# A Novel Orientation-Dependent Potential for Protein Structure Prediction

Venkatesh Sivaraman[1], Eshel Faraggi[2,3], and Andrzej Kloczkowski[3,4]

[1]Bexley High School

[2]Indiana University School of Medicine

[3]Nationwide Children's Hospital

[4]Ohio State University

November 11, 2015

## Abstract

Protein structure prediction is one of the major unsolved problems in computational biology, primarily due to the inefficiency and short time scales of detailed simulations. We describe a novel statistical potential, called Segmented Positional Analysis of Residue Contacts (SPARC), which approximates the energy of a protein structure based on orientational inter-residue interactions as well as a packing density-like hydrophobicity term. Relative orientations are measured from a sample of 91,995 structures from the Protein Data Bank. The accuracy of SPARC is demonstrated first by comparison to CHARMM, a physics-based force field, and then to other statistical potentials on the gapless threading problem. SPARC showed an accuracy of 68%, superior to a distance-dependent method and comparable to GOAP, an all-atom anisotropic potential.

In addition to the new statistical potential, a novel adaptation of the Metropolis-Hastings algorithm has been designed for conducting Monte Carlo simulations of protein folding. The new method dynamically segments the protein into clusters to control variability and selects conformations with "permissible" orientations to ensure chain connectivity. We apply this algorithm to the structure prediction of bovine pancreatic trypsin inhibitor, yielding a structure with root mean square deviation of 8.78 Å from the native structure after only 2,000 iterations. Together, SPARC and the new simulation technique will help biologists study the causes of protein folding disorders, model enzyme behavior, and engineer new proteins.

# 1 Introduction

Predicting the 3-dimensional structure of proteins remains a challenge despite advances in theory and computational power over the past three decades. Modeling protein folding has numerous biological applications, including active site detection, protein design, and visualizing the formation of complexes, ligand binding, and protein-membrane interactions [2, 15, 22]. However, the most detailed simulations must incorporate hundreds of atoms at picosecond time intervals, which currently prohibits the timescale on which these simulations can be computed. Therefore, many biological applications would benefit from a coarse-grained approach that preserves as much detail as possible from atomistic methods.

The thermodynamic hypothesis, proposed by Anfinsen, stipulates that the native structure corresponds to the free energy minimum of all possible conformational states of the protein [1]. Therefore, the crux of structure prediction methods is to accurately express the energy of a protein in a given state. Energy functions generally fall under two categories: "physics-based" and "knowledge-based potentials" [17]. Physics-based potentials and force fields, such as CHARMM [5], AMBER [10], and GROMOS [29], evaluate conventional bonded and nonbonded energy terms (e.g. bond stretch, dihedral angle, and Coulombic potentials) for all atoms in a structure [6]. These atomistic potentials can be made coarse-grained by modeling residues as one or a few particles, or by considering groups of residues as rigid subparts [3, 26, 11, 22]. However, physics-based potentials necessitate either the inclusion of explicit solvent molecules in the simulation [25] or the use of implicit solvent techniques such as the generalized Born/solvent-accessible surface area (GBSA) model [12, 28]. This tends to render all-atom force fields computationally inefficient if not intractable in large-scale molecular dynamics (MD) applications.

The second type of energy function, knowledge-based potentials (KBPs), have developed as an approximative substitute for physics-based potentials and are more efficient at comparative tasks such as the gapless threading problem, which involves choosing the most likely native structure out of a series of candidates [33]. KBPs, also known as "statistical potentials," are based on statistical analysis of known protein structures rather than distinct physical or chemical interactions. The use of KBPs was pioneered by Tanaka and Scheraga [32], then Miyazawa and Jernigan [19], who estimated inter-residue interaction energies by counting contacts (pairs of residues located within a certain cut-off distance of each other) between types of amino acids in known protein structures.

The underlying assumption in assembling these statistical energy functions is that the known structures, usually obtained by X-ray crystallography or NMR, correspond to equilibrium states [7]. As a consequence, the frequency of a local structure (e.g. a contact, distance, or relative orientation) can be related to its conformational energy according to the Boltzmann distribution. This yields the following commonly-used expression for calculating energies by statistical analysis:

$$E(s) = -kT \ln P(s) \tag{1}$$

where $s$ represents a local conformation, $k$ is Boltzmann's constant, $T$ is the temperature, and $P(s)$ is the probability of the state occurring in equilibrium [31].

Beyond contact potentials, various methods have also been developed based on the distribution of Euclidean distances between residues, such as DOPE [30], DFIRE [37], GOAP [36], and others [17, 35]. A few studies have incorporated anisotropic factors by comparing the orientations of the sidechains [34, 23] or by measuring polar, spherical and/or Euler angles between two contacting residues [20, 7]. In general orientation-dependent energy terms based on bond angles are also combined with dedicated distance-dependent components, though the additivity of statistical energy terms concerning orientation and distance has been questioned [30]. Moreover, these statistical potentials may be limited by their concern with only the backbone or the sidechains, as well as failure to consider the tendency of hydrophobic residues to move toward the protein core [24].

This paper presents a new anisotropic statistical potential, called Segmented Positional Analysis of Residue Contacts (SPARC), that addresses concerns with other statistical methods by generalizing the expression of amino acid orientations through local coordinate system transformations. We describe the methods used to derive SPARC from the database of known protein structures, as well as a "coordination number"-based implicit solvent interaction model. SPARC is then evaluated based on its performance on the gapless threading problem in comparison to other statistical potentials, and tested in a new segmented dynamic Monte Carlo simulation of protein folding.

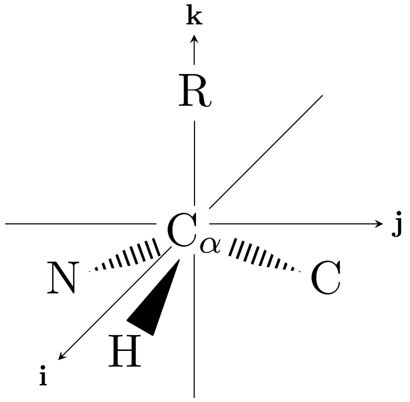

**Figure 1** – The core structure of an amino acid around the $\alpha$-carbon. We utilize the bond angles predicted by VSEPR theory to assign each amino acid a local coordinate system.

