Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Molecular Graphics and Modelling 22 (2004) 441–450

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

Continuous anisotropic representation of coarse-grained potentials for proteins by spherical harmonics synthesis

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

[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

Abstract

A new method is presented for extracting statistical potentials dependent on the relative side chain and backbone orientations in proteins. Coarse-grained, anisotropic potentials are constructed for short-, medium-, and long-range interactions using the Boltzmann method and a database of non-homologous protein structures. The new orientation-dependent potentials are analyzed using a spherical harmonics decomposition method with real eigenfunctions. This method permits a more realistic, continuous angular representation of the coarse-grained potentials. Results of tests for discriminating the native protein conformations from large sets of decoy proteins, show that the new continuous distance- and orientation-dependent potentials present significantly improved performance. Novel graphical representations are developed and used to depict the orientational dependence of the interaction potentials. These new continuous anisotropic statistical potentials could be instrumental in developing new computational methods for structure prediction, threading and coarse-grained simulations.

Keywords: Coarse-grained potentials; Side chain orientation; Statistical potentials; Harmonic analysis; Boltzmann device

1. Introduction

The ability to predict protein structures, even at a low resolution level, has become important in the field of structure-based molecular biology. Despite advances in all-atom molecular simulation methods, it is still difficult to predict in detail protein folding dynamics and thermodynamics. To gain insight into the dynamics of folding and protein–protein interactions, it is desirable to develop a series of spatially coarse-grained models. A key ingredient needed for these models is an effective set of interaction potentials. Following the seminal work of Tanaka and Scheraga [1], there is a growing interest in obtaining reasonably accurate force fields. 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 [3–7]. Tanaka and Scheraga [1] proposed that the frequencies of amino acid pairing could be used to determine potential interaction parameters. Subsequently, with the exception of a few studies [8], most of the “knowledge-based” potentials have been obtained solely in terms of residue–residue contacts [6,9–11].

An explicit distance dependence of the statistical mean force potentials was introduced by Sippl [8,12] 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. From this assumption, it follows that the distribution of the distance, r, between two side chains (SC), should correspond to the equilibrium Boltzmann distribution. Other structural parameters including internal coordinates, such as dihedral angles, can also be used in this treatment. However, most statistical potentials developed using this approach and other methods [11,13,14] have only focused on distance-dependent probability density functions.

It is known that the relative orientation of side chains is an important determinant of the local (secondary structure) geometry as well as three-dimensional (3-D) (tertiary structure) topology [5,15,16]. By analyzing various families of structures, we observed that certain orientational order parameters are prominent [17]. In this paper, we present a method for building a set of continuous orientation-dependent coarse-grained statistical potentials for proteins, from the statistics of orientational distributions extracted from PDB structures. The method is implemented for extracting potentials on three distance intervals by

[*] Corresponding author.

E-mail addresses: straub@bu.edu (N.-V. Buchete), straub@bu.edu (J.E. Straub), thirum@glue.umd.edu (D. Thirumalai).

[1] Co-corresponding author.

1093-3263/$ – see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2003.12.010

considering short- (2.0–5.6 Å), medium- (5.6–9.2 Å) and long-range (9.2–12.8 Å) side chain–side chain (SC–SC) interactions. The near globularity of protein structures implies that backbone (BB) contacts should also play an important role. The explicit consideration of the backbone interactions is also supported by the results of previous statistical derivations of backbone potentials that used virtual bond and torsion angles [18] and secondary structure information [19]. To capture the effect of the number of side chain–backbone (SC–BB) contacts, we include an extra anisotropic backbone interaction center located at the peptide bond. A spherical harmonic analysis (SHA) and synthesis (SHS) of these new potentials is used to express the orientation-dependent potentials in a more realistic, smoothed representation. The effectiveness of these potentials in recognizing the native states is assessed using decoy tests [20] and compared to their raw, non-smoothed version. The results show that the new continuous orientation-dependent potentials present a significantly improved performance.

These new coarse-grained anisotropic potentials could be useful in structure prediction studies when being used in conjunction with a either a simplified SC–BB energy function [21] or with statistical information on SC–BB orientations from a detailed backbone-dependent rotamer library [22].

2. Methods

2.1. Coarse-grained model

As shown schematically in Fig. 1, in order to get parameters for the orientational dependence of the coarse-grained potentials, it is useful to define local reference frames (LRFs) for each side chain [17] For any SC, a LRF can be constructed by considering at least three non-collinear points (P_1, P_2 and P_3) that uniquely define the orientation of the LRF. A fourth point, usually denoted by S_i for the ith side chain, specifies the location of the LRF origin. 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. In this coarse-grained model for peptides and proteins, we include an additional interaction center located on the backbone [23,24] at the geometric center of each peptide bond (see Fig. 2). In this description, we assume that the local conformation of a given residue i is sufficiently well described by the corresponding C_α^i, S_i and Pep^i interaction centers.

The method for building the LRF for each side chain is summarized below, based on the notation used in Fig. 1b. 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 and a second direction $\hat{O}_y^*$ (pointing towards the O_y axis) can be constructed as

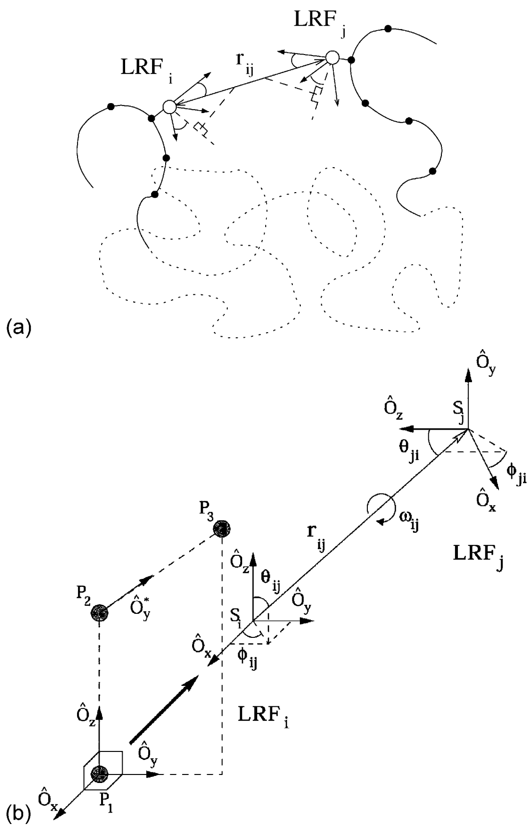


Fig. 1. Coarse-grained model for the quantitative study of the relative side chain–side chain and side chain–backbone in proteins: (a) schematic representation of the relative SC–SC 3-D coordination geometry; (b) local reference frames (LRFs) for two interaction sites i and j (P_1, P_2 and P_3 are needed to define the orientation of LRF_i, which is centered in S_i).

