the oxygen–oxygen radial distribution function correspond well with experimental measurements.

Keywords: Proteomics; Protein packing; Coarse-grained potentials

1. Introduction

Large-scale protein folding and protein structure prediction is essential in research fields like proteomics and structure-based molecular biology. Despite substantial advances in both all-atom molecular simulation methods and computational technologies, it is infeasible to perform three-dimensional, off-lattice protein simulations on thousands of proteins at a time to describe in detail processes like protein–protein interactions or even protein folding pathways. Many drug companies and molecular biology laboratories use all atom structure-based methods for drug design, but this approach is too computationally intensive for large scale applications.

To address these shortcomings, there is an on-going effort to develop a class of interaction potentials between amino acids that can describe in a simplified, yet accurate, manner the essential intra- and inter-protein interactions that dictate the thermodynamic and kinetic biochemical properties. Construction of such models requires the determination of interaction potentials between amino acid residues. More than twenty years ago, Tanaka and Scheraga [1] proposed that the frequencies of amino acid pairing can be used to determine potential interaction parameters. Since that pioneering work, the wealth of structural data on a number of proteins in the Protein Data Bank (PDB) [2] has been a source for obtaining interaction potentials. [3–5] With the exception of a few studies [6], most 'knowledge-based' potentials have been obtained solely in terms of residue–residue contacts.

Sippl [6] introduced an explicit distance dependence in the database-derived mean force potentials using the Boltzmann formula. This method, known as the 'Boltzmann device', assumes that the known protein structures from the PDB correspond to classical equilibrium states. Therefore, the distribution of the distance r between two side chains, should correspond to the equilibrium Boltzmann distribution

* Corresponding author. Tel.: +1-617-353-6816; fax: +1-617-353-6466.
E-mail address: straub@bu.edu (J.E. Straub).
[1] Current address: Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892, USA.

0032-3861/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.polymer.2003.10.093

corresponding to a particular potential of mean force. Given the distribution, the potential may be derived. Sippl [7] suggested that other structural parameters, such as dihedral angles, can be treated in a similar manner.

In this paper, we present new methods for extracting orientation-dependent statistical potentials for coarse-grained representations of groups of atoms such as side-chains, atoms involved in the peptide link or water. The method is used to develop a novel set of coarse-grained distance- and orientation-dependent residue–residue statistical potentials [8] as well as a statistical potential for water molecules. We present results obtained by including an extra anisotropic backbone interaction center located at the peptide bond and by studying their performance in discriminating the native protein structures in tests that employ multiple protein decoy sets [9]. This new approach can be extrapolated to build a statistical description of water–water interactions that could be useful in coarse-grained simulations, as suggested by results obtained from Monte Carlo simulations.

2. Statistical potentials for proteins

2.1. The underlying models of the peptide chains

In Fig. 1 are presented the examples of models of different levels of description that are commonly used in the coarse-grained representations of polypeptides. All the models aim to employ the simplest description that can provide a high degree of accuracy and realism of the predicted properties of the peptide systems. The simplest

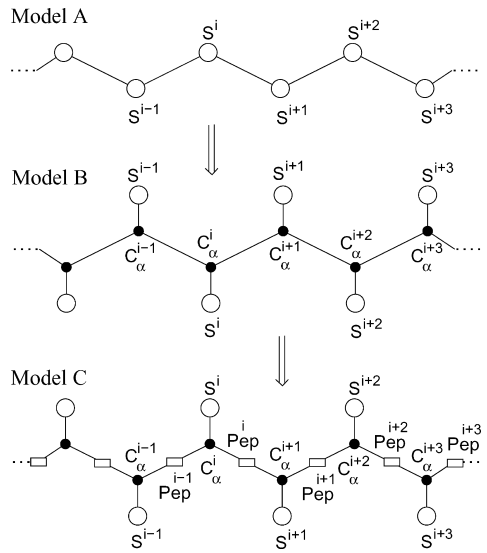

Fig. 1. Levels of commonly employed coarse-grained re presentations of peptides. While the main goal is the simplest description (as in A), an accurate description of the backbone needs to include the C_α positions (model B) and even specific sites for backbone–backbone and backbone–side chains interactions (model C).

models used for residue–residue interactions consider the peptides as simple chains of interacting beads (as in Fig. 1, model A) [10]. In order to account for the various sizes and specific packing features of the 20 different types of amino acids, more detailed models must be employed in estimations of the relative magnitudes of residue–residue interactions. In such models (Fig. 1, model B) [1,11,12] the backbone is described using one type of non-interacting backbone site located at the positions of the C_α atoms, with a second type of interaction centers S^i that are corresponding to each side chain (SC). The S^i interaction centers are typically located at the center of mass of the heavy atoms in each side chain, with the exception of Gly, where it coincides with the position of the C_α atom. The backbone sites C_α^i are used to describe the backbone structure, but only the S^i interaction centers are considered to interact with each other. Models similar to B have been successfully used to obtain contact-based side chain–side chain (SC–SC) interaction potentials, distance-dependent potentials [6,7] and, more recently, distance- and orientation-dependent potentials [8]. While useful, these models do not allow for the explicit treatment of side chain–backbone and backbone–backbone interactions. Recent estimates [13] suggest that the number of backbone–backbone contacts can range from 12% to as much as 35% depending on the protein class (e.g. alpha, beta, mixed alpha–beta [14]) and of the topological interaction level along the sequence that is considered (e.g. $|i-j| \geq 3$, $|i-j| \geq 4$, etc.). The importance of including the backbone interactions is also supported by the results of previous statistical derivations of backbone potentials that used virtual bond and torsion angles [15] and secondary structure information [16]. Therefore, we employ a more complex model (Fig. 1, model C) that includes an additional interaction center located on the backbone [11] at the geometric center of each peptide bond (Pep^i). In this description, we assume that the local conformation of a certain residue i is sufficiently well described by the corresponding C_α^i, S^i and Pep^i interaction centers. Our results from tests on decoy sets suggest that model C offers important improvements over model B in the ability to recognize native-like states of proteins.

2.2. Local reference frames of side chains

Previous studies have demonstrated the importance of orientational dependence of side chain–side chain interactions [8,17–20]. To extract quantitative parameters for the orientational dependence of coarse-grained potentials from PDB structures [2] we define local reference frames (LRFs) for each amino acid by using the approach described in Ref. [8] and summarized next.