# 2 Methods

## 2.1 Construction of local coordinate system

Given a protein $A$ of $n$ amino acids, we can assign a Cartesian local coordinate system (LCS) to each residue to quantify its orientation with respect to an arbitrary global coordinate system (GCS). To determine a set of basis vectors that is consistent across all 20 sidechain types, we turn to the ideal bond angles predicted by valence-shell electron-pair repulsion (VSEPR) theory [13]. As shown in Fig. 1, the bond geometry around the $C_\alpha$ atom is approximately tetrahedral, with bond angles of $\cos^{-1} 1/3 \approx 109.5°$.

Defining $\mathbf{n}$ and $\mathbf{c}$ as the normalized vectors in the GCS pointing from $C_\alpha$ to the amine nitrogen and carbonyl carbon, respectively, we define $\mathbf{j} \equiv \pm(\mathbf{n} - \mathbf{c})$, with the sign that minimizes the angle between $\mathbf{j}$ and $\mathbf{c}$. We proceed with the $\mathbf{k}$ vector, which should lie along the bond leading to the sidechain (or a hydrogen atom in the case of glycine). Solving for a unit vector $\mathbf{k}$ from its bond angles with $\mathbf{c}$ and $\mathbf{n}$,

$$\mathbf{n} \cdot \mathbf{k} = -1/3$$
$$\mathbf{c} \cdot \mathbf{k} = -1/3$$

we obtain two solutions, each of which has an associated vector $\mathbf{i} = \mathbf{j} \times \mathbf{k}$. The pair that maximizes the angle between **i** and **c** produces an LCS consistent with Fig. 1.

## 2.2 From orientational frequencies to energies

The Cartesian LCS for each amino acid enables us to describe the relative orientation between any two amino acids **p** and **q** as simply the coordinates of **p** in the LCS of **q**, denoted $\mathbf{p}_q$, and vice versa. To obtain a distribution of relative orientations for each sidechain type, we consider only those amino acids within a contact sphere of radius 10 Å of each residue, a set we denote $A^*$. The orientation space is coarse-grained into intervals of 1 Å for a total of 8,000 possible "bins" into which each relative location might fall. (Because the contact shell is spherical and contained within a rectangular orientation space, in reality only about 4,200 bins would be used.)

As in all statistical potentials, SPARC assumes that these pairwise orientations follow the Boltzmann distribution in an ensemble of native protein structures [31], motivated by the fact that even at a free energy minimum, unstable *local* structures can still be found. This permits us to adapt Eq. (1) to calculate a dimensionless energylike quantity from the orientational frequencies:

$$S(\mathbf{p}) = -\sum_{\mathbf{q}\in A_p^*} \left( \ln \frac{f(\mathbf{p}_q) f_{pq,\text{tot}}}{\bar{f}_{pq}\bar{f}_{\text{tot}}} + \ln \frac{f(\mathbf{q}_p) f_{pq,\text{tot}}}{\bar{f}_{pq}\bar{f}_{\text{tot}}} \right) \tag{2}$$

In this equation, which sums over all amino acids within the contact sphere $A_p^*$, $f(\mathbf{p}_q)$ refers to the frequency of an interaction between amino acids of type $p$ and $q$ with that particular relative location. $\bar{f}_{pq}$ represents the mean frequency over all bins, and $f_{pq,\text{tot}}$ the total number of contacts for these amino acid types (e.g. alanine and tyrosine, glycine and glutamine, etc.). $\bar{f}_{\text{tot}}$ is the mean number of contacts over all combinations of amino acid types. The result of this formulation is that contacts whose relative orientations are above average receive a more negative (stable) score. Furthermore, for a perfectly "average" orientation, a more negative score will result if contacts between amino acid types $p$ and $q$ are more frequent in the native ensemble.

Another important consideration when constructing SPARC was chain connectivity, since it is presumed that amino acids adjacent to each other along the peptide backbone interact differently (due to the presence of connecting covalent bonds) than non-consecutive residues. Therefore, the consecutive and non-consecutive amino acids found within the 10-Å shell were separated and analyzed separately, producing two distinct score functions $S_c$ and $S_{nc}$.

## 2.3 Solvent interaction model

In a simplified sense, the intermolecular interactions that determine the stability of a protein structure can be decomposed into those between residues and those between residues and the solvent. The latter interactions cannot be modeled by orientation in an implicit solvent, so they are collectively modeled by SPARC in terms of the "coordination number" or packing density of the residues. This approximation is similar to the coordination numbers used by Miyazawa and Jernigan [19]; however, instead of calculating the values indirectly from the volume of each residue type, we are able to count the number of amino acids within the contact sphere because of the large sample size and variety in our structures (see section 2.5). The analog for eq. (2) for the solvent interactions then becomes

$$S_{\mathrm{solv}}(\mathbf{p}) = -\ln \frac{f(|A_p^*|)}{\bar{f}_p}, \tag{3}$$

where $\bar{f}_p$ is the mean frequency over the possible coordination numbers for amino acid type $p$. The energylike quantities represented by $S_c$, $S_{nc}$, and $S_{\mathrm{solv}}$ are finally combined linearly to produce an overall SPARC score value.

## 2.4 Dynamic Monte Carlo simulation

The statistical nature of SPARC makes it especially conducive to a Markov chain Monte Carlo (MCMC) simulation because it provides direct local stability comparisons which can be used to sample the conformational space. Monte Carlo simulations have been used with some frequency to predict protein structure [14, 11], though not as often or as successfully as deterministic molecular dynamics (MD). The most widely-used MCMC technique is the Metropolis-Hastings algorithm, which uses a weighted form of rejection sampling to approximate a distribution over a large number of iterations [18]. We implement a modified version of the Metropolis-Hastings algorithm based on the distribution of relative orientations in SPARC, which should improve the speed and viability of Monte Carlo methods for protein structure prediction.