$$\hat{O}_z = \frac{\vec{r}_{P_2} - \vec{r}_{P_1}}{|\vec{r}_{P_2} - \vec{r}_{P_1}|} \quad \text{and} \quad \hat{O}_y^* = \frac{\vec{r}_{P_3} - \vec{r}_{P_2}}{|\vec{r}_{P_3} - \vec{r}_{P_2}|}. \qquad (1)$$

In the second step, the $\hat{O}_x$ and $\hat{O}_y$ axis vectors are defined using the cross products $\hat{O}_x = \hat{O}_y^* \otimes \hat{O}_z$ and $\hat{O}_y = \hat{O}_z \otimes \hat{O}_x$.

The positions of the three reference points P_1, P_2 and P_3 are identified for side chains with the positions of the C_α, C_β and C_γ atoms [17]. The positions of the interaction centers S_i are identified with the geometric centers (GC) of the heavy atoms in the side chains. Exceptions to these rules are made for the following special cases. (1) For Gly there is no C_β so we used the position of the midpoint between

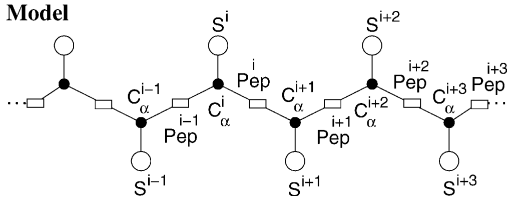


Fig. 2. Coarse-grained model: three types of particles (C_α, S, and Pep) are needed to study the SC–SC, SC–BB and BB–BB interactions.

the neighboring N^i and C^i atoms on the backbone as P_1 and C_α^i is taken to be P_2. In this way, the local O_z axis is defined by the bisector of the angle defined by N^i, C_α^i and C^i. (2) Because Gly and Ala do not have C_γ atoms, we used the position of the backbone atom C^i as P_3. In this way, the local O_y axis is pointing in the direction defined by the backbone atoms C_α^i and C^i. (3) For Cys and Ser the corresponding coordinates of the S and O atoms are substituted for the coordinates of the missing C_γ and are used, therefore, for defining P_3. (4) For Ile and Val, the coordinates of the midpoint between the two C_γ atoms are used for P_3.

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 towards more "remote" C_γ atoms in the SC. For small side chains, O_y points towards the next SC on the backbone sequence.

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.

2.2. Orientational probability maps

To extract and build orientation-dependent potentials from PDB structures we need to obtain the relative SC–SC, SC–BB and BB–BB orientational distributions from protein structures [17]. This data can be expressed as normalized relative orientational probability maps that are specific for each pair of interaction sites. For the set of non-homologous proteins used by Scheraga and co-workers [24–26], the orientational histograms were collected using $N = 12$ bins for the range of the θ angle and $2N$ bins for ϕ in the corresponding LRFs. Since all the protein structures analyzed have a resolution of 2 Å or better, the choice of bin sizes ensures a high confidence level of correct angular bin assignment (80% at a distance of at least 4.5 Å) [17].

The extracted SC–SC pair frequencies are transformed to SC–SC distance- and orientation-dependent interaction probabilities P^{ij} (r, ϕ, θ) by normalization. In the case of 3-D orientation-dependent data, the measured frequencies must also be divided by $\sin(\theta_k)$ to correct for the smaller volume elements near the poles when k equiangular intervals are used for the θ angle in the corresponding LRF. Because the amount of data available is relatively small for conventional statistical procedures, we employed the "sparse data correction" formula of Sippl [8,12] that builds the correct probability densities as linear combinations between the measured data and the reference, total probability densities obtained by averaging over all 20 SC types. As in previous studies [8,27,28], we used the value 1/50 for the constant σ, which corresponds to how many actual measurements must be observed such that both the actual probabilities and the reference would have equal weights.

2.3. The orientation-dependent potentials: the Boltzmann device

We used the Boltzmann device [8,12] to construct statistical orientational potentials from the orientational probability maps. This approach is based on the assumption that the known protein structures from protein databases (such as PDB [2]) correspond to classical equilibrium states. The SC–SC potentials can be, therefore, related to position pair distribution functions $g(r)$ by the relation

$$U_{\mathrm{D}}^{ij}(r) = -kT \ln\left[\frac{g^{ij}(r)}{g_{\mathrm{ref}}(r)}\right] \tag{2}$$

for the distributions depending only on distances. We define a more general distance- and orientation-dependent potential

$$U_{\mathrm{DO}}^{ij}(r, \phi, \theta) = -kT\left[\frac{P^{ij}(r, \phi, \theta)}{P_{\mathrm{ref}}(r, \phi, \theta)}\right] \tag{3}$$

Here, we use U_{DO} for the statistical potentials that are both distance- and orientation-dependent, and U_{D} for potentials that depend solely on inter-residue distances. To be consistent with previous studies, we consider the reference pair distribution functions g_{ref} to be the corresponding radial or angular pair distributions that are obtained through an analysis of all 20 residue types. Databases of non-homologous proteins are necessary for estimating the pair distributions and for extracting amino acid specific interaction potentials that are consistent with various protein architectures.