The LRFs were constructed by considering at least three non-collinear points (P_1, P_2 and P_3) that uniquely define the orientation of the LRF, and a fourth point (usually denoted by S^i for the ith side chain) that specifies the location of the LRF origin. In the coarse-grained representation, the S^i

points can be considered to be the 'interaction centers' since all the relative side chain–side chain distances and orientations are measured with respect to them. Let $\vec{r}_{p_1}$, $\vec{r}_{p_2}$, $\vec{r}_{p_3}$, and $\vec{r}s_i$ be the position vectors of the points P_1, P_2, P_3, and S^i, respectively. The $\hat{O}_z$ axis vector can be defined as

$$\hat{O}_z = \frac{\vec{r}_{p_2} - \vec{r}_{p_1}}{|\vec{r}_{p_2} - \vec{r}_{p_1}|}. \tag{1}$$

A second direction $\hat{O}_y^*$, pointing toward the Oy axis can be similarly constructed as

$$\hat{O}_y^* = \frac{\vec{r}_{p_3} - \vec{r}_{p_2}}{|\vec{r}_{p_3} - \vec{r}_{p_2}|}. \tag{2}$$

Finally, the $\hat{O}_x$ and $\hat{O}_y$ axis vectors are defined in terms of the cross products

$$\hat{O}_x = \hat{O}_y^* \otimes \hat{O}_z \text{ and } \hat{O}_y = \hat{O}_z \otimes \hat{O}_x. \tag{3}$$

For side chains, the positions of the three reference points P_1, P_2 and P_3 are identified with the positions of the C_α, C_β and C_γ atoms [8]. The position of the interaction centers S^i are identified with the geometric center (GC) of the heavy atoms in the side chain. Exceptions to this rule are made for the special cases of Gly, Ala, Cys, Ser, Ile and Val as described in Ref. [8]. These definitions have the advantage that, while being side-chain dependent, the positive O_z axis is always oriented away from the local backbone while the positive O_y axis points toward more 'remote' C_γ atoms in the SC. For small side chains, O_y will point towards the next SC on the backbone sequence.

Significant improvements are obtained by considering a virtual backbone interaction center in the middle of the peptide bond (Pep). We were motivated to include this additional, twenty-first interaction center by the observation that folded structures are stabilized by a substantial number of side chain–backbone contacts. For Pep, the positions of the three reference points P_1, P_2 and P_3 are identified with the positions of the carbonyl C atom, its O atom and the peptide bond N atom. The interaction center S^i for Pep is placed in the middle of its C–N peptide link. These definitions of the LRFs permit the investigation of relative coordination probabilities (e.g. for hydrogen bonding) as well as of hydropathic effects in side chain packing.

2.3. From orientational probabilities to statistical potentials: the Boltzmann device

Once the local reference systems for special groups of atoms (e.g. the heavy atoms in side chains, or the C, O and N for the peptide link Pep) are defined, the statistics collected from a database of nonhomologous proteins can be used to estimate the pair distributions for each specific type of site–site interaction. We used a standard, reproducible approach, by employing the set of nonhomologous proteins that was used by Scheraga et al. [18–20] for similar purposes. A larger training set of protein structures could be used if higher accuracy is necessary. The pair distributions are further normalized by considering the corresponding volume element and the total number of observations for building orientational probabilities $P^{ij}(r, \phi, \theta)$ for each type of interaction. In this notation, $P^{ij}(r, \phi, \theta)$ represents the probability to observe a side chain of type j in the spherical volume element corresponding to the set of coordinates $r, \phi, and \theta$ in the local frame LRF_i of side chain i.

The construction of the orientational statistical potentials is done using the 'Boltzmann device' [6,8]. By using the basic assumption that the known protein structures from protein databases (such as PDB) correspond to classical equilibrium states, we can define

$$U_{\mathrm{DO}}^{ij}(r, \phi, \theta) = -kT \ln\left[\frac{P^{ij}(r, \phi, \theta)}{P_{\mathrm{ref}}(r, \phi, \theta)}\right] \tag{4}$$

as the distance- and orientation-dependent statistical potential for the ij pair. Here $P_{\mathrm{ref}}(r, \phi, \theta)$ is the reference probability density which can be obtained from the interactions between all the residue types.

The total potential for the residue pair ij is

$$U_{\mathrm{DO}}^{ij}(r_{ij}, \phi_{ij}, \theta_{ij}, \phi_{ji}, \theta_{ji}) = U_{\mathrm{DO}}^{ij}(r_{ij}, \phi_{ij}, \theta_{ij}) + U_{\mathrm{DO}}^{ij}(r_{ji}, \phi_{ji}, \theta_{ji}) \tag{5}$$

where pairwise additivity is assumed. The UDO notation is used for statistical potentials that are both distance- and orientation-dependent. Eq. (5) is based on the major assumption of pairwise additivity of the inter-residue potentials in proteins. For Boltzmann equilibrium, this separability is consistent with the probabilistic relation between the individual probabilities $P^{ij}(r_{ij}, \phi_{ij}, \theta_{ij})$ and $P^{ji}(r_{ji}, \phi_{ji}, \theta_{ji})$ estimated from the observed frequencies of interaction, and the total interaction probability $P^{ij}(r_{ij}, \phi_{ij}, \theta_{ij}, \phi_{ji}, \theta_{ji})$ [8]. Also, in this implementation, the dependence of the U_{DO}^{ij} potentials on the torsional angle around r_{ij} (see Fig. 3 in Ref. [8]) is averaged out. The results suggest that the effect on the accuracy of the UDO potentials of the assumption that the interaction terms can be truncated as in Eq. (5) is not very large.

An important issue that appears when using probability density functions with the Boltzmann device is 'the problem of small data sets'. As noted by Sippl [6], dividing the SC–SC pair frequencies by both side chain type and distance intervals results in situations when the available data is too small for conventional statistical procedures. This problem was solved by Sippl by proposing a 'sparse data correction' formula that builds the correct probability densities as linear combinations between the measured data and the reference. The total probability densities are obtained by averaging over all twenty SC types.

For the general, orientation-dependent probability

densities the sparse data correction can be written as

$$P^{ij}_{\text{corr}}(r,\phi,\theta) = \frac{1}{1+m'\sigma}P_{\text{ref}}(r,\phi,\theta) + \frac{m'\sigma}{1+m'\sigma} \times P^{ij}(r,\phi,\theta) \quad (6)$$

where P^{ij} are the actual probability densities obtained from the database for the ij pair of side chains, and P^{ij}_{corr} are the corrected probabilities. P_{ref} is the reference probability density obtained by averaging over all the residue types. A modification introduced by the orientational dependence in our case is that the number of measurements m becomes $m' = m/\sin(\theta_k)$, as k equiangular intervals are used for the θ angle. This is necessary for accounting for the azimuthal dependence of volume elements in spherical coordinates. The σ parameter, which is a constant, controls how many actual measurements m' must be observed so that both the actual probabilities and the reference would have equal weights. As in other studies, we used $\sigma = 1/50$ [6,21,22].

2.4. The importance of the relative side chain–side chain orientations: results from decoy tests using model B

To asses the importance of the relative side chain–side chain orientations using only the simple model B (see Fig. 1), we performed tests using the distance- and orientation-dependent statistical potentials (UDO) as scoring functions. Their performance in correctly recognizing the native structure was assesed using a standard database of decoys developed by Samudrala and Levitt [9]. The results are shown in terms of the values of the energy and root-mean-square distance Z-scores (Z_{E} and Z_{RMSD}) that are defined next. The root-mean-square distance (RMSD) is calculated with respect to the C_α atoms. The general definition of the Z-score of a statistical quantity x is

$$Z = \frac{x - \bar{x}}{\sigma} \quad (7)$$

where σ is the standard deviation and $\bar{x}$ is the mean of the distribution of x values.