An overview of the simulation algorithm is presented in Fig. 2. During each iteration, segments containing 1–5 residues are chosen at random, with short segment lengths favored in unstable structures and longer segments favored in stable structures. This allows us to control the preservation

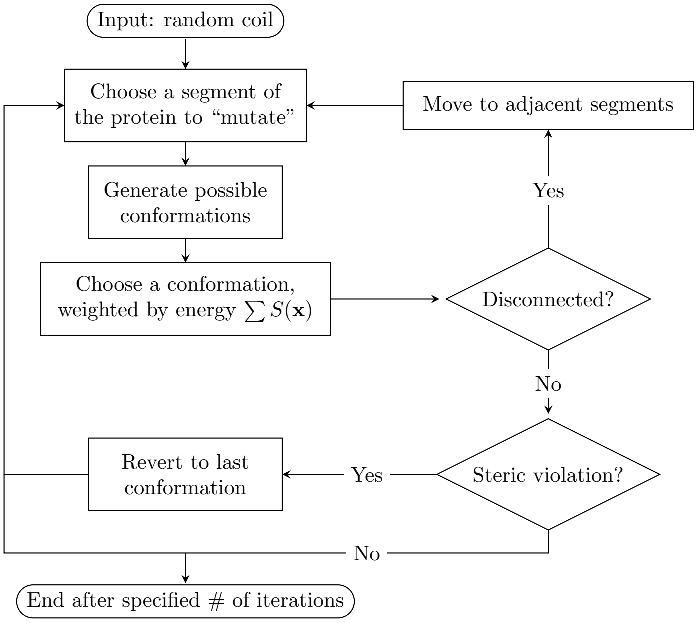
Input: random coil
Choose a segment of the protein to "mutate"
Move to adjacent segments
Generate possible conformations
Yes
Choose a conformation, weighted by energy $\sum S(\mathbf{x})$
Disconnected?
No
Revert to last conformation
Yes
Steric violation?
No
End after specified # of iterations

**Figure 2** – An overview of a Markov chain Monte Carlo (MCMC) simulation algorithm developed using SPARC. The procedure is similar to that of the Metropolis-Hastings algorithm in that new structures are sampled so that better-scoring conformations are chosen more often.

of locally-stabilizing interactions. The SPARC scores of each possible "mutation" to the chosen segment are transformed into probabilities:

$$P(C) = \exp\left(-\sum_{\mathbf{a}_i \in C} S(\mathbf{a}_i)\right) \tag{4}$$

A cumulative distribution function is then obtained from the normalized values of $P(C)$, and sampled to determine the next local conformation.

To maintain chain connectivity, two additional modifications to the Metropolis-Hastings approach are necessary. First, the locations of consecutive residues are constrained to "permissible" orientations, which are found with at least a certain frequency (0.5%) in the native ensemble; this ensures that the bonding between residues is realistic. Second, the neighboring segments of the protein are adjusted to amend the breaks caused by each mutation.

## 2.5 Materials and software

All calculations and programs were run on an off-the-shelf laptop computer and a desktop. Written in Python 2.7 and comprising about 6,000 lines of code, this first version of the SPARC software

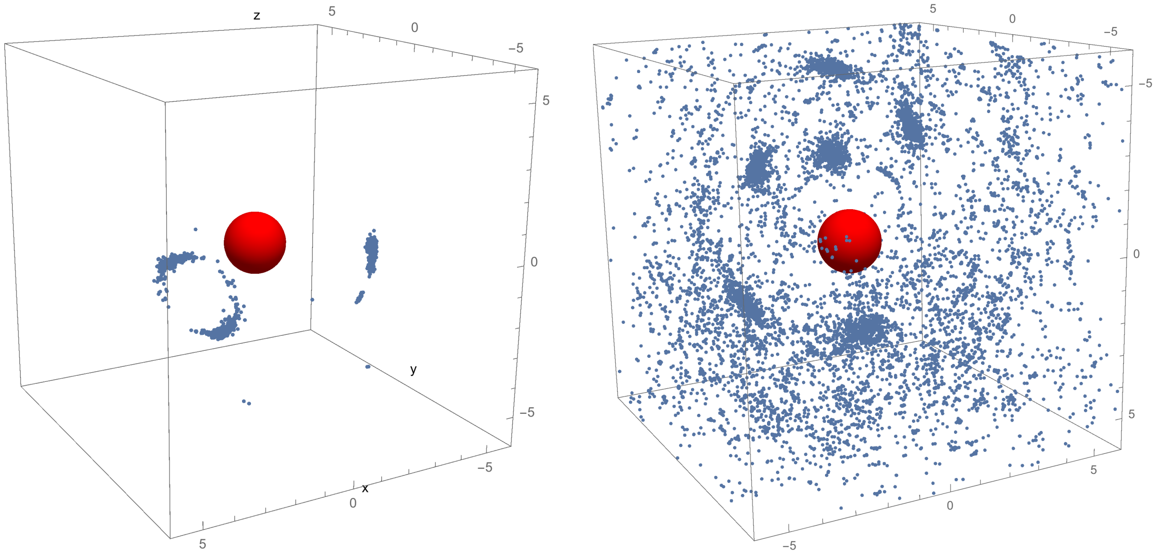

**Figure 3** – A portion of the distribution of C$_\alpha$ relative locations between two alanine residues, consecutive (left) and non-consecutive (right). The consecutive distribution is much more specific due to the constraints imposed by peptide bond, while the nonconsecutive distribution permits many more possible locations.

contains tools for reading and writing PDB files, analyzing structures for orientation data directly from the Protein Data Bank, and running the dynamic Monte Carlo simulations as well as other utilities, with all modules designed to be subclassable and extendible for future modifications.

To calculate the orientational frequencies, a large dataset of native protein structures determined using X-ray crystallography (91,995 total structures) was obtained from the Protein Data Bank. When performing relative orientation calculations, a hash table is utilized to map the amino acids into spatial "buckets" of dimension 5 Å based on the location of the $\alpha$-carbon. This resulted in a major performance improvement over searching the entire protein conformation for pairwise contacts, from $O(n)$ for each calculation to $O(1)$.