The total potential for the residue pair ij is

$$\begin{aligned} &U_{\mathrm{DO}}^{ij}(r_{ij}, \phi_{ij}, \theta_{ij}, \phi_{ji}, \theta_{ji}) \\ &\quad = U_{\mathrm{DO}}^{ij}(r_{ij}, \phi_{ij}, \theta_{ij}) + U_{\mathrm{DO}}^{ji}(r_{ij}, \phi_{ji}, \theta_{ji}) \end{aligned} \tag{4}$$

where pairwise additivity is assumed. Eq. (4) 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 $P^{ij}(r_{ij}, \phi_{ij}, \theta_{ij}, \phi_{ji}, \theta_{ji})$ [17]. The dependence of the U_{DO}^{ij} potentials on the torsional angle around r_{ij} (see Fig. 3 in [17]) is averaged out. The results suggest that there is no effect of the assumption that the interaction terms can be truncated as in Eq. (4) on the accuracy of the U_{DO} potentials.

2.4. Spherical harmonic analysis (SHA) and synthesis (SHS) of discrete potentials defined on spherical domains

The orientational dependence of the new inter-residue coarse-grained potentials can be expressed in terms of functions defined on spherical domains. For each interaction

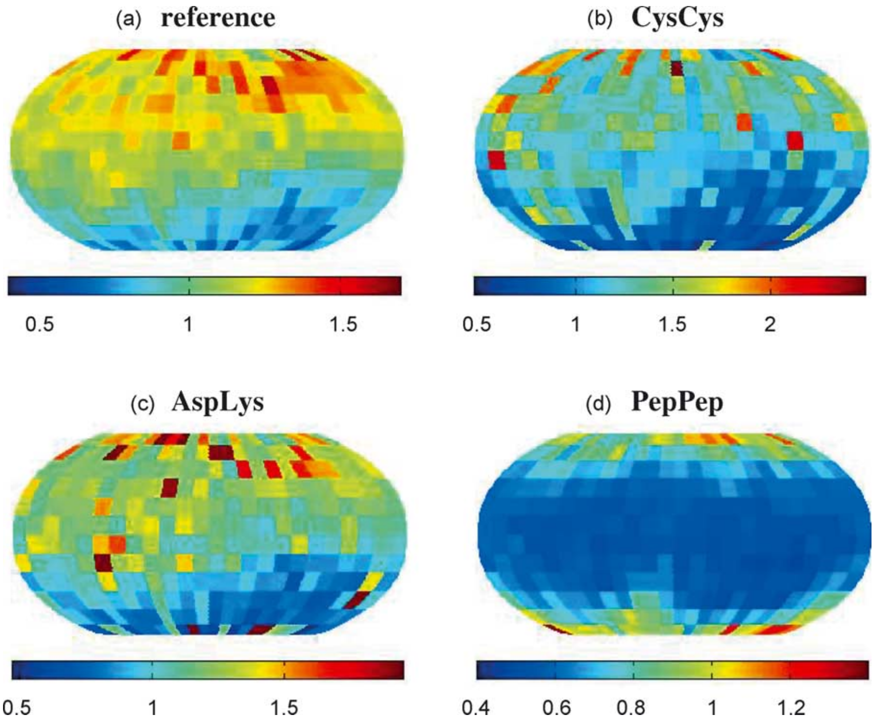


Fig. 3. Examples of orientation-dependent probability maps constructed for short-range interactions on a 12×24 angular grid. These graphical representations provide a global view of the interaction probabilities.

range, the angular dependent U_{DO} potentials are functions of the θ and ϕ polar angles defined in the corresponding LRFs of the amino acids [17]. These potential functions can be decomposed using

$$U(\theta, \phi) = \sum_{m,n} c_{mn} Y_{nm}(\theta, \phi) \tag{5}$$

where Y_{nm} are complex spherical harmonics [29] and c_{mn} are the expansion coefficients. This formula is valid only for functions $U(\theta,\phi)$ that have "well-behaved" continuity properties over the entire angular range. In practice, it is convenient to use a series with real even and odd eigenfunctions, namely,

$$U(\theta, \phi) = \sum_{m,n} [a_{mn} Y^{\mathrm{o}}_{nm}(\theta, \phi)] + b_{mn} Y^{\mathrm{e}}_{nm}(\theta, \phi). \tag{6}$$

This approach was successfully used for the accurate description of the geomagnetic field of the Earth [29].

We employed the technique developed by Adams and Swarztrauber and implemented in the FORTRAN package, Spherepack [30,31] which addresses problems associated with orthogonality at grid points and the non-uniform distribution of discrete data points. Though they were initially developed for geophysical processes, the Spherepack routines are general and can be successfully used to analyze the data extracted from protein structures, as follows. Let N be the number of grid points corresponding to sampling the data along the θ angle. We use $2(N-1)$ grid points for ϕ [31]. These sampling points are placed on the following equiangular grid

$$\theta_i = i\Delta\theta - \frac{\pi}{2}, \quad i = 0, 1, \dots, N-1,$$
$$\Delta\theta = \frac{\pi}{N-1}; \quad \phi_j = j\Delta\phi,$$
$$j = 0, 1, \dots, 2N-1, \quad \Delta\phi = \Delta\theta \tag{7}$$

Assuming that the angular dependent potential function is sufficiently smooth, one can perform its spherical harmonic analysis (SHA) and find the corresponding coefficients

$$a_{mn} = \alpha_{mn} \int_0^{2\pi} \int_{-\pi/2}^{\pi/2} U(\theta, \phi) P_n^m(\cos\theta)(\cos m\phi)\cos\theta \, \mathrm{d}\phi \, \mathrm{d}\theta \tag{8}$$

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

where P_n^m are the associated Legendre functions and $\alpha_{nm} = [(2n+1) \times (n-m)!]/[2\pi(n+m)]$ [29,30]. If the coefficients a_{nm} and b_{nm} are known, the corresponding smooth potential function $U(\theta,\phi)$ can be reconstructed using spherical harmonics synthesis (SHS)

$$U(\theta, \phi) = \sum_{n=0}^{N} \sum_{m=0}^{n'} P_n^m(\cos\theta)[a_{mn}\cos(m\phi)] + b_{mn}\sin(m\phi)] \tag{10}$$

The prime notation [30] on the sum indicates that the first term corresponding to $m = 0$ must be multiplied by 0.5.

