JCTC Journal of Chemical Theory and Computation

Transferable Coarse Grain Nonbonded Interaction Model for Amino Acids

Russell DeVane,[†] Wataru Shinoda,[‡] Preston B. Moore,[§] and Michael L. Klein*[,†]

Center for Molecular Modeling and Department of Chemistry, University of Pennsylvania, 231 South 34th Street, Philadelphia, Pennsylvania 19104-6323, Research Institute for Computational Sciences (RICS), National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba Central 2, Umezono 1-1-1, Tsukuba, Ibaraki 305-8568, Japan, and Department of Chemistry and Biochemistry, University of the Sciences in Philadelphia, 600 S. 43rd Street, Philadelphia, Pennsylvania 19104

Received October 17, 2008

Abstract: The large quantity of protein sequences being generated from genomic data has greatly outpaced the throughput of experimental protein structure determining methods and consequently brought urgency to the need for accurate protein structure prediction tools. Reduced resolution, or coarse grained (CG) models, have become a mainstay in computational protein structure prediction performing among the best tools available. The quest for high quality generalized CG models presents an extremely challenging yet popular endeavor. To this point, a CG based interaction potential is presented here for the naturally occurring amino acids. In the present approach, three to four heavy atoms and associated hydrogens are condensed into a single CG site. The parametrization of the site–site interaction potential relies on experimental data thus providing a novel approach that is neither based on all-atom (AA) simulations nor experimental protein structural data. Specifically, intermolecular potentials, which are based on Lennard-Jones (LJ) style functional forms, are parametrized using thermodynamic data including surface tension and density. Using this approach, an amino acid potential data set has been developed for use in modeling peptides and proteins. The potential is evaluated here by comparing the solvent accessible surface area (SASA) to AA representations and ranking of protein decoy data sets provided by Decoys 'R' Us. The model is shown to perform very well compared to other existing prediction models for these properties.

I. Introduction

With the rate at which genomic data have become available over the last several years, the need to accurately predict protein structures from the amino acid sequence has become paramount. Protein structure prediction has employed many approaches to attempt this feat with a few noted here.[1−16] Many of these models are based on a reduced resolution or coarse grained (CG) representation of the protein.[1−3,5,17−30] Such models are used because they provide a means to expand the capabilities of existing computational resources. However, this enhanced efficiency comes necessarily at the cost of reduced resolution in the description of the system since typically the details of several atoms are essentially averaged into a single CG interaction site. When this reduction of resolution is introduced, it is crucial to accurately capture the phenomenon of interest for the specific system under investigation or the resulting CG description is simply a toy model capable of providing only generic behavior.

The history of CG models for amino acids dates back over three decades at least, with the pioneering work of Levitt

* Corresponding author e-mail: klein@lrsm.upenn.edu.

[†] University of Pennsylvania.

[‡] National Institute of Advanced Industrial Science and Technology (AIST).

[§] University of the Sciences in Philadelphia.

10.1021/ct800441u CCC: $40.75 © 2009 American Chemical Society
Published on Web 07/17/2009

and Warshel, where they used reduced resolution models to investigate protein dynamics.[1] Since then, much work has been put forth to develop reduced resolution models to overcome the computational barriers of studying proteins. Much of this work has been built on knowledge based models which take advantage of the existing structural data acquired from experimental techniques including X-ray crystallography and NMR spectroscopy. Tanaka and Scheraga were the first to develop knowledge based contact potentials from the frequency of contacts between residues in known structures.[2] This approach was then extended by many groups to include for example distance and direction dependence.[3,5,17−20] Others such as Huang et al. have broken from this paradigm to use an approach that does not rely on known protein structural data other than the general nature of the packing of hydrophobic and hydrophilic residues in protein structures.[9,31]

Further, the use of CG protein models for molecular dynamics (MD) simulations has helped extend the capabilities of current computational resources allowing access to much larger temporal and spatial scales.[32−47] In particular, the MARTINI model has become a widely used CG model with an array of applications to protein and membrane simulations.[33−35] The later approach employees a LJ potential for nonbonded interactions with parameters in a tabulated form based on the interaction types (hydrophobic/hydrophobic, hydrophilic/hydrophobic, etc.). Many other groups have pushed CG models to a still higher level with approaches such as force matching[38−40] and other schemes which have enjoyed success.[41−53]

Herein we present the application of a recently proposed methodology to the development of CG nonbonded interaction potentials for amino acids.[44,48−50] The current model is essentially a distance dependent potential which distinguishes the 20 naturally occurring residues to a high level such that only a few residues are modeled with the same parameters. Molecular simulations are used to parametrize CG sites that are used here in a nondynamical representation of proteins. Experimental thermodynamic data are used as the target for parametrization of the CG site interactions. This approach builds on previous work by Nielsen et al.[52] who used a similar approach to obtain CG parameters for a series of alkanes and Shinoda et al.[49,50] who extended the CG parameters to polyethylene glycol surfactants. Related approaches have been reported recently but with notable differences.[54,55] The work by Basdevant et al.[54] uses a hybrid $1/r^6$ repulsive term and Gaussian attractive term that is parametrized using AA force fields. The work by Han et al.[55] uses a thermodynamic based parametrization approach but with a higher resolution mapping.

Herein, the current model is used to predict the native structures from several protein decoy sets with a level of success on par with knowledge based potentials. Finally, although not explicitly demonstrated here, the nonbonded CG model presented could be coupled with an appropriate intramolecular force field and used for MD simulations of peptides and proteins in a spirit similar to the MARTINI force field. Indeed, an initial application of the present approach was previously utilized in an MD simulation investigation of peptide nanoring self-assembly.[32] In Section II, the methods are discussed including the force field details and parametrization. Section III presents results and discussion including evaluation of the model through solvent accessible surface area (SASA) calculations and the ranking of protein decoys. Finally, Section IV closes with the conclusions.