# 3 Results

## 3.1 Frequency distributions

The large sample size of proteins used to construct SPARC allowed us to establish fairly small bins without considerable loss in accuracy. For consecutive amino acids, it was expected that the vast majority of relative orientations would be located in a small number of bins, corresponding to the

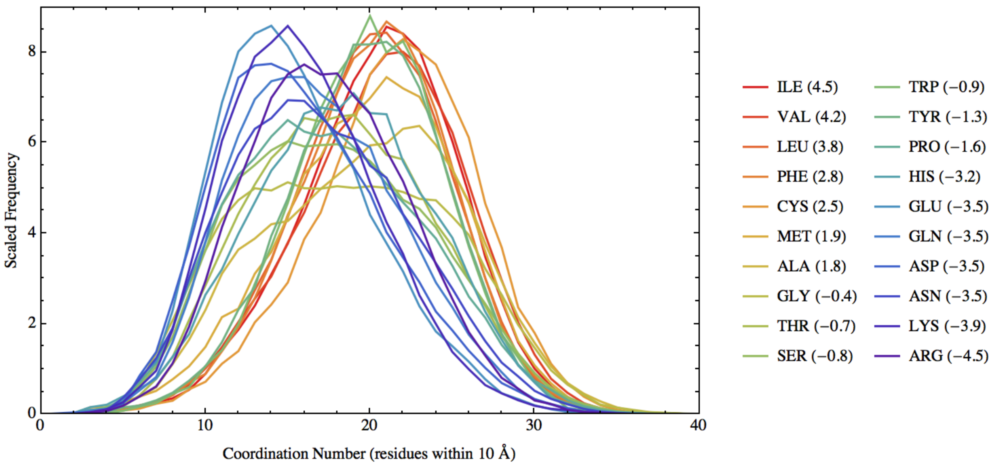

**Figure 4** – The distributions of coordination numbers of each amino acid type, scaled from 1–10. The Kyte-Doolittle hydrophobicity values are given in parentheses; the blue hues are more polar residue types, and the red hues are more nonpolar.

allowed angles of the peptide bond; this hypothesis was confirmed in Fig. 3, left. On average, 505 bins were filled and only 27 contained more than 50 occurrences. The nonconsecutive distributions showed much more variability, also as expected because of the lack of a restrictive covalent bond (Fig. 3, right): 4,032 bins were non-empty, almost all of the ~4,200 possible bins.

While other techniques we attempted (orientations with respect to nearby residues, distance to the geometric center of the protein) failed to capture hydrophobicity sufficiently, the coordination-number approach does seem to accurately reflect the conventional hydrophobicity scale reported by Kyte and Doolittle [16]. Fig. 4 illustrates the distributions of coordination numbers or packing densities around each amino acid type. Nonpolar residues such as Ile and Val, the red curves, show a clear preference for a greater packing density (both Ile and Val had 21 residues within the contact shell on average). Overall, the mean coordination numbers agreed with the Kyte-Doolittle ranking scale with a correlation coefficient $R^2 = 0.75$, as depicted in Fig. 5. The main exceptions to this correlation were Trp and Tyr, which appeared more hydrophobic on the SPARC scale than they are conventionally described, and three of the charged amino acids (Lys, Asp, and Glu), which appeared more hydrophilic. Regardless of the relative magnitudes of the means, the distributions proved useful in the simulation because the energy with respect to polarity of a specific residue could be easily approximated according to the coordination number distribution for its sidechain type.

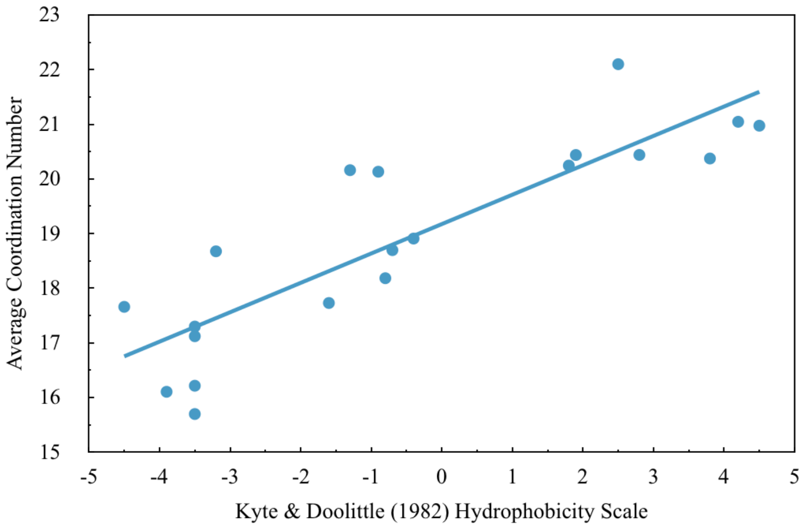

**Figure 5** – A comparison of the Kyte and Doolittle hydrophobicity scale and the SPARC coordination number approach illustrates the correlation ($R^2 = 0.75$) between the packing density of residues within the structure and their polarity.

## 3.2 Validation with CHARMM

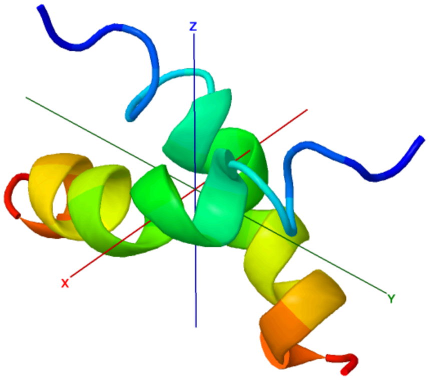

**Figure 6** – The structure of magainin, a small antimicrobial protein. The distance between the two chains is varied to compare the scores yielded by CHARMM and SPARC.