This method of spherical harmonic analysis provides a realistic representation, through spherical harmonic synthesis, of the orientation-dependent statistical potentials as smoothed, continuous functions.

3. Results

3.1. Orientational probability density maps

Orientational probability density maps were constructed by dividing the interaction range into three regions, and using a 12 × 24 equiangular grid ($\theta \times \phi$) as described above, following the method introduced in [17]. Fig. 3 shows probability density maps in the short range (2.0–5.6 Å) of interactions. The color mapping is directly proportional to the probability of finding another side chain at a given orientation, as shown in the color bars. High interaction probability values appear as red while small probabilities are represented as blue.

The representations in Fig. 3 use pseudo-cylindrical orthophanic projections (a.k.a. Robinson projections) of the data values over the entire spherical (θ, ϕ) domain. These projections are commonly employed to represent mapping data for spherical geoids. The probability map in Fig. 3a was used as reference, constructed by averaging over interaction frequencies counted for all the 20 amino acid types. It is noticeable that there are relatively higher interaction probabilities toward the "north pole" (i.e. the positive O_z axis, pointing away from the local backbone) for this reference state. This is a manifestation of the finite size and compact packing of the protein structures. The probability maps constructed for Cys–Cys, Asp–Lys and Pep–Pep interactions are shown in Fig. 3b–d. The specific locations of statistically preferred interaction loci are observable. In particular, the Asp–Lys representation presents a few preferred directions for this type of SC–SC interaction. Orientations preferred for hydrogen bonding are clearly visible in the Pep–Pep probability maps. Propensity for disulfide bond formation manifests itself in the high probability in the polar region of the Cys–Cys probability map. For BB–BB interactions, we observe high interaction probabilities along the O_z direction, as expected. These features become more pronounced in the representations of the corresponding statistical potentials constructed from these probability maps.

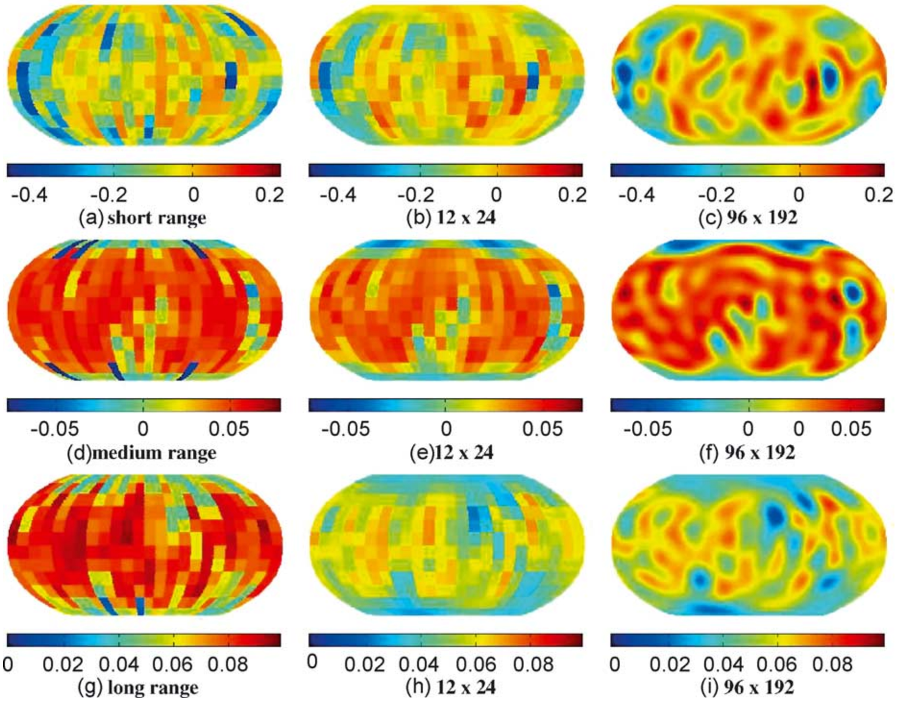
3.2. Continuous representations of orientation-dependent potentials

The orientation-dependent statistical potentials derived using the Boltzmann device were further analyzed using spherical harmonic analysis. Spherepack routines [30,31] were adapted and employed for the numerical analysis of the potential data, which was first constructed on a 12 × 24 equiangular grid on spherical domains corresponding to the three (i.e. short, middle and long) interaction ranges. a_{mn} and b_{mn} expansion coefficients were computed up to order $n = 13$ ($m \leq n$). The analysis of all 21×21 types of orientational potentials was performed and the a_{mn} and b_{mn} coefficients were stored. Calculation of the expansion coefficients (a_{mn} and b_{mn}, see Eq. (6)) is vital because it permits the rapid calculation of each specific orientational potential by spherical harmonic synthesis for any value of the LRF orientational parameters θ and ϕ. Importantly, not many a and b coefficients have large amplitudes suggesting that further filtering methods can be applied, and that efficient computational methods employing the new smooth potentials resulting from SHS can be developed. In Fig. 4 are shown potential projection maps. For Cys–Cys interactions to investigate the orientational preferences expected for disulfide bonds. The images in each row correspond to the same radial interaction range (e.g. the first row is for short-range, second row for middle-range, and third row for long-range interactions). The first column (i.e. Fig. 4a,d and g) represents the "raw" values (U_{DO21} for Cys–Cys) of the statistical potentials constructed directly from the corresponding probability maps using the Boltzmann device. The second column (i.e. Fig. 4b,e and h) represents the values of the statistical potentials reconstructed by using the SHA/SHS method (U_{DO21s}). Finally, in the third column (i.e. Fig. 4c,f and i) are shown the values of the statistical potentials reconstructed by using the same spherical harmonic analysis and synthesis method (on a much more detailed 96 × 192 equiangular grid)

For comparison, in Fig. 5 are shown the corresponding potential projection maps constructed for Gly–Gly interactions. The arrangement in columns and rows has the same significance as in Fig. 4. It is noticeable that due to the very small size of Gly, Gly–Gly interactions are described by a weak orientational preference when compared with Cys–Cys interactions. The magnitudes of the interaction ranges are shown in the horizontal color bar under each figure. Note that the magnitudes of the Gly–Gly interactions are generally much smaller, and present a different distance-dependence than in the case of Cys–Cys interactions, as expected.