For comparing the performance of various interaction potentials on sets of decoy structures, we have computed both the energy and root mean square distance Z-scores (Z_{E} and Z_{RMSD}) for the distribution of the total energies for each protein decoy set. [8] The energy Z-score (Z_{E}) is a relative measure of the value of the energy of the native state with respect to the distribution of all decoy energies. The RMSD Z-score (Z_{RMSD}) is a relative measure of the value of the C_α root-mean-square distance (RMSD) from the native state of the decoy with the lowest energy with respect to the distribution of RMSDs of the other decoys. Both Z_{E} and Z_{RMSD} are important [9]. The data shown in Fig. 2 illustrates the method of calculating energy Z-scores for the set of 500 decoys of the 2cro protein from the 'fisa' family [9,23]. The two distributions correspond to the distance-dependent statistical potential (red histogram, left) for a 20×20 interaction scheme, and to the present smoothed distance- and orientation-dependent 21×21 potential (blue, right) that includes backbone interactions. The values for the corresponding energies of the native 2cro state and the mean values are also shown for illustrating the definitions of the energy Z-scores (Z_{E}) used in our analysis. For ideal potentials it is expected that the structure corresponding to the native state would have the most negative energy Z-score (Z_{E}). We also expect that, a good potential scoring function, one consistent with a single 'funnel-like' energy landscape, should also assign a very negative C_α root mean square distance Z-score (Z_{RMSD}) to the decoy structure that has the lowest energy.

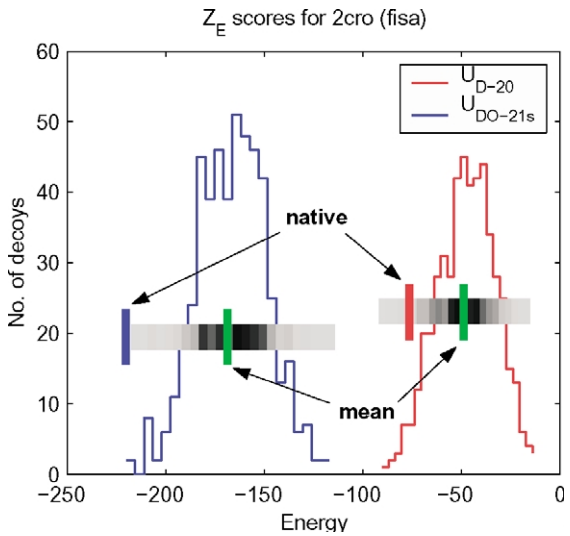

Fig. 2. The distributions of energy values for the set of 500 decoys of the *2cro* protein from the "fisa' family [9,23]. The two distributions correspond to the distance-dependent statistical potential (red histogram, left) for a 20×20 interaction scheme, and to the present smoothed distance- and orientation-dependent 21×21 potential (blue, right) that includes backbone interactions. The values for the corresponding energies of the native 2cro state and the mean values are also shown to illustrate the definitions of the energy Z-scores (ZE) used in our analysis. The Z-scores are a measure of how far the native energy values are displaced from the mean of the corresponding distribution as compared to the standard deviation.

In Fig. 3 are shown energy (Z_{E}) and C_α RMSD Z-scores (Z_{RMSD}) calculated for the multiple decoy sets 'lmds', 'fisa_casp3', fisa and '4state' [9].

Z-scores calculated for the distance-only dependent statistical potentials U_{D} (dark bars) and for the new distance- and orientation-dependent U_{DO} potential (white bars) values are compared. In all the cases presented here regarding Z-scores, more negative values are better, and the cases in which U_{DO} leads to better results than U_{D} are emphasized using arrows. Based on considerations regarding the limited experimental resolutions of the nonhomologous structures that are analyzed in the training set [8], for the U_{D} potentials used here we employed 20 radial distance bins of width $L = 1.2$ Å for distances starting at 2 Å. For the U_{DO} potentials the analysis was performed using 12 angular bins for θ and 24 bins for ϕ. It is observed that in a large

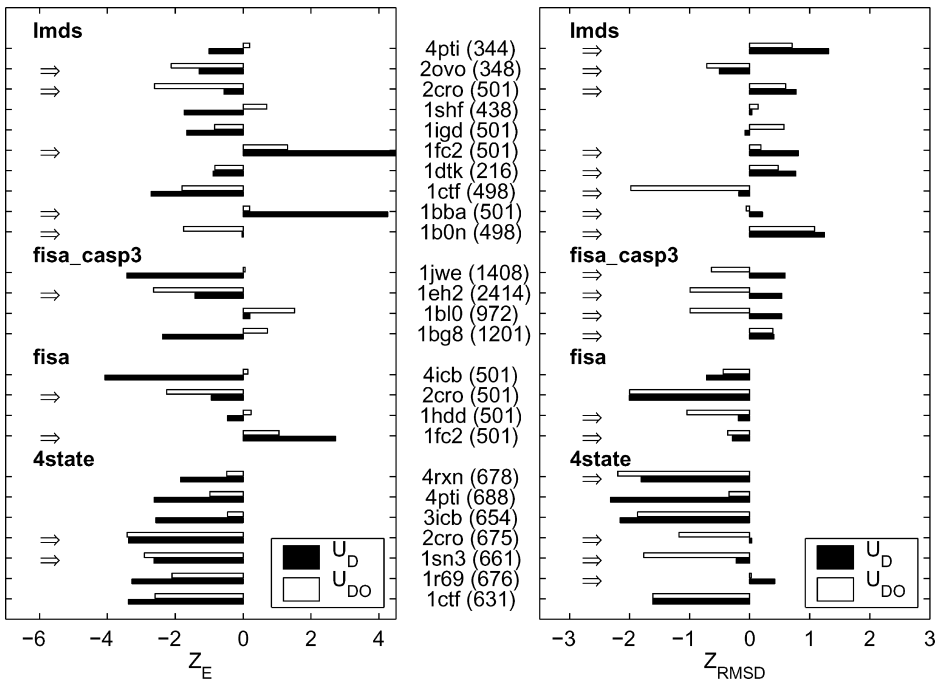

Fig. 3. Energy (Z_E) and C_α RMSD Z-scores (Z_{RMSD}) calculated for the multiple decoy sets 'lmds', 'fisa casp3', 'fisa' and '4state' [9]. The numbers in brackets represent the number of decoys in each set, including the native structure. Z-scores calculated using statistical potentials dependent only on distance U_D (dark bars) and distance- and orientation-dependent potentials (U_{DO}, white) are compared. More negative Z-scores are better, and the cases in which U_{DO} leads to better results than U_D are emphasized by the arrows on the left. When both Z_E and Z_{RMSD} Z-scores are considered, the inclusion of orientational information improves the performance in a majority of cases.

number of cases the inclusion of orientational information improves the performance of both Z_E and Z_{RMSD} scores. An especially interesting case is the one of the 'lmds' set in which all-atom distance-dependent scores were shown to perform poorly [9,24]. In this case, when both the Z_E and Z_{RMSD} Z-scores are considered, the new distance- and orientation-dependent potentials U_{DO} performed much better than the distance-only dependent UD in a majority of cases. A more extensive analysis of these tests is presented elsewhere [8].