The correlation between knowledge-based potentials (KBPs) and all-atom force fields such as CHARMM is often tenuous because of the thermodynamic assumptions that are made on the native ensemble [24], although some reports have demonstrated good correlations between the force fields and their potentials [21, 3]. While SPARC is constructed on similar principles to many other KBPs, we hypothesized that because of the combination of the orientation-based residue interaction model and a solvent interaction term, it would correlate closely with the detailed all-atom potentials. This was tested according to the procedure used by Basdevant et al. [3], which is based on the structure of the small antimicrobial protein magainin (PDB code 1DUM), consisting of two $\alpha$-helices in anti-parallel as shown in Fig. 6. The B-chain was translated incrementally in both the $xy$ and the $yz$ directions; for each distance, the SPARC inter-residue scores were calculated as well as the all-atom potentials using the CHARMM27 force field *in vacuo* in the GROMACS command-line tool [4]. The structures were relaxed first using a

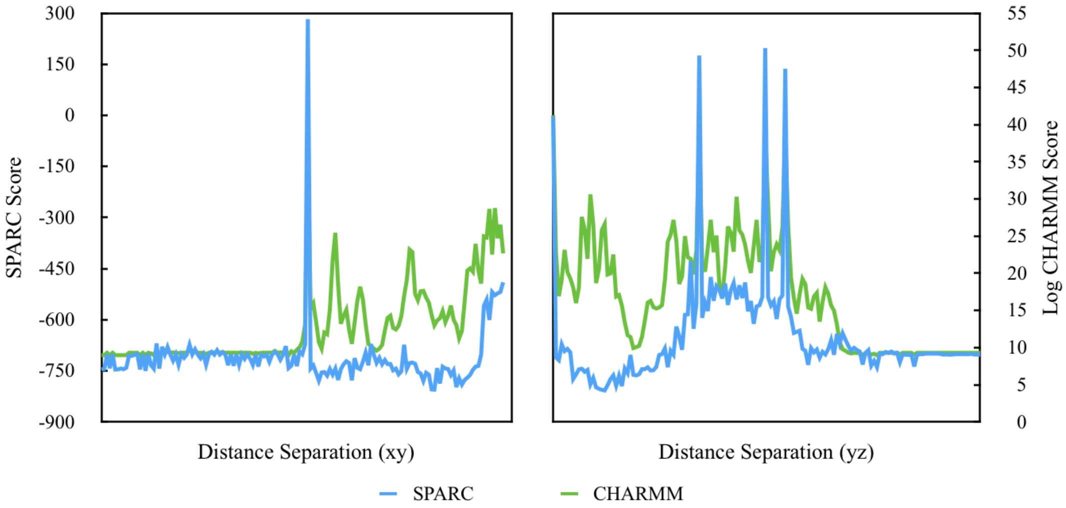

**Figure 7** – The graphs for various translations of the two helices of the magainin structure. Notably, SPARC mimics the local contours of the potential curve even though the absolute energy values do not show a strong correlation. (The distance values are not shown because they are irregular, due to the energy minimization performed before evaluation.)

short MD simulation to prevent extremely high CHARMM energy values.

The comparison revealed an interesting relationship between the CHARMM and SPARC potential values. As shown in Fig. 7, the SPARC score mimics the contours of the logarithm of the physics-based potential, especially in the $xy$-direction. The large jump in the $xy$ graph, which occurs at around 15 Å, is probably due to a steric clash and is exactly captured by SPARC; other blips are also evident in both curves. However, the energy values do not strongly correlate with each other without the inclusion of the distance factor ($R^2 = 0.29$). This could be due to a lack of resolution in unstable structures, which obviously were not prevalent in the native ensemble used to construct our potential; SPARC would therefore be most useful for local comparisons between similar structures.

## 3.3 Evaluation with gapless threading

The most common way to assess the discriminating ability of a statistical potential is through its performance on the gapless threading problem: to identify the most stable conformation out of an ensemble of "decoy" structures onto which the same amino acid sequence is threaded. These databases vary considerably in difficulty, so SPARC was tested on decoy sets from the Decoys 'R' Us database [27] as well as the I-TASSER Decoy Set II [35]. The decoy sets vary in size as well;

| Decoy Set | # Decoys | DFIRE | GOAP | SPARC |
|---|---|---|---|---|
| 4state_reduced | 7 | 6 (-3.48) | 7 (-4.38) | 6 (-4.06) |
| fisa_casp3 | 5 | 4 (-4.8) | 5 (-5.27) | 3 (-3.60) |
| lmsd | 10 | 7 (-0.88) | 7 (-4.07) | 5 (0.52) |
| lattice_ssfit | 8 | 8 (-9.44) | 8 (-8.38) | 7 (-4.04) |
| hg_structal | 29 | 12 (-1.97) | 22 (-2.73) | 18 (-1.74) |
| ig_structal | 61 | 0 (0.92) | 47 (-1.62) | 33 (-1.60) |
| I-TASSER | 56 | 49 (-4.02) | 45 (-5.36) | 48 (-7.44) |
| Total | 176 | 86 | 141 | 120 |

**Table 1** – SPARC was tested on the gapless threading problem, which entails predicting the native structure out of an ensemble of decoys "threaded" with the same structure. The decoy sets vary in size and difficulty, but overall SPARC was superior in accuracy to the distance-dependent DFIRE potential (data obtained from [36]) and slightly inferior to GOAP, an all-atom orientation-based potential.

is the largest with _ decoys per structure on average.

In Table 1, we present the results of this test alongside the published statistics of other statistical potentials. GOAP [36] and Miyazawa & Jernigan [20] are anisotropic potentials, using conceptually similar approaches to SPARC but at a more fine-grained level. (The potential developed by Buchete et al. [7, 9] is coarse-grained like SPARC, but the numerical accuracies on these decoy sets were not given in the original publication.) For comparison with non-orientational methods we include DFIRE, an all-atom distance-dependent potential that utilizes a distance-scaled ideal gas reference state.

Because the energylike quantities obtained from Eqs. (2) and (3) cannot necessarily be added together, a weighting procedure was performed to determine the optimum linear combination of the three energy terms (consecutive, nonconsecutive, and solvent). The integer weights that yielded the highest number of correct native structures was 4 for consecutive, 9 for nonconsecutive, and 3 for solvent interactions.