We also show in Fig. 6 the corresponding potential projection maps constructed for Asp–Lys interactions. It is noticeable that the salt bridges that are likely to be formed between Asp and Lys, confer to the Asp–Lys potentials strong orientational statistical preferences in this interaction range. The strength of the attractive (blue) regions, as shown in the color bars, is also significantly larger than for Gly–Gly potentials shown in Fig. 5.

In Fig. 7 are shown 3-D representations of the orientation-dependent potentials for Pep–Pep (Fig. 7a) and for Gly–Gly (Fig. 7b) interactions. As above, the attractive regions are blue and the repulsive regions are red. The orientation-dependent potential values for short-range interactions are projected on the surface of a spherical geoid centered in the middle of the peptide bond for Pep–Pep or in the C_α for Gly–Gly interactions. This type of 3-D representation offers a useful way to visualize the locations of the orientational interaction loci with respect to the atomic

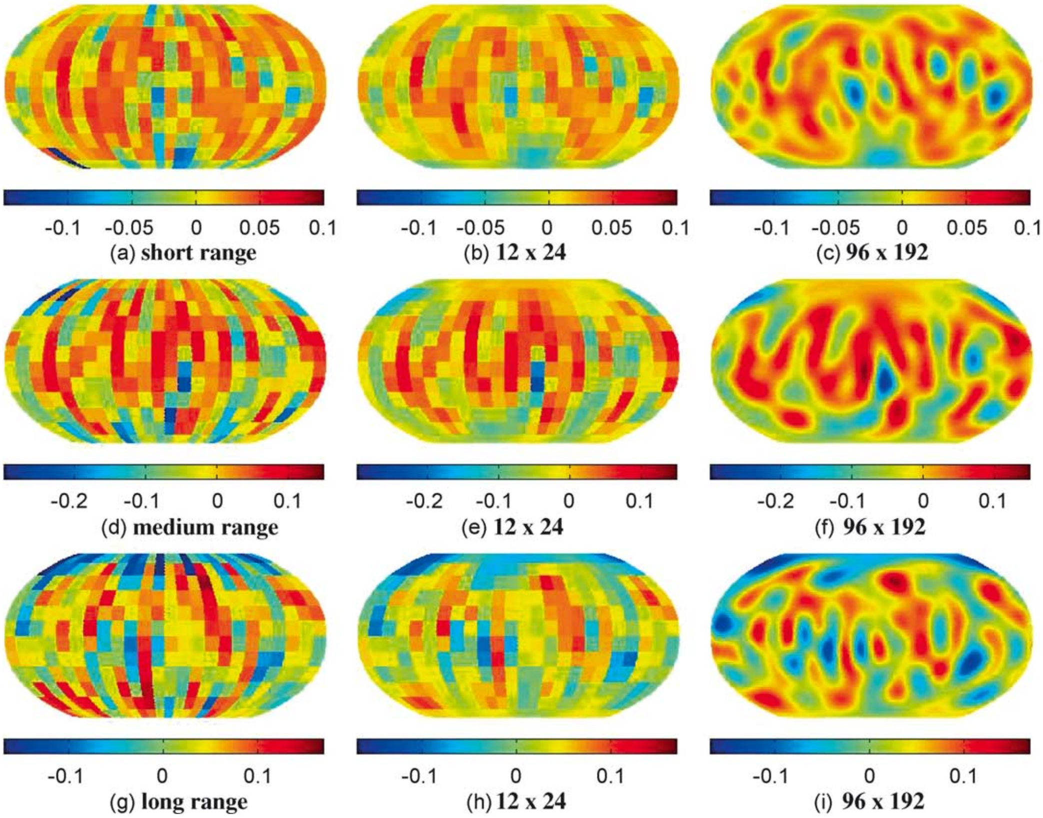


(a) short range (b) 12 x 24 (c) 96 x 192

(d)medium range (e)12 x 24 (f) 96 x 192

(g) long range (h) 12 x 24 (i) 96 x 192

Fig. 4. Cys–Cys potentials. The images in each row correspond to the same radial interaction range. The first column represents the raw potentials, in the second column are potentials reconstructed with SHS on a 12 × 24 angular grid, and in the third are shown potentials reconstructed on a 96 × 192 angular grid.



(a) short range (b) 12 x 24 (c) 96 x 192

(d) medium range (e) 12 x 24 (f) 96 x 192

(g) long range (h) 12 x 24 (i) 96 x 192

Fig. 5. Gly–Gly potentials. The images in each row correspond to the same radial interaction range. The first column represents the raw potentials, in the second column are potentials reconstructed with SHS on a 12 × 24 angular grid, and in the third are shown potentials reconstructed on a 96 × 192 angular grid.

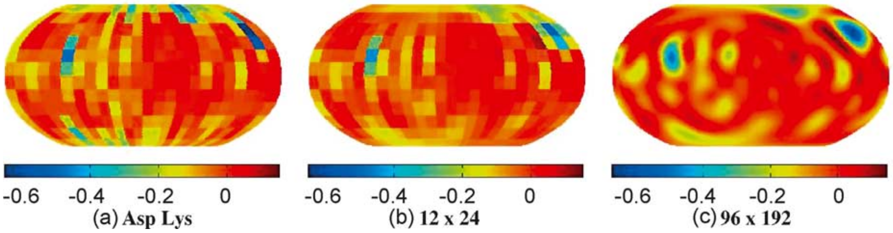


Fig. 6. Asp–Lys potentials for short-range radial interactions: (a) the raw potentials U_{DO21}, (b) potentials reconstructed with SHS on a 12 × 24 angular grid, and (c) the orientation-dependent U_{DO21s} potentials reconstructed on a 96 × 192 angular grid.