Our results show that the information encoded in the relative side chain–side chain orientations in native-like structures of proteins could be as important (and in many cases more important) than the information that can be extracted from relative distances alone. Even though in many structures the essential features of inter-residue interactions can be extracted by analyzing only the residue–residue distances, the complementary information contained by relative orientations is also important in a significant number of cases.

2.5. Including backbone interactions in a smooth, continuous version of the coarse-grained potentials

It is important to have a more realistic, spatially continuous description of the statistical potentials that also includes the backbone interactions explicitly (see model C in Fig. 1). However the amount of statistical data that is available in the training set of nonhomologous structures is limited. To overcome this limitation, besides the implementation of the 'sparse data correction' method (see Eq. (6)), we have also reduced the number of radial interaction ranges to only three. As shown by other studies [25], the short-range distance-dependence of the statistical potentials in globular proteins reflects specific differences between hydrophobic and hydrophilic side chains. We observed

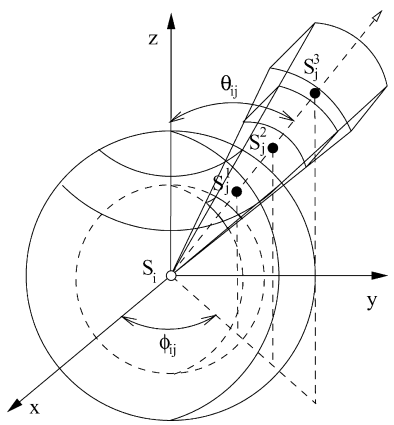

Fig. 4. The definitions of local reference frames for specific groups of atoms (e.g. heavy atoms in side chains and backbone atoms involved in the peptide link) permit the collection of data on relative distances and orientations between those groups. To obtain sufficient data to permit a smooth, continuous description of the relative orientational dependence, we employed only three distance ranges, corresponding to short-, medium- and long-range interactions. As shown in this figure, the volume elements at the given orientation have different size; the appropriate normalization must be performed for building orientation-dependent probabilities.

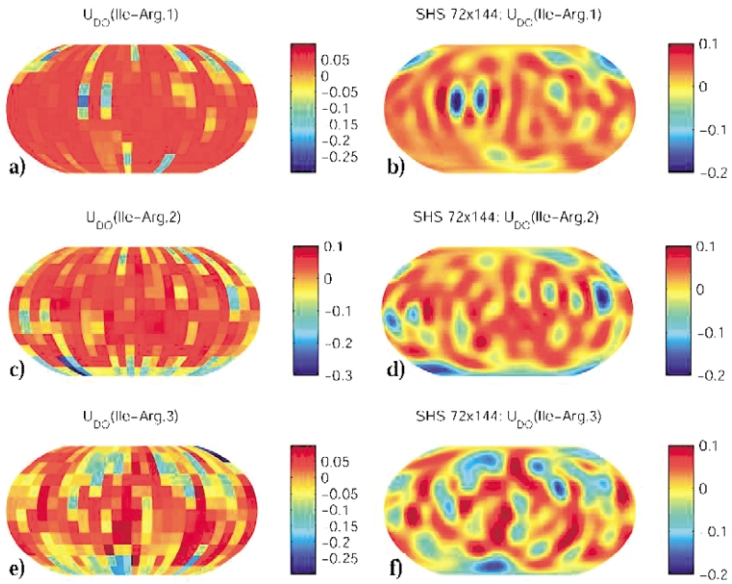

Fig. 5. The smooth Ile-Arg orientational potentials represented for short-range (top), middle-range (middle) and long-range (bottom) interactions. The potential values, calculated originally for a 12×24 equiangular grid, are shown on the right (a, c and e), and the corresponding smooth potentials computed for a 72×144 grid using spherical harmonic synthesis are shown on the left (b, d and f). Blue regions correspond to attractive (i.e. negative) potentials, while red regions are positive, thus less likely to correspond to interaction loci.

similar qualitative differences (Figs. 4 and 5 in Ref. [8]) in the orientational probability maps extracted from protein structures. Due to the orientation dependence, the coarse graining of the distance dependence does not prevent the observation of hydropathic effects in the short-range U_{DO} potentials.

In Fig. 4 are illustrated the three levels of volume elements in spherical coordinates that were employed for collecting data on side chain–side chain, side chain–backbone and backbone–backbone interactions. When computing the interaction probabilities, the raw histogram data must be normalized not only by dividing by the corresponding volume of each shell, but also by $\sin(\theta)$ in order to eliminate the 'pole effects'.

As shown in Fig. 4, in our approach we have considered three specific interaction ranges, corresponding to short-, medium- and long-range interactions (i.e. $2.0 \rightarrow 5.6$ Å, $5.6 \rightarrow 9.2$ Å, and $9.2 \rightarrow 12.8$ Å). The new orientation-dependent potentials present sufficient continuity properties to allow for spherical harmonic analysis (SHA) [13,26]. The numerical spherical harmonic analysis of the new 21×21 (i.e. backbone dependent) potentials $U_{\mathrm{DO\text{-}21}}$ was performed using the technique developed by Adams and Swarztrauber and implemented in the FORTRAN package Spherepack 3.0 [26,27]. We used $2(N-1)$ grid points for ϕ, where $N = 13$ is the number of grid points corresponding to sampling the data along the θ angle [27]. These sampling points are placed on the following equiangular grid

$$\theta_i = i\Delta\theta - \pi/2, \quad i = 0, 1, \ldots, N-1, \quad \Delta\theta = \frac{\pi}{N-1} \tag{8}$$

$$\phi_i = j\Delta\phi, \quad j = 0, 1, \ldots, 2N-1, \quad \Delta\phi = \Delta\theta.$$

Since the angular dependent potential functions are sufficiently smooth, we performed their spherical harmonic analysis and find the corresponding coefficients

$$a_{mn} = \alpha_{mn} \int_0^{2\pi} \int_{-\pi/2}^{\pi/2} U(\theta, \phi)\, P_n^m(\cos\theta)\cos(m\phi)\cos\theta \, \mathrm{d}\phi \, \mathrm{d}\theta \tag{9}$$

$$b_{mn} = \alpha_{mn} \int_0^{2\pi} \int_{-\pi/2}^{\pi/2} U(\theta, \phi)\, P_n^m(\cos\theta)\sin(m\phi)\cos\theta \, \mathrm{d}\phi \, \mathrm{d}\theta \tag{10}$$

where

$$\alpha_{nm} = \frac{2n+1}{2\pi} \frac{(n-m)!}{(n+m)!} \tag{11}$$

and P_n^m are the associated Legendre functions [26,28]. Alternatively, if the coefficients a_n^m and b_n^m are known, the corresponding smooth potential function $U(\theta, \phi)$ can be reconstructed using the spherical harmonics synthesis (SHS) formula

$$U(\theta, \phi) = \sum_{n=0}^{N} \sum_{m=0}^{n\prime} P_n^m(\cos\theta)[a_{mn}\cos(m\phi) + b_{mn}\sin(m\phi)]. \tag{12}$$

The prime notation [26] on the sum indicates that the first term corresponding to $m = 0$ must be multiplied by 0.5. We use the notation $U_{\mathrm{DO\text{-}21s}}$ for the 21×21 distance- and orientation-dependent potentials that were reconstructed by using the SHA/SHS procedure.