In addition to the standard decoy tests, the CASP11 decoy set was used to test the correlation between the root-mean-square deviation (RMSD) from the native structure and the SPARC energy score. RMSD is calculated from two protein structures $\mathbf{x}$ and $\mathbf{y}$ as $\sqrt{\frac{1}{n}\sum_{i=1}^{n}||\mathbf{x}_i - \mathbf{y}_i||^2}$ over the $\alpha$-carbons in each structure, and serves as a measure of the difference between the structures. In Fig. 8, two characteristic RMSD graphs are shown for the CASP11 decoys.

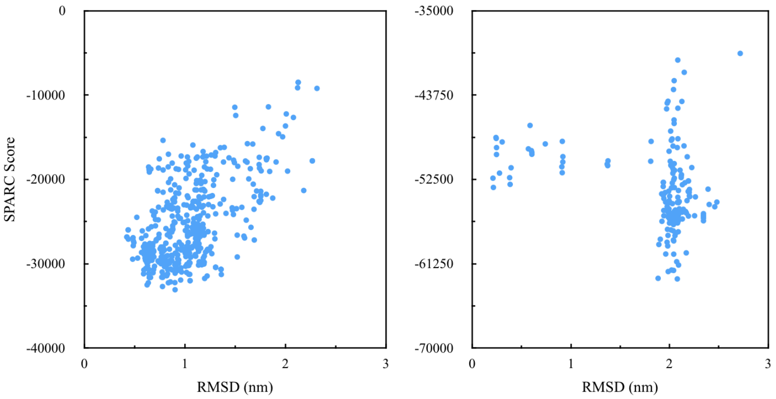

**Figure 8** – Two illustrative root-mean-square deviation (RMSD) graphs from the CASP11 set. In the left graph (PDB 2MQ8), there is a clear correlation between the RMSD and the energy score. However, in the right graph (PDB 4Q5T), a conformation cluster at around RMSD = 2 nm appears more stable than the native structure. See Discussion.

| Segment Length | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Mean | -93.38 | -93.52 | -93.84 | -89.68 | -90.08 |
| Minimum | -136.60 | -123.77 | -129.88 | -132.45 | -140.87 |
| Variance | 159.57 | 174.75 | 209.01 | 499.12 | 303.67 |

**Table 2** – Statistics on the energy scores given by SPARC over 500 iterations of the simulation algorithm at various segment lengths. Until $\ell = 4$, the variance monotonically increases, suggesting that grouping amino acids together leads to more erratic simulation.

### 3.4 Simulation of a small globular protein

Using the weighted potential terms from the decoy set results, we apply SPARC to *ab initio* protein structure prediction. Bovine pancreatic trypsin inhibitor (BPTI, PDB code 1QLQ), a 58-amino acid protein, was chosen for this study because of its simple structure (two $\alpha$-helices and a pair of antiparallel $\beta$-strands) and because it is conventional for testing protein structure prediction algorithms.

First, the effect of manipulating the segment length $\ell$ was tested on the BPTI structure by running 500 iterations at segment lengths from 1–5. The average, minimum, and variance of the energy scores are given in Table 2. The mean energy over the course of the simulation increases sharply at $\ell = 4$, and while the minimum is not affected, the variance shows a stark increase with $\ell$ until 5.

According with this increase in variance with increasing segment length, a random selector was designed to choose a certain value of $\ell$ with a weight proportional to

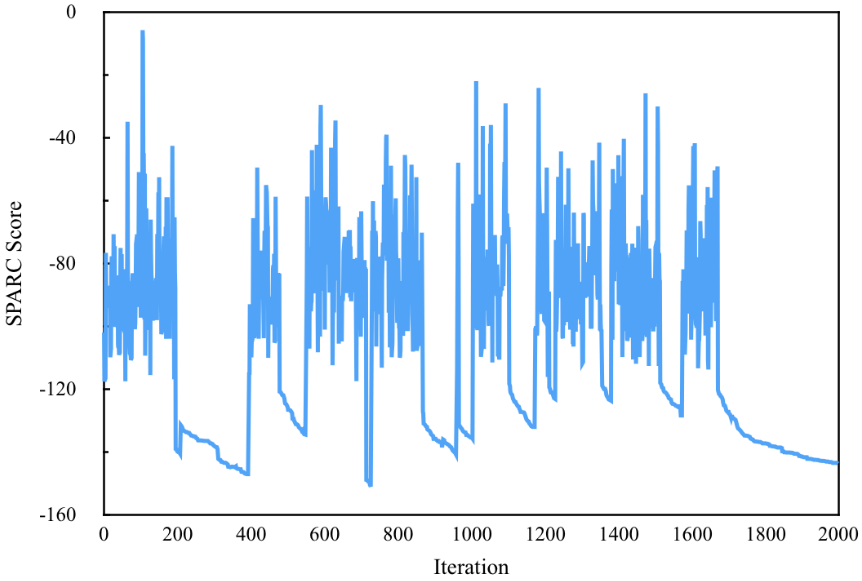

**Figure 9** – SPARC energy scores over the course of a short simulation of bovine pancreatic trypsin inhibitor (BPTI) folding. The steady, deep regions correspond to "gentle" mutations, while the rapidly fluctuating iterations are "erratic" and designed to introduce variability into the structure.

$$-\frac{1}{9}\left(x - e^{\frac{4}{225}S+2}\right)^2 + 4, \tag{5}$$

which favors slightly greater segment lengths when the potential energy score $S$ is high. In addition, the behavior of the simulation was split into two modes, dubbed "erratic" and "gentle;" the erratic mode helps prevent the simulation from being trapped in local energy minima, and the gentle mode refines the structure to cultivate more stable interactions.

After 2,000 iterations, the SPARC energy scores followed the curve shown in Fig. 9. The result closest to the native structure in terms of RMSD was found at the 1,278th iteration with a score of -106.4 and RMSD 8.78Å , compared to -140.9 for the native structure. As intended, the "erratic" and "gentle" modes produced markedly different energy patterns; gentle groups of iterations were relatively inflexible and tended to decrease the energy scores slightly, while erratic iterations were more chaotic (average variance 60.0 compared to 365.6).