II. Methods

A. Coarse Grain Model. Mainly for ease of implementation, the present CG model employs LJ style nonbonded potential functions.[49] For the models developed herein, all CG beads interact via a LJ(9-6) potential given in eq 1. The CG sites are charged, and this is also included in the electrostatic contribution given by eq 2

$$v(r)_{9-6} = \frac{27}{4}\epsilon\left(\frac{\sigma^9}{r^9} - \frac{\sigma^6}{r^6}\right) \tag{1}$$

$$v(r)_{elec} = \frac{1}{4\pi\epsilon_0}\frac{q_1q_2}{r} \tag{2}$$

The choice of prefactor for the LJ function is selected such that $v(\sigma) = 0$ and ϵ is the minimum energy. The choice of the LJ functional form (for example the 9−6 in eq 1) is essentially an adjustable parameter used in the fitting procedure. We have previously explored various options including 6−4, 8−4, 10−4, 9−6, and 12−4.[49] For alkane interactions the choice of the 9−6 functional form was validated by comparison of the CG liquid structure to AA MD results. The adoption of the 9−6 functional form reflects a desire to maintain consistency with existing CG models.[49,50] A simple truncation, implemented at a distance of $rcut$ = 15 Å, is employed for the long-range cutoff, with no smoothing or shifting. This length is sufficient to avoid gross artifacts resulting from the discontinuity; however, the cutoff length, $rcut$, does affect the thermodynamic properties and so is treated as a CG fitting parameter. The parameters in the LJ function are fixed by reproducing thermodynamic data. For bulk solutions, both ϵ and σ can be unambiguously fixed with a combination of density and surface tension. The cross interactions arising from nonidentical CG sites can be generated using the combining rules given by eqs 3 and 4

$$\sigma_{ab} = \frac{\sigma_{aa} + \sigma_{bb}}{2} \tag{3}$$

$$\varepsilon_{ab} = \sqrt{\varepsilon_{aa}\varepsilon_{bb}} \tag{4}$$

Here, ϵ_{aa} and σ_{aa} represent the self-interaction values for ϵ and σ values, and ϵ_{AB} and σ_{AB} represent those for the cross interactions. Finally, the electrostatic interactions are calculated using an effective dielectric constant of 80 and without employing a cutoff for electrostatic interactions. The dielectric constant is given below in the form of scaled charges with the value of 0.1118.

Although the intramolecular force field is not discussed or used in detail here, it was necessary to employ bonds and angles in the development of the nonbonded interactions for PHE, TRP, and TYR (throughout the paper, the 3-letter amino acid labels will be used) side chains. For those cases, the intramolecular interactions are modeled via a harmonic potentials given by eqs 5 and 6

$$V(r)_{bond} = K_b(r - r_0)^2 \tag{5}$$

$$V(r)_{angle} = K_a(\theta - \theta_0)^2 \tag{6}$$

where r_0 and k_b represent the equilibrium bond length and force constant for bonds, and θ_0 and k_a represent the equilibrium angle value and force constant for bends.

B. Molecular Dynamics Simulations. CG MD simulations were performed using the LAMMPS code developed at Sandia National Laboratory and extended by our group (now a part of the standard LAMMPS release) to implement our CG models.[56] An integration time step of up to at least 25 fs can be used to evaluate the nonbonded interactions. Intramolecular degrees of freedom have to be dealt with on a case by case basis as the force constant will determine the time step necessary. For these cases, the multitime step integrator rRESPA can be used to separate the various degrees of freedom.[57]

Electrostatic interactions were calculated using the Particle−Particle Particle-Mesh method implemented in the LAMMPS MD code.[58−61] Nonbonded interactions are excluded for bonded CG sites separated by two or less bonds (1−2, 1−3). Interactions between CG sites separated by 3 or more bonds are included without scaling. For the sake of evaluating these exclusions, the connectivity (bonds) of the model was considered although the bond energy was not calculated in the potential evaluation. Only the nonbonded van der Waals and electrostatic energy were used. The bulk system MD simulations included roughly 500 CG sites each. For surface tension simulations, roughly 50 ns simulation time was used. Systems were set up by equilibrating with an isothermal−isobaric (NPT) simulation. The equilibrated system box dimension was then extended in the z-direction creating a vacuum region large enough such that no interactions extended through the vacuum via the periodic boundary conditions to the opposite side. The extended system was then used in a MD run in a canonical (NVT) simulation. The surface tension, γ, was then calculated via

$$\gamma = \frac{L_z}{2}\left[P_{zz} - \frac{(P_{xx} + P_{yy})}{2}\right] \tag{7}$$

where L_z is the box dimension in the z-direction, and P_{aa} is the isotropic pressure tensor for each Cartesian direction where $a =$ x,y,z.[62] Density calculations were performed with NPT simulations averaged over roughly 2 ns simulation time. Finally, all MD simulations were performed at a temperature of 303.15 K, unless otherwise specified.

C. Solvent Accessible Surface Area. One property of interest in proteins is the SASA which is proportional to protein solvation free energy and represents an important property in protein research and design. In the work of Lee and Richards, it was shown that the hydrophobic residues tend to have a larger reduction of SASA when going from extended chains to the native conformation compared to the polar residues.[63] Succinctly, the SASA is the area of the surface traced out by rolling a probe sphere, representing the solvent, over the surface of the protein. Any area accessible to the probe that does not require the probe to overlap with neighboring atoms is considered solvent accessible. The SASA was calculated here via the implementation in VMD.[64] For the AA level calculations, a probe radius of 1.4 Å and VMD default atomic radii for the protein atoms were used. For the CG SASA calculations, a probe radius of 2.5 Å was used. This value was calculated from the minimum of the potential energy function, r, where $V(r)/r = 0$, for the CG water−water interaction. CG site radii for the amino acids were calculated in a similar fashion. Note that the minima of the LJ functional forms used herein are not equal to the standard LJ (12−6) minimum of $r = 2^{1/6}\sigma$ but are given by $r = 1.5^{1/3}\sigma$ and $r = 3^{1/8}\sigma$ for the LJ (9−6) and LJ (12−4), respectively.

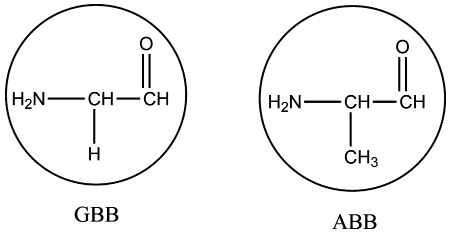

Figure 1. Shown here are the CG mapping for the backbone sites. The standard backbone CG site (GBB) is shown on the left, and the ALA CG site (ABB) is shown on the right.

D. Parameterization. In the CG model presented here, the mapping includes roughly three to four heavy atoms and adjacent hydrogens per CG site. Each amino acid is divided into a backbone and side chain section. The mappings for all of the naturally occurring amino acids are given in Figures 1, 2, and 3. All amino acids in this model (including GLY) share the same backbone representation with the exception of ALA. The standard backbone CG site includes the carbonyl carbon and oxygen, nitrogen, α carbon, and related hydrogens. GLY is of course represented by the single standard backbone CG site. Due to its small side chain, the ALA amino acid is mapped to a single site containing the backbone and side chain. The mappings for the backbone sites are shown in Figure 1 with the standard CG backbone site (labeled GBB in Table 3 used for all amino acids except ALA on the left) and the ALA CG model (labeled ABB in Table 3 given on the right.