To illustrate the process of calculating the potential values by SHS, we show in Fig. 5 the reconstructed Ile-Arg

orientational potential using 12 × 24 equiangular bins (left) and a 72 × 144 grid (right), for short-range (top), middle-range (middle) and long-range (bottom) interactions. By comparing the SHS potential values reconstructed on the 72 × 144 grid to the original orientational potential values for Ile-Arg shown in Fig. 5 (left), the smoothing effect of the SHA/SHS procedure becomes apparent.

The results of the same type of SHA/SHS process are shown in Fig. 6 for the anisotropic virtual backbone interaction centers (Pep) located in the middle of the peptide bond (see Fig. 1c). The smooth Pep-Pep orientational potentials are represented for short-range (top), middle-range (middle) and long-range (bottom) interactions. Blue regions correspond to attractive (i.e. negative) potentials, while red regions are positive, thus less likely to correspond to interaction loci. From both Figs. 5 and 6 we can see that, for each interaction range, there are specific anisotropic features of the orientation-dependent statistical potentials. Some of the attractive or repulsive angular regions are conserved from one interaction shell to the other. However, some present significant changes that may account for the specific features of residue–residue, residue–backbone and backbone–backbone interactions.

The continuous and smooth properties of the reconstructed statistical potentials allow us to represent them as three dimensional contour maps (Fig. 7). This figure also illustrates the relative LRF orientation of the virtual backbone interaction centers (Pep), with respect to the reconstructed statistical potential shown in Fig. 6. For the objects shown here, the color is directly proportional to the amplitude of the potentials. The negative attractive potential values are indicated as dark blue angular regions. The presence of other Pep particles are favored at these orientations. The red regions represent unfavorable and repulsive regions around Pep. This type of three dimensional representation can be used to effectively investigate all the possible side chain–side chain, side chain–Pep, and Pep–Pep orientation-dependent statistical potentials and their complex features.

2.6. Testing the smooth potentials: results from decoy sets using model C

One of the main features of the SHA/SHS approach is that specific values of the orientational potentials can be calculated (reconstructed) from the a_{mn} and b_{mn} coefficients for each specific value of the orientational parameters θ and ϕ. As such, the discontinuities that were originally present in the binned orientational potentials are eliminated. Our results indicate that the SHA/SHS procedure can be successfully used to describe the orientation-dependent potentials in a uniform and computationally convenient manner. One needs to keep in mind that there is an intrinsic information loss introduced by the SHS/SHA procedure that needs to be examined before the smooth reconstructed potentials can successfully replace the coarse, raw statistical data in coarse grained simulations [26,27].

After using the corresponding spherical harmonics coefficients to reconstruct the potentials, we performed tests for discriminating the native state from multiple decoy sets [8,9]. In Fig. 8 we present the results that were obtained for a test of the efficacy of our statistical potentials in

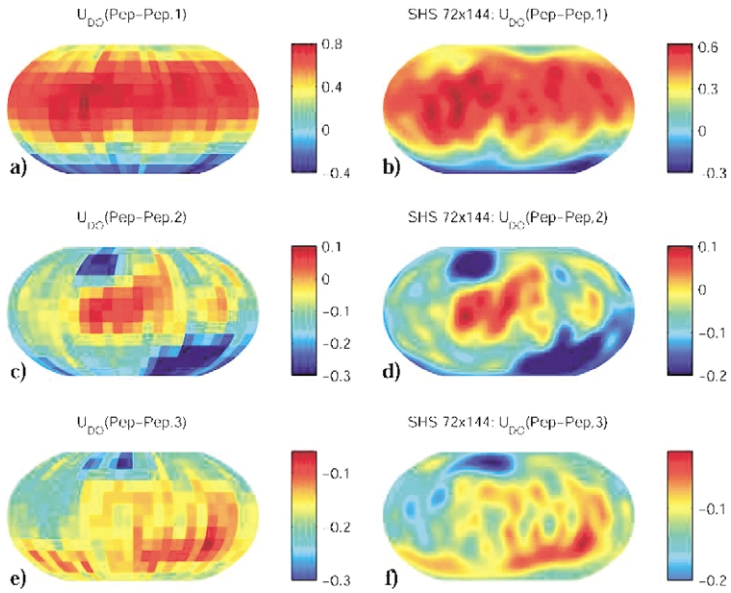

Fig. 6. The smooth Pep–Pep orientational potentials represented for short-range (top), middle-range (middle) and long-range (bottom) interactions. The potential values, calculated originally for a 12 × 24 equiangular grid are shown on the right (a, c and e), and the corresponding smooth potentials computed for a 72 × 144 grid using spherical harmonic synthesis are shown on the left (b, d and f). Blue regions correspond to attractive (i.e. negative) potentials, while red regions are positive, thus less likely to correspond to interaction loci.

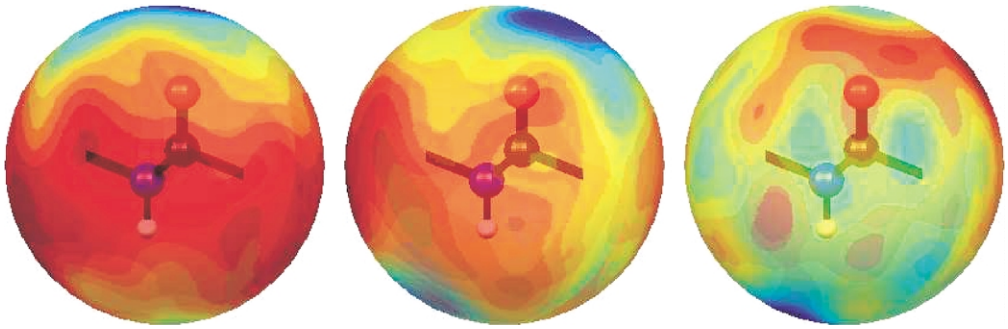

Fig. 7. Three-dimensional representations of the statistical potential field for the smooth short-range (right), middle-range (middle) and long-range (left) Pep–Pep interactions. The relative orientation of the Pep group of atoms is also shown. The blue, attractive regions responsible for hydrogen bonding are apparent for mid-range interactions (middle).

discriminating the native structure of a protein from the Samudrala and Levitt [9] decoy structures.

The results for the energy Z-score (Z_E) and C_α RMSD Z score (Z_{RMSD}) calculated for the multiple decoy sets [9,23, 29–32] lmds, fisa, 'fisa casp3' and 4state are compared in Fig. 8. The values of the distance- and orientation-dependent potentials (U_{DO}) were calculated using both the old 20×20 method ($U_{DO\text{-}20}$, dark bars) and the new 21×21 interaction scheme ($U_{DO\text{-}21}$, white). The cases where the new $U_{DO\text{-}21}$ potentials perform better are emphasized by the arrows on the left. It is observed that for a large majority of decoy sets (84% when considering the energy score Z_E) the performance is improved by including the backbone interaction centers.