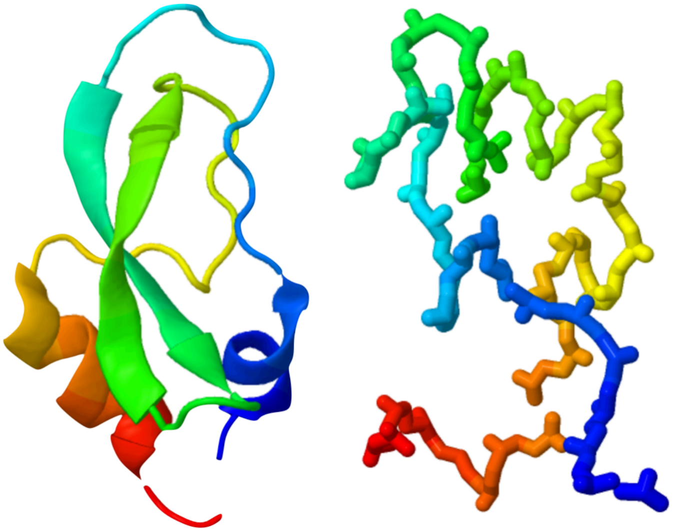

**Figure 10** – Comparison of the native structure of BPTI (left) with the closest structure (RMSD = 8.78Å) predicted by SPARC and the Monte Carlo simulation. There are clear differences in the predicted conformation, but the beginnings of secondary structure and the formation of contacts can be seen.

## 4 Discussion

Unlike other statistical potentials, SPARC utilizes a combination of orientational inter-residue interactions and coordination number-based solvent interactions to thoroughly describe the energy of a protein. When designing a coarse-grained approach to any algorithmic task, it is important to minimize the loss in accuracy over the improvement in performance. Current potentials range from all-atom [3] to solely contact-based [8], but SPARC uniquely captures the orientational stability without losing the performance of a coarse-grained method.

SPARC runs quickly once the distributions are calculated, but the construction of the potential itself was extremely time-consuming, taking roughly two weeks to compute all of the relative orientations in the 91,995-structure sample on the two machines. The additional time spent analyzing this dataset (which is an order of magnitude larger than the sets used for other potentials) improved the potential's resolution, since it better differentiated the low-frequency bins. The correlations between CHARMM and SPARC on the magainin test did show some lack of definition for CHARMM energies greater than [], but these extremely high energies would be rare in a simulation. One major change that would improve the accuracy of SPARC even further, at the expense of performance, would be to combine the relative orientations in each LCS as they occurred in the native ensemble.

As Eq. (2) shows, the frequencies from each residue's LCS are added independently, when in reality certain orientations are likely to be conditional to specific orientations of the other residue. We did not implement this change because it would effectively have squared the potential tables' memory requirement; however, it could be an area for extension in a future version.

The comparison of CHARMM and SPARC using the magainin protein illustrates a clear correspondence between the scores given by each potential. These simulations were performed *in vacuo*, so the CHARMM energies should strictly consist of bonded, electrostatic, and van der Waals interactions. Given the diversity of the terms which compose an all-atom potential, it is remarkable that the solely statistical SPARC model is able to approximate the contours of the CHARMM energy function. Also notable is that SPARC fits the *logarithm* of the physics-based potential; this was probably due to the extreme outliers in the CHARMM scores when a steric clash occurred.

The test to identify the native structures in the gapless threading problem illustrates SPARC's strengths and weaknesses in comparison to similar knowledge-based methods. SPARC achieved an accuracy of 68%, compared to 48.9% by DFIRE and 80.1% by GOAP. DFIRE, a distance-dependent potential, was less effective than either of the two anisotropic potentials, confirming that distance alone is not sufficient to describe the energy of a protein. (The same was found in the test with magainin, where distances varied in different directions resulted in different energy curves.) GOAP, which is based on orientations of heavy atoms within each amino acid as well as atomic distances, has a higher accuracy than SPARC, most likely because of its all-atom components. Since SPARC does not consider the exact locations of atoms within each residue, only the VSEPR-derived Cartesian coordinates, it cannot distinguish between structures with stable sidechains and those with infeasible packing arrangements. For our potential to perform better on the gapless threading test, we could add additional terms to describe the sidechain arrangements. However, our Monte Carlo simulation algorithm, designed for SPARC, circumvents this issue by checking for steric violations at every iteration and allowing only permissible consecutive orientations.

Although the simulations we performed were relatively short, they are representative of the characteristic of the algorithm since the output does not depend on previous iterations. The energy scores fluctuated frequently because new, unstable structures were introduced by the mutation system. We did find that mutating longer segments led to more chaotic simulations and greater variance in energy scores, which is most likely a consequence of the restriction to permissible orientations. Since the mutation in each iteration was cascaded through the chain in either direction to maintain chain connectivity (see Fig. 2), longer segments led to larger deviations in each cascade.

The experiments with segment length, as well as other tests, allowed us to probabilistically control the introduction of variation into the structure. In the dynamically-adjusted simulation of BPTI, the energy scores become much steadier and the conformation is conserved to a greater extent when the algorithm recognizes it as being stable. However, because this simulation is not in physical time, it is appropriate that eventually the mutations interrupt the "gentle" iterations to experiment with other shapes. For instance, the native structure of BPTI gave a SPARC energy score of -140.9, and for 258 points in the run the score dropped below -140. However, the closest structure in terms of RMSD had a considerably higher score by SPARC, which suggests a lack of stabilizing contacts and appropriate secondary structures. The lowest-energy structures on the other hand, display tighter helices and even some $\beta$-sheets. We believe that in a longer simulation, even better conformations would appear and be refined, approaching the shape of the native structure and forming even better orientations than the experimentally-determined BPTI structure.

## 5 Conclusion

The statistical potential developed in this paper, Segmented Positional Analysis of Residue Contacts (SPARC), contains innovations in several aspects. It is based on relative orientations using Cartesian local coordinate system transformations, which describe the protein structure more thoroughly than contact-based or distance-based representations. On the other hand, SPARC is coarse-grained, which permits it better performance at a small price in accuracy, at least on the gapless threading problem. SPARC was shown to replicate the contours of the energy function given by a conventional all-atom physics-based force field, and the packing density criterion correlates with a well-known hydrophobicity scale. Furthermore, the development of an off-lattice simulation algorithm using a novel segmentation technique proved useful for generating realistic candidate conformations in less than two thousand iterations.