Most side chains are represented by a single CG site. The exceptions to this are the PHE, TYR, LYS, and ARG side chains, which are modeled with two CG sites, and TRP which is modeled with three. Finally, Figures 2 and 3 show the mapping for all of the neutral and charged side chains, respectively.

The side chain analogue molecules used in the model development of the neutral amino acids are given in Table 1. To parametrize the LJ9-6 potentials for the self-interactions of the CG sites (SER-SER, LEU-LEU, etc.) the surface tension and density were reproduced in bulk solutions of the representative side chain analogue where possible. For example, this was not possible for the TRP side chain analogue 3-methylindole due to the complexity (too many parameters to fix simultaneously with only two observables)

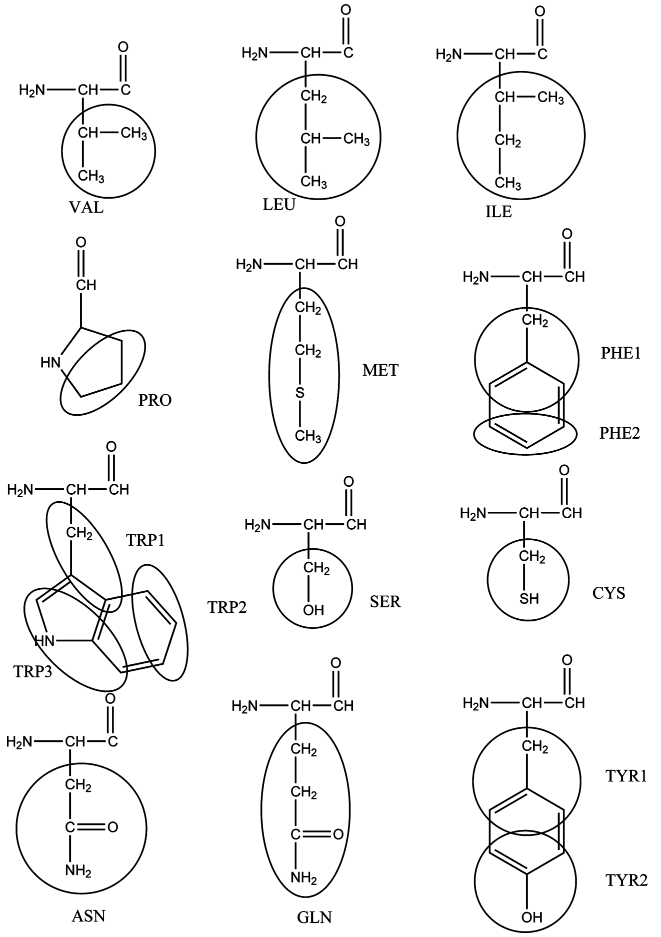

Figure 2. Shown are the mappings for the amino acid side chains. The label here corresponds to the labels used in the interaction potential Table 3.

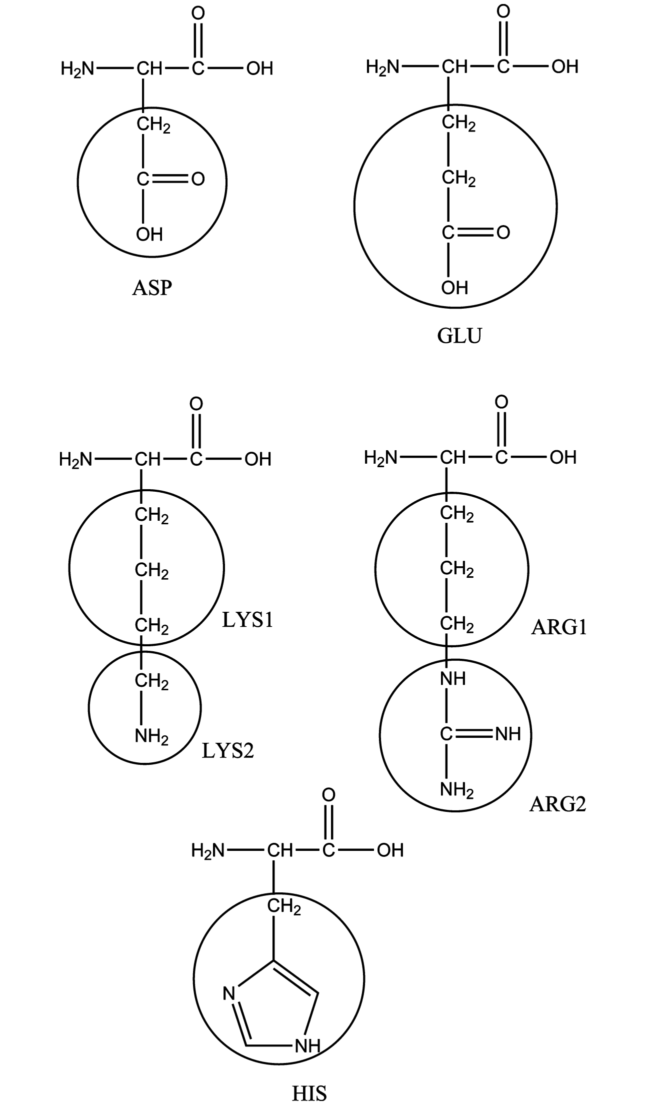

Figure 3. The CG mappings for the charged residues are shown here. ASP, GLU, LYS2, and ARG2 are all modeled with the same interaction parameters. LYS1 and LYS2 are modeled with the VAL CG site paramters. Note that HIS is included in this figure although it is not modeled with a charge in the current model.