The results of the tests to discriminate the native states from decoy sets show that the new 21×21 smoothed potentials perform better in a majority of cases. The results suggest that the anisotropic backbone interactions play an important role that might not be fully captured by simpler 20×20 models that consider only the interactions and conformations of side chains in an explicit manner.

3. A statistical potential for water

The derivation of knowledge-based inter-residue potentials for proteins is based on the assumption that the influence of the solvent environment is included implicitly in the interaction scheme, through a proper consideration of

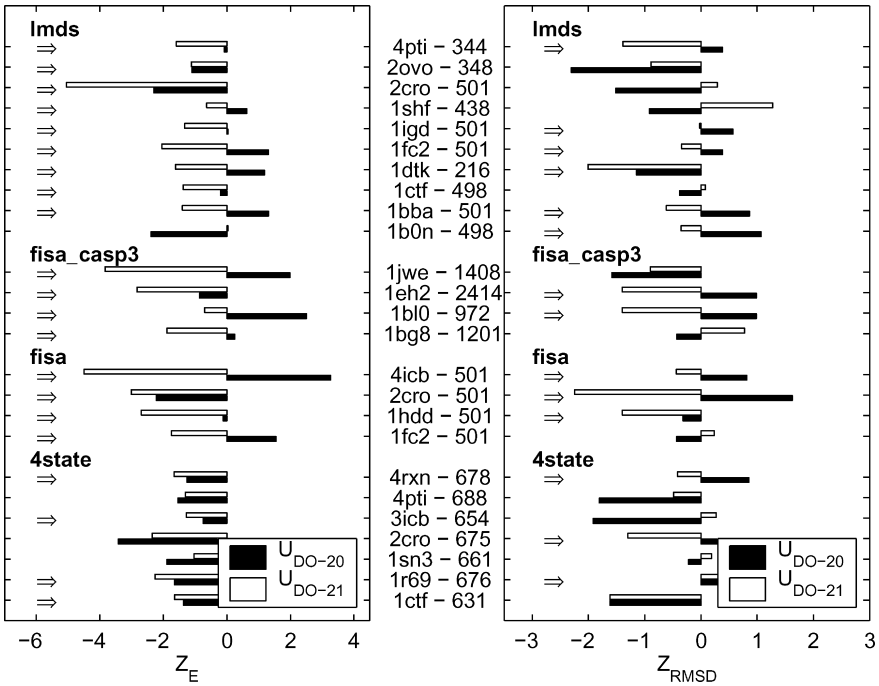

Fig. 8. Results from decoy tests. The energy (Z_E) and C_α RMSD Z-score (Z_{RMSD}) calculated for multiple decoy sets [9,23,30,31] 'lmds', 'fisa casp3', 'fisa' and '4state' are compared. More negative Z-scores are better. The name of each protein and the number of decoys in its corresponding set are shown in the center. The values of the distance- and orientation-dependent potentials (U_{DO}) were calculated using both the old 20×20 method ($U_{DO\text{-}20}$, dark bars) and the new 21×21 interaction scheme ($U_{DO\text{-}21}$, white). The cases where the new $U_{DO\text{-}21}$ potentials perform better in discriminating the native state from decoys are emphasized by the arrows on the left. For a majority of decoy sets (84% for Z_E and 56% for Z_{RMSD}) the performance is improved by including the backbone interaction centers.

the reference state [22,33]. However, in many cases (e.g. large scale coarse-grained simulations) it might be necessary to include the solvent explicitly. While there are many effective 2D [34,35] and 3D water models [36,37] that are currently used in molecular simulations, none is in full agreement with the measured equilibrium structural properties of water. Moreover, there is an on-going interest in investigating how the specific features of the underlying pair potential functions influence the structural and dynamic properties of various water models. [38] Our new methods for extracting orientation-dependent statistical pair-potentials for residue–residue interactions can be directly applied to investigating the effects of orientational anisotropy in water simulations. This approach serves both as a new test of the feasibility of building and using knowledge-based orientation-dependent potentials for molecular systems, as well as a novel method to build a simple, yet accurate, coarse-grained model for intermolecular interactions in liquid water. In this preliminary investigation we describe the most direct approach: deriving a single orientation-dependent interaction potential 'shell' and testing it in a small scale Monte Carlo NPT simulation [39,40]. The successful features and the limits of this simple statistical potential for liquid water are presented and discussed in this work. The simulated oxygen–oxygen radial distribution agrees reasonably well with recent experimental measurements [41,42].

3.1. Extracting the new water–water statistical potential

We applied the orientation-dependent 'Boltzmann device' procedure described in the previous sections to investigate the equilibrium statistical properties of liquid water. We performed a standard large molecular dynamics NPT simulation [39], of an equilibrated box of transferable inter-molecular potential with three points (TIP3P) [43] water at 310 K using NAMD [44]. We defined local reference frames for the individual water molecules, similar to the one previously employed [37], and we used these simulation results to extract a short range ([2,4] Å) statistical interaction potential for water. The results are shown as 2D and 3D maps in Fig. 9. The red and blue areas on these orientational probability maps (appearing as dark regions on a gray-scale plot of this figure) correspond to attractive regions (i.e. with a higher probability to coordinate another H_2O molecule) and, respectively, repulsive regions (i.e. with a lower interaction probability).

The orientational probability maps for liquid water are further used to build the corresponding orientation-dependent statistical potential maps for water, by employing the Boltzmann device. The water–water potentials can therefore be related to probability pair distribution functions $P(r,\phi,\theta)$ by the generalized distance- and orientation-dependent relation for the potential

$$U^{ij}_{\text{DO-wat}}(r,\phi,\theta) = -kT\ln\left[\frac{P^{ij}_{\text{wat}}(r,\phi,\theta)}{P_{\text{ref}}(r,\phi,\theta)}\right]. \quad (13)$$

We use the U_{DO} notation for the statistical potentials that are both distance-and orientation- dependent. For these studies of liquid bulk water, we consider the reference probability functions $P_{\text{ref}}(r,(\phi,\theta)$ to be a uniform radial or angular pair distribution.

In Fig. 10 is presented an orientational statistical

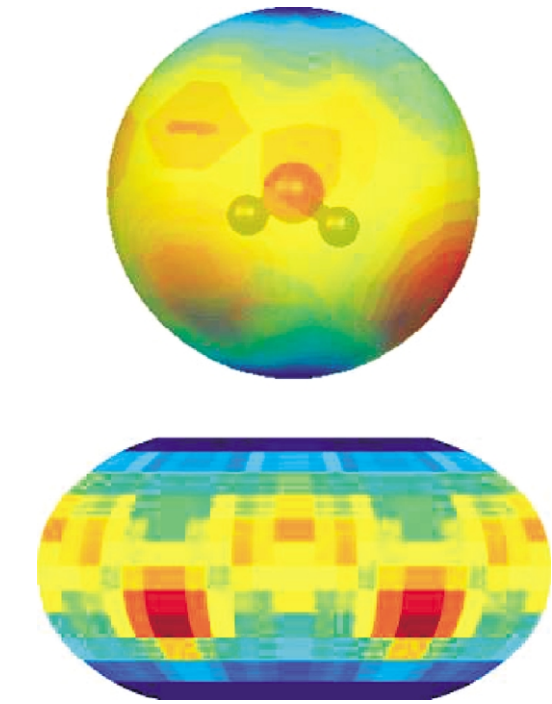

Fig. 9. Orientational interaction probability map for water. The 3D representation (top) and the corresponding orientation-dependent interaction probability map (bottom) for water molecules constructed on a 12×24 equiangular grid. The most and least probable interaction regions (red and blue, appearing both as dark on a gray-scale plot of this figure) are observed at locations related to the hydrogen bond formation loci.