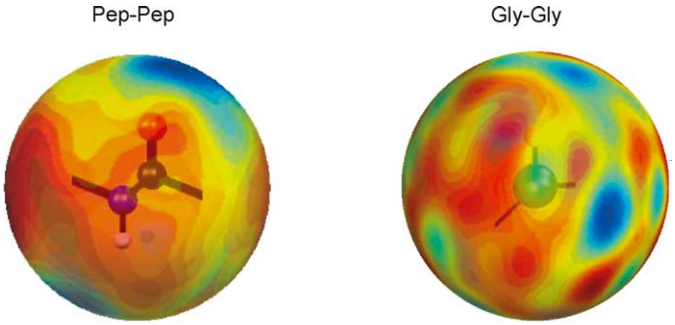


Fig. 7. 3-D representations of the orientation-dependent potentials for Pep–Pep and Gly–Gly interactions. The attractive (blue) and repulsive (red) potentials values are projected on the surface of a spherical geoid centered in the middle of the peptide bond for Pep–Pep or in the C_α for Gly–Gly interactions. The combined graphical representation of the "ball and stick" peptide bond and glycine α-carbon with a translucent sphere of projected potential energy, clearly represents the correlation between structure and orientational dependence of the interaction potential energy.

positions. However, such a representation is relatively difficult to implement for long side chains.

Due to the local reference frame definitions, the centers of the potential geoid surfaces should be located in the geometric center of the heavy atoms of the side chain.

In Fig. 8, is shown an alternative spherical contour plot representation of the orientation-dependent potentials for Cys–Cys interactions, from two diametrically opposite points of view. This representation is useful in cases when only a few interaction loci are found.

Finally, in Fig. 9 are shown 3-D representations of the reconstructed short-range orientation-dependent potentials U_{DO21s} for several types of interactions. Fig. 9a shows the values of the U_{DO21s} potentials reconstructed for Asp–Lys on a 12 × 24 equiangular grid. In Fig. 9b, the same Asp–Lys potentials are reconstructed with a resolution that is eight times more detailed than the original raw data. The same 96 × 192 equiangular grid is used for the representations of Asp–Lys, Ile–Arg, Pep–Pep, and Cys–Cys interactions in Fig. 9c–f. In these representations, the magnitude of the potentials is proportional to both the radius from the center of each local reference frame and to the color (i.e. red for repulsive and blue for attractive regions). It is therefore, possible to create 3-D shapes that would correspond both

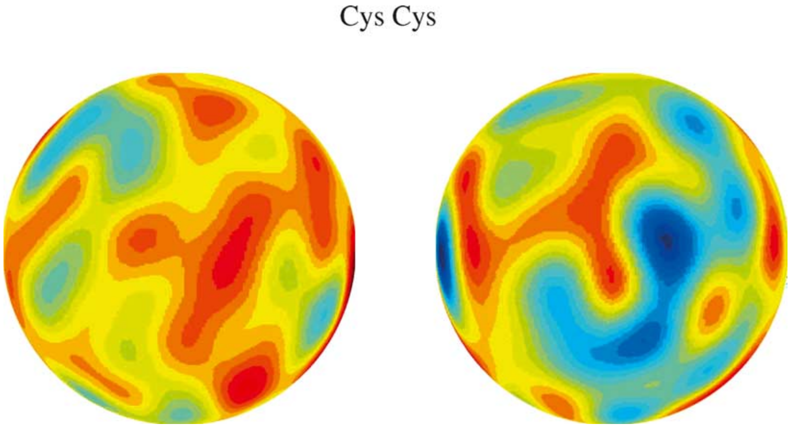
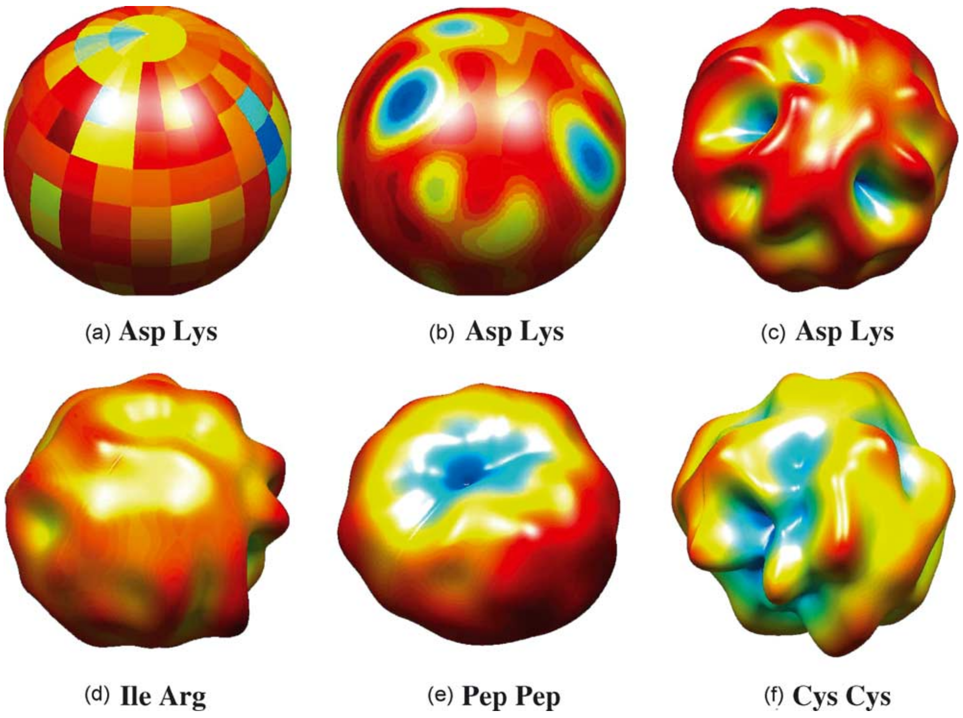


Fig. 8. Front and back views of the orientation-dependent potentials for Cys–Cys interactions.