Table 1. Side Chain Analogue Molecules

residue	analog	residue	analog	residue	analog
ALA	none	MET	methyl-ethyl sulfide	THR	ethanol
VAL	propane	PHE	toluene	CYS	methanethiol
LEU	isobutane	TRP	3-methylindole	ASN	acetamide
ILE	butane	GLY	none	GLN	propionamide
PRO	propane	SER	methanol	TYR	cresol

and the fact that it is a solid at room temperature with a melting point of around 370 K. Instead, the CG sites for this side chain were developed in a stepwise process (see below). On the other hand, having slightly higher melting point did not alter the approach for ASN and GLN side chains. A slightly higher temperature was used in the simulations for the parametrization of this CG sites. Strictly speaking, the CG models are parametrized at a specific temperature and thus not transferable to different temperatures; however, there is a small range of temperature which will not alter the properties to a large degree. Results for the surface tension and density of each side chain analogue is compared to experimental values in Table 2.[65] The parameters for the VAL side chain are from previous work.[49] The parameters for LEU and ILE were developed using the analogues 2,5-dimethylhexane and 3,4-dimethylhexane, respectively, since the true analogues are gases at room temperature. Equilibrium bond lengths of 3.87 Å and 3.22 Å with force constant of 5.0 kcal/mol were used for each, respectively.

For PHE, TYR, and TRP, it was necessary to implement a bond between the CG sites in order to parametrize the overall CG assembly. For PHE and TRP the chosen minimum energy bond lengths, 2.5 Å, and the force constant of 150 kcal/mol were used for all bonds. For TYR, the minimum energy bond length, 2.9 Å, and the force constant of 150 kcal/mol were used. The choice of a two-site CG model for PHE and TYR was used to maintain consistency with the mapping of roughly 3 to 4 heavy atoms per CG site. This is admittedly a gross overestimation of these side chains and the implications arising are discussed below. The PHE side chain was developed from the CG sites for p-xylene (XYL) and benzene (BEN). Symmetric two site models were made for each of these two molecules and parametrized to reproduce the respective surface tension and density. These sites were then combined to produce the PHE

Table 2. Thermodynamic Data from CG Simulations Compared with Experiment

	surface tension (mN/m)		density (g/cm^3)	
molecule	simulation	experiment	simulation	experiment
VAL (PRO) (hexane)	17.0±0.5	17.4	0.65	0.65
LEU (2,5-dimethylhexane)	19.0±0.5	18.8	0.69	0.69
ILE (3,4-dimethylhexane)	20.3±0.5	20.7	0.72	0.71
MET	24.0±0.5	23.2	0.83	0.83
PHE	26.7±0.5	27.3	0.86	0.86
SER	22.0±0.5	22.9	0.78	0.78
THR	22.7±0.5	22.7	0.78	0.78
CYS	22.0±0.5	23.0	0.86	0.86
ASN[a]	37.5±0.5	39.4	1.04	1.00
GLN[a]	30.7±0.5	30.6	0.95	0.93
TYR[b]	35.1±0.5	35.1	1.00	1.03

[a] 358.0 K. [b] 315.5 K.

Table 3

	LJ9-6 parameters		radii	
side chain	ϵ (kcal/mol)	σ (Å)	CG (Å)	PDB (Å)
VAL	0.469	4.59	2.63	2.35
LEU	0.536	5.06	2.90	2.6
ILE	0.487	5.18	2.96	2.6
PRO	0.469	4.59	2.63	2.34
MET	0.835	5.03	2.88	2.71
PH1	0.435	4.91	2.81	2.57
PH2	0.413	4.33	2.48	2.3
TR1	0.473	4.41	2.52	2.33
TR2	0.413	4.33	2.48	2.3
TR3	0.753	4.66	2.67	2.51
SER	0.580	3.68	2.11	1.96
THR	0.666	4.24	2.43	2.27
CYS	0.656	4.16 (2.40)	2.38	2.18
ASN	0.870	4.15	2.38	2.42
GLN	1.192	4.74	2.71	2.66
TYR1	0.435	4.91	2.81	2.57
TYR2	0.700	4.10	2.35	2.48
ASP	0.497	4.00	2.29	2.4
GLU	0.497	4.00	2.29	2.64
LYS1	0.469	4.59	2.63	2.33
LYS2	0.497	4.00	2.29	1.91
ARG1	0.469	4.59	2.63	2.34
ARG2	0.497	4.00	2.29	2.32
HIS	1.400	5.40	3.09	2.73
GBB	0.870	4.15	2.38	2.4
ABB	1.200	4.74	2.71	2.69

side chain model. Figure 4 shows the CG sites and corresponding AA representations (without hydrogens for clarity) used in the stepwise process. The blue and red spheres represent the BEN and XYL CG sites, respectively. In the PHE side chain CG model (far right of Figure 4), the XYL CG site, represented by the red sphere, is labeled PHE1, and the BEN CG site, represented by the blue sphere, is labeled PHE2 in the interaction database (see Figure 4). The combination rules, eqs 3 and 4, from the Methods section were used to develop the XYL-BEN cross interaction. The result of using this approach can be seen in the reproduction of the bulk properties for the PHE side chain model in Table 2. The XYL CG site was then combined with a new site to create the cresol (TYR side chain) CG model in a similar fashion. The labels in the TYR CG model are TYR1, modeled with the XYL CG site and TYR2 modeled with the new −OH containing cresol CG site (see Figure 4). Finally, the benzene CG site was used to develop a CG model for the TRP side chain.

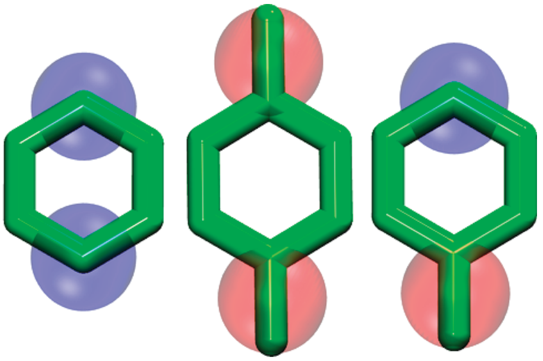

Figure 4. The models used to develop the phenyl based rings (PHE, TYR, and TRP) are shown here. The CG model for benzene (left and blue) and p-xylene (middle and red) were combined to produce the PHE CG model (right). These sites were also used in the TYR and TRP CG models as described in the text.

The analogue molecule 3-methylindole was used to represent the TRP side chain. The CG mapping for the TRP side chain is shown in Figure 2. The new CG sites TRP1 and TRP3 represent propene and allylamine, respectively. Each of these molecules is represented by a single distinct CG site that was parametrized in a similar fashion to the amino acids with side chains represented by single CG sites. In a stepwise process the TRP side chain was built up first by combining the BEN and ALL CG sites to create the CG model for aniline. The bond was modeled with the equilibrium bond length and force constant stated above. This process yielded the BEN-ALL cross interaction from the bulk aniline simulations. Then, the BEN and ALL beads were combined with the PPE CG site to create the TRP side chain CG model. The combination rules were employed to fix the interactions BEN-PPE and ALL-PPE.