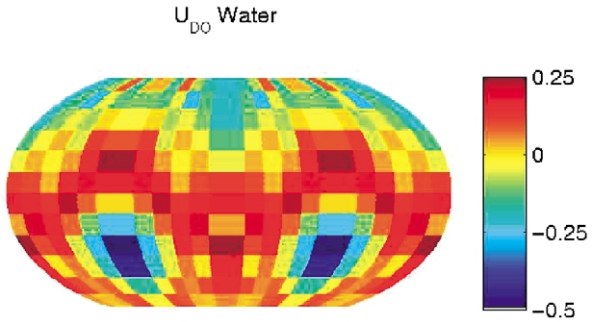

Fig. 10. Orientational statistical potential map for water. The orientation-dependent statistical potential map for water molecules constructed on a 12×24 equiangular grid for the [2,4] Å interaction range. The attractive and repulsive regions (red and blue, or appearing as dark on gray-scale representations of this figure) are related to the positions of hydrogen bond formation loci.

potential map for water constructed with this 'Boltzmann device.' This potential map for liquid water molecules was built on a 12×24 equiangular grid for the [2,4] Å interaction range. The attractive and repulsive regions (red and blue, or appearing as dark in gray-scale representations of this figure) are related to the positions of hydrogen bond formation loci.

3.2. Discussion: the static equilibrium features of the new water potential

We test the quality of this statistical potential by investigating the structural properties measured by the radial distribution functions [36,39]. In Fig. 11 is shown a comparison between an experimentally measured [41,42] radial distribution function for water at room temperature (ALS, borken curve) and the data (MC125) that we obtained from a Monte Carlo simulation using a small box of only 125 molecules with periodic boundary conditions. The potential employed in the MC simulation consisted of a hard core repulsive region up to 2 Å (region A), an orientation-dependent statistical potential shell from 2 to 4 Å (region B), and a very weak attractive region (C) with an isotropic potential of only -0.05 kT. The continuous line data (MC125-dr2') was collected with a bin size $dr_2 = 0.4$ Å and the dotted line (MC125-dr1') corresponds to a bin size $dr_1 = dr_2/10$.

A good correlation exists between the positions of the minima and maxima of the experimental and the simulation-derived radial distribution functions. The differences noticeable for the shape and position of the first peak are due to the fact that the orientation-dependent potential 'shell' has attractive regions that span the entire [2,4] Å interaction range. It is, therefore, an artifact of the simple, most direct application of the statistical potentials to this case. Such artifacts can be corrected by a more detailed model that employs multiple interaction ranges and which, at close separations (<2 Å), are better fitted for water–water interactions than the simplest hard core repulsive region A that is employed here.

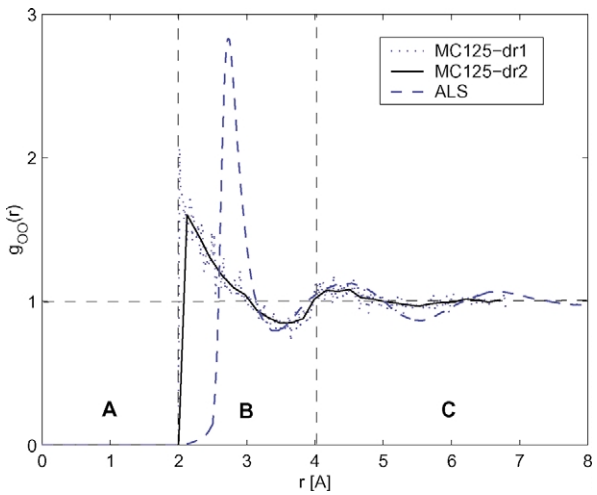

Fig. 11. Oxygen–oxygen radial distribution functions for water. Comparison between an experimentally measured radial distribution function for water at room temperature (ALS, dashed curve) and the data (MC125) from a Monte Carlo simulation using a small box of 125 molecules with periodic boundary conditions. The potential employed in the MC simulation consisted of a hard core repulsive region up to 2 Å (region A), an orientation-dependent statistical potential shell from 2 to 4 Å (region B), and a very week attractive region (C) with an isotropic potential of only -0.05 kT. The continuous line data (MC125-dr2) was collected with a bin size $dr_2 = 0.4$ Å and the dotted line (MC125-dr1) corresponds to a bin size $dr_1 = dr_2/10$.

These calculations demonstrate that a good modeling of the structural characteristics of liquid water may be obtained by employing statistically derived orientation-dependent potentials. The specific packing and local coordination geometry are controlled by the orientational anisotropy of the interaction potentials, an important factor that is usually treated by other potentials using distributed sites interacting via centrosymmetric Lennard–Jones and coulombic potentials. Our approach shows that specific structural features, such as the shape, position and number of the main peaks of the radial distribution function, which are generally thought to be directly related to various parameter values of distance-dependent-only or of more complex three-dimensional potentials, can be also obtained by employing a simple orientation-dependent 'knowledge-based' approach.

4. Conclusions

This paper presents results obtained by developing a new statistical method for building coarse-grained potentials using a generalized distance- and orientation-dependent Boltzmann device. Our successful application of this method to develop simple conformational models of proteins and small peptides demonstrates that the performance of energy based scoring functions can be improved by using statistical information extracted from the relative residue–residue orientations. The ability of the scoring functions to discriminate native-like protein structures is significantly enhanced by including the orientational dependence of side chain–side chain interactions as well as by including explicit anisotropic interaction centers that can model the side chain–backbone and backbone–backbone potentials. We have also demonstrated that the statistical data extracted from protein databases can be successfully used to build orientation-dependent potentials that have sufficient continuity properties to make possible their spherical harmonic analysis. Our novel smooth, continuous interaction potential is defined using separate spherical harmonic expansions of the orientation-dependent potential for short-, medium- and long-range interactions.

The new potentials were tested on a standard data base of artificially generated decoy structures [9] and the results demonstrate that the orientational information has both common and complementary significance as compared to the information that can be extracted from the relative

residue–residue distances alone. From a computational point of view, there are potential benefits both for free energy calculations and for coarse-grained dynamical simulations that might employ the continuous, smoother statistical potentials. The memory requirements for storing the spherical harmonic coefficients, as opposed to the raw orientational data, are smaller. In addition, the values of the potentials can be readily computed for any values of the (θ and ϕ orientational parameters specified over the entire spherical domain.

We have investigated further the feasibility of applying the orientation-dependent Boltzmann device to develop a new class of anisotropic statistical potentials for liquid water. These new potentials were extracted from detailed molecular dynamics simulations and were tested in a standard NPT Metropolis Monte Carlo simulation of liquid water [39,40]. The preliminary use of the new coarse-grained water model in simulations shows that specific features of the liquid water structure can be correctly reproduced. While other, more detailed anisotropic water potential models [36,37] have been also developed, our approach has the advantage of simplicity and the ability to be generalized to cases were a coarse-grained representation of groups of molecules is desirable.