(a) Asp Lys (b) Asp Lys (c) Asp Lys

(d) Ile Arg (e) Pep Pep (f) Cys Cys

Fig. 9. 3-D representations of the short-range orientation-dependent potentials U_{DO21s} constructed for: (a) Asp–Lys on a 12×24 equiangular grid, (b) Asp–Lys on a 96×192 equiangular grid, (c) Asp–Lys, (d) Ile–Arg, (e) Pep–Pep, and (f) Cys–Cys. In the graphical representations (c) to (f), the magnitude of the interaction potentials is proportional to both the radius from the center of each local reference frame and to the color (i.e. red for repulsive and blue for attractive regions).

qualitatively and quantitatively to the relative strengths and specific features of each SC–SC, SC–BB and BB–BB interaction type.

3.3. Decoy tests: improved Z score values

To assess the efficacy of the reconstructed orientational potentials, we performed tests for discriminating the native state from multiple decoy sets [17,20]. The results were obtained for testing the ability of our statistical potentials to discriminate the native structure of a protein from large sets of multiple decoy structures generated for the same protein sequence, using the decoy database of Samudrala and Levitt [20]. As in [17], the results are shown in terms of the values of the energy Z scores (Z_{E}), defined as

$$Z_{\mathrm{E}} = \frac{E - \overline{E}}{\sigma_{\mathrm{E}}} \tag{11}$$

where σ_{E} is the standard deviation and $\overline{E}$ is the mean of the distribution of E energy values corresponding to each decoy structure. For comparing the performance (and for studying the effect of smoothing) of the interaction potentials on sets of decoy structures, we calculate the energy Z_{E} scores both for the raw, backbone-dependent U_{DO21} potentials, and for their reconstructed and smoothed versions U_{DO21s}. Note that for a successful test of the interaction potentials, the Z_{E} score must be negative (i.e. the energy of the native state must have a lower value than the mean energy of the decoys).

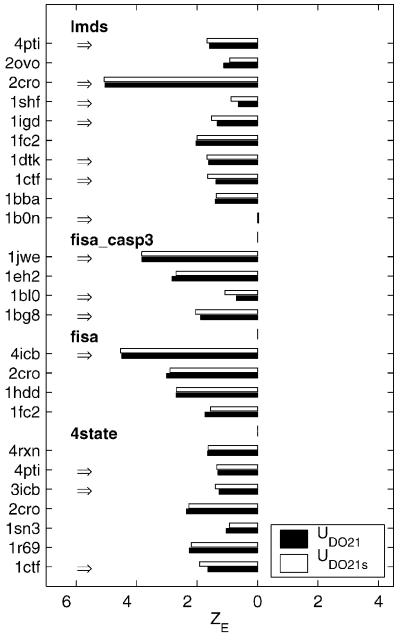


Fig. 10. Results from decoy tests. The energy Z scores (Z_{E}) calculated for multiple decoy sets [20,32–35] “lmds”, “fisa_casp3”, “fisa” and “4state” are compared before (U_{DO21}) and after (U_{DO21s}) applying the SHA/SHS method. The PDB code for each protein decoy set is shown on the left. The dark bars correspond to U_{DO21} and the white bars are for U_{DO21s}. The cases where the U_{DO21s} potentials perform better in discriminating the native state from decoys are emphasized by the arrows on the left. For a majority of decoy sets, the performance of the Z_{E} score is actually improved by using the spherical harmonic representation.

The data in Fig. 10 shows the results of these decoy tests. The energy Z scores (Z_E) were calculated for the multiple decoy sets [20,32–35] "lmds", "fisa_casp3", "fisa" and "4state". We compared the Z_E values obtained before (U_{DO21}) and after (U_{DO21s}) applying the smoothing reconstruction.

The dark bars correspond to U_{DO21} and the white bars are for U_{DO21s}. The cases where the U_{DO21s} potentials perform better in discriminating the native state from decoys are emphasized by the arrows on the left. While both potentials (U_{DO21} and U_{DO21s}) perform similarly well (i.e. they have negative Z_E scores), for a majority of decoy sets, the performance is actually improved by using the spherical harmonic representation. While there is an intrinsic information loss introduced [30,31], the potential smoothing that results appears to marginally improve the performance of the orientation-dependent potentials.

These results show that the smoothing of the orientational potential using the spherical harmonic analysis and synthesis approach does not necessarily lead to a loss of accuracy. In practice, it can actually lead to continuous, more realistic and computationally efficient representations of the orientation-dependent, coarse-grained interactions.

4. Conclusions

We have developed a method for building coarse-grained potentials using a generalized distance- and orientation-dependent statistical approach. We have successfully applied this method to develop a simple conformational model of proteins and small peptides that includes in an explicit manner the relative orientations of the SC–SC, SC–BB, and backbone–backbone (BB–BB) interactions. We have shown [17] that the performance of energy based scoring functions can be improved by using statistical information extracted from the relative residue–residue orientations. Our new results, obtained for this new set of anisotropic potentials with only three radial interaction ranges (the previous version [17] had more radial bins), demonstrate that the statistical data extracted from protein structural databases can be successfully used to build orientation-dependent potentials that have sufficient continuity properties to make possible their spherical harmonic analysis. The resulting smooth, continuous interaction potentials are represented 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 database of artificially generated decoy structures [20]. Although there is an intrinsic information loss introduced by the spherical harmonic analysis and synthesis [30,31], the new continuous orientation-dependent potentials lead to results that are consistent with, and in many cases marginally improved, when compared to the raw potentials constructed directly from orientational interaction probabilities. These results show that the smoothing of the orientational potentials using the SHA/SHS approach does not necessary lead to a loss of accuracy. In practice, it can lead to continuous, more realistic and efficient representations of the orientation-dependent, coarse-grained interactions.