For the charged residues (ASP, GLU, LYS, and ARG) a primitive model was developed including charges. Figure 3 shows the CG mapping for the charged residues. It should be noted that although HIS is included here, it is not modeled with a charge in the current model. The CG sites labeled ASP, GLU, LYS2, and ARG2 are identical in the current model. These sites were parametrized to reproduce the surface tension of guanidium salt solutions at various concentrations. A charge of 0.1118 (with $\epsilon_0 = 1$ in eq 2 scaled down from unity to compensate for the solvent screening that would take place normally but is absent due to the CG water model being represented by a single chargeless nonpolar spherical site) is included on the sites ASP, GLU, LYS2, and ARG2. All other sites are modeled without charge. The addition of a scaled charge is part of a recent development designed for modeling charged surfactants. Details will be part of a forthcoming manuscript. As stated above, the current model will not distinguish ASP from GLU or LYS from ARG. For LYS and ARG, the charged CG sites, labeled LYS2 and ARG2 (modeled with a charge), are combined with a CG site equivalent to the VAL side chain CG site (and modeled neutral) and labeled LYS1 and ARG1,

Table 4. Comparison of SASA for CG and AA Models[a]

PDB I.D.	CG (Å^2)	AA (Å^2)	ratio (CG/AA)
lysozyme (2vb1)	6479.4	6676.4	0.97
myoglobin (2jho)	8040.1	8222.4	0.98
ribonuclease (1cv9)	7085.9	6968.8	1.02
flavodoxin (1flv)	7539.8	7713.3	0.98

[a] The SASA was calculated for the CG and AA representations with a probe radius of 2.5 Å and 1.4 Å, respectively.

respectively in Figure 3, to create the two bead CG site for these side chains.

III. Results and Discussions

A. Surface Area. Evaluation of force fields for complex molecules such as proteins is a difficult task. Indeed it is an ongoing process where many systems must be explored in order to provide a good representation of possible interactions that can arise. Nonetheless, here we set out to provide at least a preliminary level of evaluation. For protein−protein interactions, steric effects arising from side chain packing can play an important role in determining the spatial arrangements. Therefore the relative sizes of amino acid side chains is a key property to preserve in any model of amino acids. Further, it is understood that CG modeling often distorts the representation of molecules especially when mapping to spherical sites. To evaluate the effects of this mapping, the SASA was calculated for each amino acid side chain and backbone unit from a set of PDB databank structures with less than 30% homology. For these calculations, each side chain or backbone unit was calculated as a solitary group of atoms (excluding all other atoms in the structure) to give a SASA for that residue side chain or backbone type (GLY or ALA) taking into account only the various conformations of each subunit. From these calculations, the radii were calculated for each subunit and are shown in Table 3. The results are compared to the radii calculated for the CG model side chain units as described in the Methods section. As shown in Table 3, the values compare well suggesting conservation of SASA with the mapping from AA to CG. It is important to note that the radii for the CG model were not based on all atom or PDB calculations but arise from the parametrization process as described above. To further evaluate the SASA predictive ability of the CG model, the SASA for full proteins were calculated. Table 4 shows the results of the SASA calculations for several proteins comparing the CG results to those from the AA level. The results show that despite the fact that the CG model uses a bulky solvent (combining 3 water molecules) the SASA is consistent in the CG and AA models. Finally, it should be mentioned that calculation of SASA is not straightforward, and many things, such as the probe radius, can affect the results. Measurements made here were done as consistently as possible using standard techniques and radii for the probe and atomic radii. Therefore, it is assumed that this is at least a reasonable comparison of the SASA.

B. Predicting Native Structures. The ranking of structures in decoy sets such as those provided by Decoys R Us is a useful test for evaluating amino acid potentials and has become a somewhat standard evaluation tool.[66] The decoy sets are composed of native protein structures with decoys for each native structure generated through various techniques. The primary use of the protein decoy sets is to test a model's ability to distinguish the native structure from the non-native decoy structures. Here we evaluated the current CG model using five decoy sets (4*state_reduced*, *fisa*, *fisa_casp*3, *lmds*, *lattice_ssfit*) provided by Decoys R Us (http//dd.stanford.edu).[14,66−69] For comparison, the MARTINI force field was also used to rank the decoy sets.[35] For each structure, the CG model was mapped onto the AA model by placing the CG sites at the center of mass of the representative heavy atoms. The nonbonded potential presented here (and the MARTINI nonbonded potential) was then used to calculate the potential energy of each structure. These energies were then ranked from lowest to highest. The observable is the ranking of the potential energy of the native structure compared to the decoys. A value of 1 indicates that the native structure was predicted to have the lowest energy (is the most stable) of the set. The higher the value the worse the ability of the model to identify the native structure from the decoys. The ranking of the native structures for each set is shown in Table 5. The results for the model presented here are shown in the columns labeled CG, and those for the MARTINI model are labeled MARTINI.[35]

During the evaluation of the decoy sets, one issue had to be addressed in an ad-hoc fashion. Disulfide bonds arising between CYS residues had to be treated as a special case. The parametrization the CYS side chain site used bulk methylmercaptan solutions leading to an effective size which is much larger than the center of mass distance that arises when the CYS residues are involved in a covalent bond. To address this issue, sigma for the CYS-CYS interaction was shortened to a value of 2.4 Å to represent the CYS disulfide bond. However, it is important to note that the cross terms (CYS-LEU, CYS-SER, etc.) generated with the combination rules used the original bulk solution sigma value of 4.16 Å. To test the effect of the CYS-CYS interactions, the structures were re-evaluated with all of the CYS-CYS interactions removed. The results did not show significant differences which indicates that the use of the modified sigma value is reasonable for the structures evaluated here. However, to make a more general potential, this issue will have to be dealt with and the CYS-CYS will need to be categorized as either covalent or noncovalent. Typically CYS-CYS interactions involved in a disulfide bond would be excluded from the nonbonded interactions; however, the current setup does not take into account these bonds and thus they had to be dealt with with a reduced sigma value. For the MARTINI model, the calculation was repeated with all of the CYS-CYS interactions turned off and showed no significant difference in the results.