Our orientation-dependent statistical potentials could be instrumental in developing more efficient computational methods for protein structure prediction as well as coarse-grained simulations on mesoscopic length scales.

Acknowledgements

This work was supported by the National Institutes of Health R01 NS41356-01 (JES and DT), the National Science Foundation CHE-9975494 (JES), and CHE-0209340 (DT). The authors are grateful to the Center for Scientific Computing and Visualization at Boston University for computational resources.

References

[1] Tanaka S, Scheraga HA. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules 1976;9:945–50.
[2] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucl Acids Res 2000;28:235–42.
[3] Lee J, Liwo A, Scheraga HA. Energy-based de novo protein folding by conformational space annealing and an off-lattice united-residue force field: application to the 10-55 fragment of staphylococcal protein A and to Apo Calbindin D9K. Proc Natl Acad Sci USA 1999; 96:2025–30.
[4] Miyazawa S, Jernigan RL. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. Proteins 1999;34:49–68.
[5] Godzik A, Kolinski A, Skolnick J. Are proteins ideal mixtures of amino-acids?—analysis of energy parameter sets. Protein Sci 1995;4: 2107–17.
[6] Sippl MJ. Calculation of conformational ensembles from potentials of mean force. J Mol Biol 1990;213:859–83.
[7] Sippl MJ. Knowledge-based potentials for proteins. Curr Opin Struct Biol 1995;5:229–35.
[8] Buchete N-V, Straub JE, Thirumalai D. Anisotropic coarse-grained statistical potentials improve the ability to identify native-like protein structures. J Chem Phys 2003;118:7658–71.
[9] Samudrala R, Levitt M. Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction. Protein Sci 2000;9:1399–401.
[10] Honeycutt JD, Thirumalai D. Metastability of the folded states of globular proteins. Proc Natl Acad Sci USA 1990;87:3526–9.
[11] Levitt M, Warshel A. Computer simulation of protein folding. Nature 1975;253:694–8.
[12] Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–52.
[13] Buchete N-V, Straub JE, Thirumalai D. Orientational potentials extracted from protein structures improve native fold recognition in preparation; 2003.
[14] Pearl FMG, Lee D, Bray JE, Sillitoe I, Todd AE, Harrison AP, Thornton JM, Orengo CA. Assigning genomic sequences to CATH. Nucl Acids Res 2000;28:277–82.
[15] Bahar I, Kaplan M, Jernigan RL. Short-range conformational energies, secondary structure propensities, and recognition of correct sequence-structure matches. Proteins 1997;29:292–308.
[16] Miyazawa S, Jernigan RL. Evaluation of short-range interactions as secondary structure energies for protein fold and sequence recognition. Proteins 1999;36:347–56.
[17] Bahar I, Jernigan RL. Coordination geometry of nonbonded residues in globular proteins. Fold Des 1996;1:357–70.
[18] Liwo A, Oldziej S, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. J Comput Chem 1997;18:849–73.
[19] Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Oldziej S, Scheraga HA. A united-residue force field for off-lattice protein-structure simulations. II. Parameterization of short-range interactions and determination of weights of energy terms by *Z*-score optimization. J Comput Chem 1997;18:874–87.
[20] Liwo A, Kazmierkiewicz R, Czaplewski C, Groth M, Oldziej S, Wawak RJ, Rackovsky S, Pincus MR, Scheraga HA. United-residue force field for off-lattice protein-structure simulations: III. Origin of backbone hydrogen-bonding cooperativity in united-residue potentials. J Comput Chem 1998;19:259–76.
[21] Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, Gottsbacher K, Casari G, Sippl MJ. Identification of native protein folds amongst a large number of incorrect models: the calculation of low energy conformations from potentials of mean force. J Mol Biol 1990;216:167–80.
[22] Thomas PD, Dill KA. Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol 1996;257:457–69.
[23] Simons KT, Kooperberg C, Huang ES, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997; 268:209–25.
[24] Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:895–916.
[25] Bahar I, Jernigan RL. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. J Mol Biol 1997;266:195–214.
[26] Adams JC, Swarztrauber PN. Spherepack 2.0: a model development facility, NCAR Tech. Note, NCAR/TN-436-STR.

[27] Adams JC, Swarztrauber PN. Spherepack 3.0: a model development facility. Monthly Weather Rev 1999;127:1872–8.
[28] Arfken GB, Weber HJ. Mathematical methods for physicists. New York: Academic Press; 1995.
[29] Keasar C, Levitt M. A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics. J Mol Biol 2003;329:159–74.
[30] Park B, Levitt M. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. J Mol Biol 1996;258: 367–92.
[31] Park B, Huang ES, Levitt M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J Mol Biol 1997;266:831–46.
[32] Fain B, Xia Y, Levitt M. Determination of optimal Chebyshev-expanded hydrophobic discrimination function for globular proteins, IBM. J Res Dev 2001;45:525–32.
[33] Betancourt MR, Thirumalai D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Sci 1999;8:361–9.
[34] Silverstein KAT, Haymet ADJ, Dill KA. A simple model of water and the hydrophobic effect. J Am Chem Soc 1998;120:3166–75.
[35] Silverstein KAT, Dill KA, Haymet ADJ. Hydrophobicity in a simple model of water: entropy penalty as a sum of competing terms via full, angular expansion. J Chem Phys 2001;114:6303–14.
[36] Liu Y, Ichiye T. Soft sticky dipole potential for liquid water: a new model. J Phys Chem 1996;100:2723–30.
[37] Chandra A, Ichiye T. Dynamical properties of the soft sticky dipole model of water: molecular dynamics simulations. J Phys Chem 1999; 111:2701–9.
[38] Stanley HE, Buldyrev SV, Giovambattista N, Nave EL, Mossa S, Scala A, Sciortino F, Starr FW, Yamada M. Application of statistical physics to understanding static and dynamic anomalies in liquid water. J Stat Phys 2003;110:1039–54.
[39] Allen MP, Tildesley DJ. Computer simulation of liquids. New York: Oxford University Press; 1990.
[40] Frenkel D, Smit B. Understanding molecular simulation: from algorithms to applications, 2nd ed. New York: Academic Press; 2002.
[41] Hura G, Sorenson JM, Glaeser RM, Head-Gordon T. A high-quality X-ray scattering experiment on liquid water at ambient conditions. J Chem Phys 2000;113:9140–8.
[42] Sorenson JM, Hura G, Glaeser RM, Head-Gordon T. What can X-ray scattering tell us about the radial distribution functions of water? J Chem Phys 2000;113:9149–61.
[43] Mahoney MW, Jorgensen WL. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. J Chem Phys 2000;112:8910–22.
[44] Kale L, Skeel R, Bhandarkar M, Brunner R, A G, Krawetz N, Phillips J, Shinozaki A, Varadarajan K, Schulten K. NAMD2: greater scalability for parallel molecular dynamics. J Comput Phys 1999; 151:283–312.