A variety of graphical representations have been developed to effectively portray the orientational dependence of the statistical interaction potentials. These representations should be of value in comparative studies of orientational dependent potential functions for molecular fluids as well as proteins.

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. The new continuous distance- and orientation-dependent statistical potentials could be instrumental in developing more efficient computational methods for protein structure prediction as well as for Monte Carlo or molecular dynamics simulations of coarse-grained models of peptides and proteins.

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. The data visualization was carried out using Matlab (The Mathworks Inc., Natick, MA, USA).

References

[1] S. Tanaka, H.A. Scheraga, Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins, Macromolecules 9 (1976) 945–950.

[2] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucl. Acids Res. 28 (2000) 235–242.

[3] J. Lee, A. Liwo, H.A. Scheraga, 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. U.S.A. 96 (1999) 2025–2030.

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

[5] I. Bahar, R.L. Jernigan, Coordination geometry of non-bonded residues in globular proteins, Fold. Des. 1 (1996) 357–370.

[6] S. Miyazawa, R.L. Jernigan, Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues, Proteins 34 (1999) 49–68.

[7] A. Godzik, A. Kolinski, J. Skolnick, Are proteins ideal mixtures of amino-acids? Analysis of energy parameter sets, Protein Sci. 4 (1995) 2107–2117.

[8] M.J. Sippl, Calculation of conformational ensembles from potentials of mean force, J. Mol. Biol. 213 (1990) 859–883.

[9] J. Skolnick, A. Kolinski, A. Ortiz, Derivation of protein-specific pair potentials based on weak sequence fragment similarity, Proteins 38 (2000) 3–16.

[10] J. Skolnick, L. Jaroszewski, A. Kolinski, A. Godzik, Derivation and testing of pair potentials for protein folding. When is the quasi-chemical approximation correct? Protein Sci. 6 (1997) 1–13.

[11] D. Tobi, G. Shafran, N. Linial, R. Elber, On the design and analysis of protein folding potentials, Proteins 40 (2000) 71–85.

[12] M.J. Sippl, Knowledge-based potentials for proteins, Curr. Opin. Struct. Biol. 5 (1995) 229–235.

[13] D. Tobi, R. Elber, Distance-dependent, pair potential for protein folding: results from linear optimization, Proteins 41 (2000) 40–46.

[14] J. Meller, M. Wagner, R. Elber, Maximum feasibility guideline in the design and analysis of protein folding potentials, J. Comput. Chem. 23 (2002) 111–118.

[15] Z. Bagci, R.L. Jernigan, I. Bahar, Residue packing in proteins: uniform distribution on a coarse-grained scale, J. Chem. Phys. 116 (2002) 2269–2276.

[16] Z. Bagci, R.L. Jernigan, I. Bahar, Residue coordination in proteins conforms to the closest packing of spheres, Polymer 43 (2002) 451–459.

[17] N.-V. Buchete, J.E. Straub, D. Thirumalai, Anisotropic coarse-grained statistical potentials improve the ability to identify native-like protein structures, J. Chem. Phys. 118 (2003) 7658–7671.

[18] I. Bahar, M. Kaplan, R.L. Jernigan, Short-range conformational energies, secondary structure propensities, and recognition of correct sequence-structure matches, Proteins 29 (1997) 292–308.

[19] S. Miyazawa, R.L. Jernigan, Evaluation of short-range interactions as secondary structure energies for protein fold and sequence recognition, Proteins 36 (1999) 347–356.

[20] R. Samudrala, M. Levitt, Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction, Protein Sci. 9 (2000) 1399–1401.

[21] R. Kazmierkiewicz, A. Liwo, H.A. Scheraga, Addition of side chains to a known backbone with defined side-chain centroids, Biophys. Chem. 100 (2003) 261–280.

[22] A.A. Canutescu, A.A. Shelenkov, R.L. Dunbrack, A graph-theory algorithm for rapid protein side-chain prediction, Protein Sci. 12 (2003) 2001–2014.

[23] M. Levitt, A. Warshel, Computer simulation of protein folding, Nature 253 (1975) 694–698.

[24] A. Liwo, S. Oldziej, M.R. Pincus, R.J. Wawak, S. Rackovsky, H.A. Scheraga, A united-residue force field for off-lattice protein-structure simulations. Part I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data, J. Comput. Chem. 18 (1997) 849–873.

[25] A. Liwo, M.R. Pincus, R.J. Wawak, S. Rackovsky, S. Oldziej, H.A. Scheraga, A united-residue force field for off-lattice protein-structure simulations. Part II. Parameterization of short-range interactions and determination of weights of energy terms by Z-score optimization, J. Comput. Chem. 18 (1997) 874–887.

[26] A. Liwo, R. Kazmierkiewicz, C. Czaplewski, M. Groth, S. Oldziej, R.J. Wawak, S. Rackovsky, M.R. Pincus, H.A. Scheraga, United-residue force field for off-lattice protein-structure simulations. Part III. Origin of backbone hydrogen-bonding cooperativity in united-residue potentials, J. Comput. Chem. 19 (1998) 259–276.

[27] M. Hendlich, P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, M.J. Sippl, 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. 216 (1990) 167–180.

[28] P.D. Thomas, K.A. Dill, Statistical potentials extracted from protein structures: how accurate are they? J. Mol. Biol. 257 (1996) 457–469.

[29] G.B. Arfken, H.J. Weber, Mathematical Methods for Physicists, Academic Press, 1995.

[30] J.C. Adams, P.N. Swarztrauber, Spherepack 2.0: A Model Development Facility, NCAR Tech. Note, NCAR/TN-436-STR.

[31] J.C. Adams, P.N. Swarztrauber, Spherepack 3.0: a model development facility, Monthly Weather Rev. 127 (1999) 1872–1878.

[32] K.T. Simons, C. Kooperberg, E.S. Huang, D. Baker, Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions, J. Mol. Biol. 268 (1997) 209–225.

[33] C. Keasar, M. Levitt, A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics, J. Mol. Biol. 329 (2003) 159–174.

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

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