The results in Table 5 show that the potential is extremely effective at picking the native structure from the decoy sets. The present model ranked the native structure as the most stable in 68% of the decoy sets and ranked the native structure as one of the top 3 most stable structures in 77% of the decoy sets. For the sake of comparison, other methods are shown including MARITINI (mentioned above and not

Table 5. Results of Ranking Native Structures for Decoy Sets from Decoys 'R' Us[a]

	DFIRE	Rosetta	ModPipe-Pair	ModPipe-Comb	DOPE	MARTINI	CG	Z-score	% NEG
				4state_reduced					
1ctf	1/3.86	1	1	1	1	2	1	1.99	99.2
1sn3	1/3.79	1	1	1	1	398	29	1.39	99.4
2cro	1/3.29	5	1	1	1	397	1	2.47	99.0
3icb	4/2.28	6	151	8	1	308	282	0.31	99.4
4pti	1/3.62	1	1	1	1	254	267	0.35	99.6
4rxn	1/3.33	1	6	1	1	613	227	0.46	99.6
				fisa					
1fc2	254/0.23	158	491	453	375	458	32	1.28	99.6
1hdd-C	1/4.50	90	293	135	1	131	3	1.95	98.2
2cro	1/6.33	26	11	19	1	58	1	4.22	98.8
4icb	1/6.91	1	196	167	1	482	1	4.20	95.4
				*fisa_casp*3					
1bg8-A	1/5.35	1068	1	282	1	1200	1	4.35	4.6
1bl0	1/4.50	960	4	86	1	972	1	NA	0.1
1eh2	NA	NA	NA	NA	NA	2413	2	0.31	0.2
1jwe	1/6.26	1177	1	6	1	1407	1	2.37	0.6
smd3	NA	NA	NA	NA	NA	1	1	1.86	0.5
				lmds					
1ctf	1/3.54	1	1	1	1	108	1	2.01	99.8
1dtk	1/2.62	9	4	1	1	197	154	−0.57	98.2
1fc2	501/−5.72	291	325	222	476	2	1	4.55	92.6
1igd	1/5.16	1	1	1	1	1	1	2.54	96.8
1shf-A	1/6.68	5	24	7	1	1	1	3.16	99.5
2cro	1/4.70	2	4	12	1	1	1	6.14	100
2ovo	1/3.21	29	5	2	1	346	34	1.18	100
				lattice_ssfit					
1beo	1/12.09	1	1	1	1	1602	1	2.97	53.9
1ctf	1/10.05	1	1	1	1	1	1	3.29	70.9
1dkt-A	1/6.87	1	1	1	1	6	3	2.88	10.2
1fca	1/7.18	1	1	1	1	8	1	2.42	90.8
1nkl	1/9.29	1	1	1	1	1086	1	3.04	26.2
1pgb	1/11.87	1	1	1	1	186	1	2.71	68.2
1trl-A	1/6.32	45	1	1	1	7	1	2.78	69.4
4icb	1/7.81	1	1	1	1	2	1	3.07	35.5
total correct	26	14	17	17	27	5	21	NA	NA

[a] The lower the value, the better the performance of the model. A value of "1" indicates that the model predicts the native structure to be the most stable of all of the decoy structures. The column labeled "MARTINI" shows the results from the MARTINI model.[35] The column heading "CG" represents the results from the current CG model. All other results were taken from Zhou and Zhou[73,74] and Shen and Sali.[13] The Z-scrore values and the percentage of structures reporting a negative potential energy value with the current model are shown under the "Z-score" and "%NEG" heading, respectively. For comparison, the ranking and Z-scores have been included for the "DFIRE" results (ranking/Z-score).

designed for this type of evaluation), DFIRE1, Rosetta, ModPipe-Pair, ModPipe-Comb, and DOPE.[14,67,70−75] The values for these methods were taken directly from the references of Zhou and Zhou[73,74] and Shen and Sali.[13] Another evaluation of the quality of ranking is the Z-score which indicates the models discriminatory ability. We define the Z-score as

$$Z = \frac{\langle E \rangle - E_N}{\sigma} \tag{8}$$

where E_N is the energy of the native structure, $<E>$ is the average energy of all structures in a given set (for example 4*state_reduced*: 1*ctf*), and σ is the standard deviation of the distribution of energy values for a given set. The Z-scores for each set are given in Table 5. To prevent the Z-scores from being skewed by structures with unreasonably high energies, any decoy structure with a positive potential energy value was not included in the Z-score calculation. The percentage of structures that had negative potential energy values as determined by the current model are listed in Table 5 under the heading "%NEG". It should be noted that the *fisa_casp*3 set had a high percentage of decoy structures with a positive potential energy according to the current model, and so the Z-scores from those sets must be judged accordingly. Further, the *lattice_ssfit* had a modest amount of structures evaluated with positive potential energies. Finally, all of the native structures had a negative potential energy. For the sets in which the native structure was ranked as the most stable structure, the average Z-score was 3.13. For comparison, Z-scores for the DFIRE model have been included in Table 5. The average Z-score from the DFIRE model for all of the systems tested here was 5.21.

The 4*state_reduced* structures 3icb, 4pti, 4rxn, and the 1dtk structure from the lmds set are ranked very poorly by the current model. However, closer examination of these structures indicates that a substantial portion of the bad contacts arise from interactions involving phenyl based residues. The problem results from the mapping of the PHE and TYR side chains to two site CG models. The planar nature of these side chains is critical for the spatial arrangements they occupy

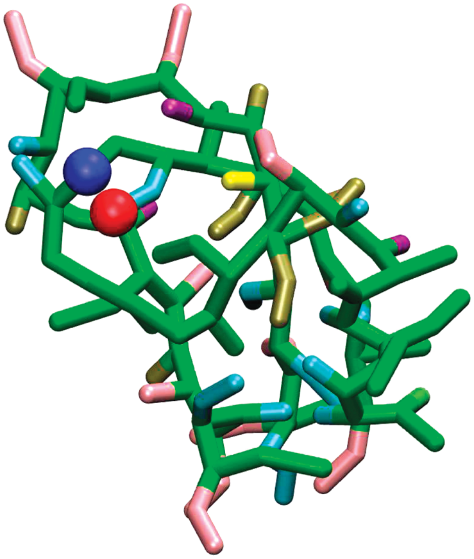

Figure 5. Shown is the CG representation of the 1dtk native structure from the lmds decoy set. The close contact is highlighted for the TYR2 site (red sphere) and the ALA site (blue sphere).

and must be conserved as much as possible in the CG model. The CG mapped structure for 1dtk native conformation is shown in Figure 5. In this figure, the TYR1 side chain site is shown as a red sphere and the ALA backbone site is shown as a blue sphere. It is easy to see here how the two site spherical representations of the phenyl based side chains leads to cases where bad contacts are generated by the CG model. This arrangement is not present in all decoys, and indeed the decoys with the lowest energy values do not have this conformation in which the TYR residue is in close proximity to other sites. The choice of a two site mapping for these residues was made to stay consistent with the mapping of roughly 3 to 4 heavy atoms. However, this issue will need to be addressed in future refinement of the model.