Using these simulation techniques and potential functions, biologists will be able to model a host of protein activities, including ligand docking, signal transduction, and nucleic acid manipulations[1]. In addition, this simulation algorithm could be be applied to protein design, which seeks to create

[1] Nucleic acids are not currently supported by SPARC, but could be easily added with a large enough dataset of nucleotide-amino acid interactions and a simple Python subclass.

protein sequences that have certain structural properties for drug design and gene therapy. In the future, we are considering expanding SPARC to take advantage of its flexible implicit solvent model, especially by utilizing it in the plasma membrane where the solvent varies through the environment. Through these adaptations, we hope to develop the orientational approach to protein structure prediction into a general research tool for modeling biomolecular systems.

## References

[1] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[2] D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science*, 294:93–96, 2001.

[3] N. Basdevant, D. Borgis, and T. Ha-Duong. A coarse-grained protein-protein potential derived from an all-atom force field. *J. Phys. Chem.*, 111(31):9390–9399, 2007.

[4] H. Berendsen, D. van der Spoel, and R. van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.*, 91(1–3):43–56, 1995.

[5] B. Brooks et al. CHARMM: The biomolecular simulation program. *J. Comput. Chem.*, 30(10):1545–1614, 2009.

[6] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4:187–217, 1983.

[7] N. Buchete, J. Straub, and D. Thirumalai. Orientation-dependent coarse-grained potentials derived by statistical analysis of molecular structural databases. *Polymer*, 45:597–608, 2004.

[8] N. Buchete, J. Straub, and D. Thirumalai. Dissecting contact potentials for proteins: Relative contributions of individual amino acids. *Proteins*, 70(1):119–130, 2007.

[9] N. Buchete, J. Straub, and D. Thirumalai. Anisotropic coarse-grained statistical potentials improve the ability to identify native-like protein structures. *arXiv*, 2008.

[10] D. Case et al. Amber 2015. Technical report, University of California, San Francisco, 2015.

[11] M. Enciso and A. Rey. Improvement of structure-based potentials for protein folding by native and nonnative hydrogen bonds. *Biophys. J.*, 101:1474–1482, 2011.

[12] M. Feig and C. L. Brooks. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Opin. Struc. Biol.*, 14:217–224, 2004.

[13] R. Gillespie. The valence-shell electron-pair repulsion (VSEPR) theory of directed valency. *J. Chem. Educ.*, 40:295–301, 1963.

[14] A. Kolinski and J. Skolnick. Monte Carlo simulations of protein folding. i. lattice model and interaction scheme. *Proteins*, 18:338–352, 1994.

[15] M. Kouza, N. T. Co, P. H. Nguyen, A. Kolinski, and M. S. Li. Preformed template fluctuations promote fibril formation: Insights from lattice and all-atom models. *J. Chem. Phys.*, 142, 2015.

[16] J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157:105–132, 1982.

[17] H. Lu and J. Skolnick. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins*, 44:223–232, 2001.

[18] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.

[19] S. Miyazawa and R. L. Jernigan. Estimation of effective inter-residue contact energies in protein crystal structures: Quasi-chemical approximation. *Macromolecules*, 18, 1985.

[20] S. Miyazawa and R. L. Jernigan. How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins? *J. Chem. Phys.*, 122(2), 2005.

[21] D. Mohanty, B. N. Dominy, A. Kolinski, C. L. B. III, and J. Skolnick. Correlation between knowledge-based and detailed atomic potentials: Application to the unfolding of the GCN4 leucine zipper. *Proteins*, 35:447–452, 1999.

[22] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink. The MARTINI coarse-grained force field: Extension to proteins. *J. Chem. Theory Comput.*, 4, 2008.

[23] A. Mukherjee, P. Bhimalapuram, and B. Bagchi. Orientation-dependent potential of mean force for protein folding. *J. Chem. Phys.*, 123(1), 2005.

[24] J. W. Mullinax, W. G. Noid, and H. A. Scheraga. Recovering physical potentials from a model protein databank. *P. Natl. Acad. Sci. USA*, 107(46):19867–19872, 2010.

[25] A. Onufriev. Implicit solvent models in molecular dynamics simulations: A brief overview. In R. A. Wheeler and D. C. Spellmeyer, editors, *Annual Reports in Computational Chemistry, Volume 4*. Elsevier, Amsterdam, 2008.

[26] R. P. F. Pontiggia and C. Micheletti. Coarse-grained description of protein internal dynamics: An optimal strategy for decomposing proteins in rigid subunits. *Biophys. J.*, 96:4993–5002, 2009.

[27] S. R and L. M. Decoys 'R' Us: A database of incorrect protein conformations to improve protein structure prediction. *Prot. Sci.*, 9:1399–1401, 2000.

[28] B. Roux and T. Simonson. Implicit solvent models. *Biophys. Chem.*, 78:1–20, 1999.

[29] W. R. P. Scott et al. The GROMOS biomolecular simulation program package. *J. Phys. Chem.*, 103(19):3596–3607, 1999.

[30] M.-Y. Shen and A. Sali. Statistical potential for assessment and prediction of protein structures. *Prot. Sci.*, 15:2507–2524, 2006.

[31] M. J. Sippl. Calculation of conformational ensembles from potential of mean force: An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, 213:859–883, 1990.

[32] S. Tanaka and H. A. Scheraga. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, 9(6):945–950, 1976.

[33] P. D. Thomas and K. A. Dill. An iterative method for extracting energy-like quantities from protein structures. *P. Natl. Acad. Sci. USA*, 93(21):11628–11633, 1996.

[34] J. Zhang and Y. Zhang. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PLOS One*, 5(10), 2010.

[35] Y. Zhang and J. Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57:702–710, 2004.

[36] H. Zhou and J. Skolnick. Goap: A generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys. J.*, 101(8):2043–2052, 2011.

[37] H. Zhou and Y. Zhou. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Prot. Sci.*, 11(11):2714–2726, 2002.