As mentioned above, the electrostatic residues were crudely parametrized. The current parameter set is unable to distinguish the residues ASP from GLU or LYS from ARG. This is demonstrated by the fact that the electrostatics make only small improvements to the results (4*state_reduced*-1ctf went from the number 3 ranked structure to number 1; lmds-1igd went from the number 2 ranked structure to number 1; others saw small or no improvement with no structures showing detriment with inclusion of electrostatics). However, it is possible that the structures studied here have conformations that are not largely dependent on electrostatic effects. Further, the percentage of charged residues is typically low in a protein, so a small impact on results is not totally unexpected.

Finally, the backbone atoms of proteins pose a particularly difficult problem as the directionality of hydrogen bonding makes the backbone an extremely anisotropic collection of atoms. This problem can be seen in the decoy sets evaluated here where high energy contacts are formed upon mapping to a CG representation. In these structures, the all atom representation shows hydrogen bonding interactions between the backbone sites leading to a stabilization of that conformation. Further, it should be mentioned that in the model used here, the backbone was not specifically parametrized but took advantage of the CG sites developed for the ASN and GLN side chains. This is a good first approximation to the interactions of the standard backbone and ALA backbone interaction sites; however, there is room for improvement of the interaction parameters of these sites. Also, although a nonpolar spherical CG site is a crude approximation to the backbone, the results shown here suggest that the approximation can perform reasonably well. Nonetheless, improvements to the performance of the backbone CG sites could possibly begin by addition of a dipole term such as that recently employed by Cascella et al.[76]

IV. Conclusions

A systematic approach has been used to develop a CG nonbonded interaction potential for amino acids. The model uses a reduced representation in which roughly 3 to 4 heavy atoms (non-hydrogen) and adjacent hydrogens are mapped to a single CG. The conservation of size in the model is reassuring for the protein−protein interaction predictive ability and for the future possibility of inclusion of a solvation contribution to the fold stability. The work here also demonstrates the consistency that can be maintained with parametrization techniques that are not based directly on protein structures. Further, it should be reiterated that no protein structural data or preconceived protein structural characteristics were used in the parametrization of this model. The encouraging results shown in the ranking of the decoy sets suggest that the current approach can be useful as a protein structure predictive tool. The decoy analysis results also highlights the importance of accurately treating the morphology of side chains especially the phenyl based residues (PHE, TRP, and TYR). As stated above, although the model is not presented here as suitable for MD simulations, there is nothing to prevent its implementation into a MD framework. Work on an intramolecular potential is currently in progress and will be presented in a forthcoming manuscript. In addition to this, further refinement of the nonbonded interactions will be a focal point of future work. The results presented herein suggest that the present approach is capable of yielding useful CG models.

Acknowledgment. The authors thank NSF, NIH, and Procter & Gamble for generous support. Also, the authors would like to thank Axel Kohlmeyer for implementation of the CG model into the LAMMPS MD code. Finally, thanks go to Vincenzo Carnevale for useful discussions. R.D. was supported by a NSF Biological Informatics Postdoctoral Fellowship under Grant No. DBI-0532800.

References

(1) Levitt, M.; Warshel, A. *Nature* **1975**, *253*, 694–698.

(2) Tanaka, S.; Scheraga, H. A. *Macromolecules* **1976**, *9*, 945–50.

(3) Miyazawa, S.; Jernigan, R. L. *Macromolecules* **1985**, *18*, 534–552.

(4) Treptow, W. L.; Barbosa, M. A. A.; Garcia, L. G.; de Araujo, A. F. P. *Proteins: Struct., Funct., Genet.* **2002**, *49*, 167–18.

(5) Sippl, M. J. *J. Mol. Biol.* **1990**, *213*, 859–883.

(6) Bahar, I.; Atilgan, A. R.; Erman, B. *Fold. Des,* **1997**, *2*, 173–181.

(7) Ueda, Y.; Taketomi, H.; Go, N. *Biopolymers* **1978**, *17*, 1531–1548.

(8) Dill, K. A.; Ozkan, S. B.; Shell, M. S.; Weikl, T. R. *Ann. Rev. Biophys.* **2008**, *37*, 289–316.

(9) Huang, E. S.; Subbiah, S.; Levitt, M. *J. Mol. Biol.* **1995**, *252*, 709–720.

(10) Schueler-Furman, O.; Wang, C.; Bradley, P.; Misura, K.; Baker, D. *Science* **2005**, *310*, 638–642.

(11) Moult, J. *Curr. Opin. Struct. Biol.* **2005**, *15*, 285–289.

(12) Zhang, Y. *Curr. Opin. Struct. Biol.* **2008**, *18*, 342–348.

(13) Shen, M. Y.; Sali, A. *Protein Sci.* **2006**, *15*, 2507–2524.

(14) Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. *J. Mol. Biol.* **1997**, *268*, 209–225.

(15) Reva, B. A.; Finkelstein, A. V.; Sanner, M.; Olson, A. J.; Skolnick, J. *Protein Eng.* **1997**, *10*, 1123–1130.

(16) Petrey, D.; Honig, B. *Mol. Cell* **2005**, *20*, 811–819.

(17) Miyazawa, S.; Jernigan, R. L. *J. Mol. Biol.* **1996**, *256*, 623–644.

(18) Miyazawa, S.; Jernigan, R. L. *Proteins: Struct., Funct., Genet.* **1999**, *36*, 357–369.

(19) Sippl, M. J. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 473–501.

(20) Sippl, M. J.; Weitckus, S. *Proteins: Struct., Funct., Genet.* **1992**, *13*, 258–271.

(21) Kolinski, A.; Skolnick, J. *Polymer* **2004**, *45*, 511–524.

(22) Clementi, C. *Curr. Opin. Struct. Biol.* **2008**, *18*, 10–15.

(23) Heath, A. P.; Kavraki, L. E.; Clementi, C. *Proteins: Struct., Funct., Bioinf.* **2007**, *68*, 646–661.

(24) Matysiak, S.; Clementi, C. *J. Mol. Biol.* **2006**, *363*, 297–308.

(25) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *Protein Sci.* **1993**, *2*, 1697–1714.

(26) Liwo, A.; Oldziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 849–873.

(27) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Oldziej, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 874–887.

(28) Liwo, A.; Kazmierkiewicz, R.; Czaplewski, C.; Groth, M.; Oldziej, S.; Wawak, R. J.; Rackovsky, S.; Pincus, M. R.; Scheraga, H. A. *J. Comput. Chem.* **1998**, *19*, 259–276.

(29) Yap, E. H.; Fawzi, N. L.; Head-Gordon, T. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 626–638.

(30) Khatun, J.; Khare, S. D.; Dokholyan, N. V. *J. Mol. Biol.* **2004**, *336*, 1223–1238.

(31) Huang, E. S.; Subbiah, S.; Tsai, J.; Levitt, M. *J. Mol. Biol.* **1996**, *257*, 716–725.

(32) Khurana, E.; DeVane, R.; Kohlmeyer, A.; Klein, M. L. *Nano Lett.* **2008**, *8*, 3626.

(33) Marrink, S. J.; de Vries, A. H.; Mark, A. E. *J. Phys. Chem. B* **2004**, *108*, 750–760.

(34) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. *J. Phys. Chem. B* **2007**, *111*, 7812–7824.

(35) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. *J. Chem. Theory Comput.* **2008**, *4*, 819–834.

(36) Bond, P. J.; Holyoake, J.; Ivetac, A.; Khalid, S.; Sansom, M. S. P. *J. Struct. Biol.* **2007**, *157*, 593–605.

(37) Treptow, W.; Marrink, S. J.; Tarek, M. *J. Phys. Chem. B* **2008**, *112*, 3277–3282.

(38) Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2005**, *109*, 2469–2473.

(39) Noid, W. G.; Chu, J. W.; Ayton, G. S.; Voth, G. A. *J. Phys. Chem. B* **2007**, *111*, 4116–4127.

(40) Zhou, J.; Thorpe, I. F.; Izvekov, S.; Voth, G. A. *Biophys. J.* **2007**, *92*, 4289–4303.

(41) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144–150.

(42) Tozzini, V.; McCammon, J. A. *Chem. Phys. Lett.* **2005**, *413*, 123–128.

(43) Masella, M.; Borgis, D.; Cuniasse, P. *J. Comput. Chem.* **2008**, *29*, 1707–1724.

(44) Klein, M. L.; Shinoda, W. *Science* **2008**, *321*, 798–800.

(45) Aksimentiev, A.; Schulten, K. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 4337–4338.

(46) Arkhipov, A.; Yin, Y.; Schulten, K. *Biophys. J.* **2008**, *95*, 2806–2821.

(47) Shih, A. Y.; Arkhipov, A.; Freddolino, P. L.; Schulten, K. *J. Phys. Chem. B* **2006**, *110*, 3674–3684.

(48) Shelley, J. C.; Shelley, M. Y.; Reeder, R. C.; Bandyopadhyay, S.; Klein, M. L. *J. Phys. Chem. B* **2001**, *105*, 4464–4470, 26.

(49) Shinoda, W.; DeVane, R.; Klein, M. L. *Mol. Simul.* **2006**, *33*, 27–36.

(50) Shinoda, W.; DeVane, R.; Klein, M. L. *Soft Matter* **2008**, *4*, 2454–2462.

(51) Bhargava, B. L.; DeVane, R.; Klein, M. L.; Balasubramanian, S. *Soft Matter* **2007**, *3*, 1395–1400.

(52) Nielsen, S. O.; Lopez, C. F.; Srinivas, G.; Klein, M. L. *J. Chem. Phys.* **2003**, *119*, 7043–7049.

(53) Shih, A. Y.; Freddolino, P. L.; Arkhipov, A.; Schulten, K. *J. Struct. Biol.* **2007**, *157*, 579–592.

(54) Basdevant, N.; Borgis, D.; Ha-Duong, T. *J. Phys. Chem. B* **2007**, *111*, 9390–9399.

(55) Han, W.; Wan, C. K.; Wu, Y. D. *J. Chem. Theory Comput.* **2008**, *4*, 1891–1901.

(56) Plimpton, S. *J. Comput. Phys.* **1995**, *117*, 1–19.

(57) Tuckerman, M. E.; Berne, B. J.; Martyna, G. J. *J. Chem. Phys.* **1991**, *94*, 6811–6815.

(58) Hockney, R.; Eastwood, J. *Computer Simulation Using Particles*; IOP: Bristol, 1988.

(59) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(60) Deserno, M.; Holm, C. *J. Chem. Phys.* **1998**, *109*, 7678–7693.

(61) Deserno, M.; Holm, C. *J. Chem. Phys.* **1998**, *109*, 7694–7701.

(62) Allen, M.; Tildesley, D. *Computer Simulation of Liquids*; Oxford Science: 1987.

(63) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379.

(64) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33.

(65) L., Y. C. *Chemical Properties Handbook*; McGraw-Hill: 1999.

(66) Samudrala, R.; Levitt, M. *Protein Sci.* **2000**, *9*, 1399–1401.

(67) Simons, K. T.; Ruczinski, I.; Kooperberg, C.; Fox, B. A.; Bystroff, C.; Baker, D. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 82–95.

(68) Keasar, C.; Levitt, M. *J. Mol. Biol.* **2003**, *329*, 159–174.

(69) Xia, Y.; Huang, E. S.; Levitt, M.; Samudrala, R. *J. Mol. Biol.* **2000**, *300*, 171–185.

(70) Melo, F.; Sanchez, R.; Sali, A. *Protein Sci.* **2002**, *11*, 430–448.

(71) Zhang, C.; Liu, S.; Zhou, H. Y.; Zhou, Y. Q. *Protein Sci.* **2004**, *13*, 400–411.

(72) Zhang, C.; Liu, S.; Zhou, H. Y.; Zhou, Y. Q. *Biophys. J.* **2004**, *86*, 3349–3358.

(73) Zhou, H. Y.; Zhou, Y. Q. *Protein Sci.* **2002**, *11*, 2714−2726.

(74) Zhou, H. Y.; Zhou, Y. Q. *Protein Sci.* **2003**, *12*, 2121–2121.

(75) Melo, F.; Feytmans, E. *J. Mol. Biol.* **1997**, *267*, 207–222.

(76) Cascella, M.; Neri, M. A.; Carloni, P.; Dal Peraro, M. *J. Chem. Theory Comput.* **2008**, *4*, 1378–1385.

CT800441U